



# THE DISCIPLINE OF ORGANIZING



edited by **ROBERT J. GLUSHKO**

The Discipline of Organizing: 4th Professional  
Edition





# THE DISCIPLINE OF ORGANIZING



edited by **ROBERT J. GLUSHKO**





The Discipline of Organizing: 4th Professional Edition by Robert J. Glushko is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/), except where otherwise noted.

# Contents

|                                                       |      |
|-------------------------------------------------------|------|
| Foreword to the First Edition                         | xi   |
| <i>Foreword</i>                                       |      |
| Preface to the Fourth Edition                         | xvii |
| <i>Preface</i>                                        |      |
| Typographic Conventions                               | xxi  |
| <i>Typographic Conventions</i>                        |      |
| Murray Maloney                                        |      |
| Abstract                                              | xxvi |
| <i>Abstract</i>                                       |      |
| <br>Part I. <u>Foundations for Organizing Systems</u> |      |
| 1. The Discipline of Organizing                       | 29   |
| 2. The “Organizing System” Concept                    | 47   |
| 3. The Concept of “Resource”                          | 51   |
| 4. The Concept of “Collection”                        | 58   |
| 5. The Concept of “Intentional Arrangement”           | 64   |
| 6. The Concept of “Organizing Principle”              | 73   |
| 7. The Concept of “Agent”                             | 83   |
| 8. The Concept of “Interactions”                      | 85   |
| 9. The Concept of “Interaction Resource”              | 89   |
| 10. Organizing This Book                              | 91   |

## Part II. Design Decisions in Organizing Systems

|                                       |     |
|---------------------------------------|-----|
| 11. Introduction (II)                 | 99  |
| 12. What Is Being Organized?          | 103 |
| 13. Why Is It Being Organized?        | 110 |
| 14. How Much Is It Being Organized?   | 119 |
| 15. When Is It Being Organized?       | 130 |
| 16. How (or by Whom) Is It Organized? | 135 |
| 17. Where is it being Organized?      | 139 |
| 18. Key Points in Chapter Two         | 143 |

## Part III. Activities in Organizing Systems

|                                           |     |
|-------------------------------------------|-----|
| 19. Introduction (III)                    | 151 |
| 20. Selecting Resources                   | 159 |
| 21. Organizing Resources                  | 177 |
| 22. Designing Resource-based Interactions | 226 |
| 23. Maintaining Resources                 | 251 |
| 24. Key Points in Chapter Three           | 290 |

## Part IV. Resources in Organizing Systems

|                                       |     |
|---------------------------------------|-----|
| 25. Introduction (IV)                 | 301 |
| 26. Four Distinctions about Resources | 311 |
| 27. Resource Identity                 | 345 |
| 28. Naming Resources                  | 360 |
| 29. Resources over Time               | 382 |
| 30. Key Points in Chapter Four        | 396 |

## Part V. Resource Description and Metadata

|                                         |     |
|-----------------------------------------|-----|
| 31. Introduction (V)                    | 403 |
| 32. An Overview of Resource Description | 411 |
| 33. The Process of Describing Resources | 431 |
| 34. Describing Non-text Resources       | 495 |
| 35. Key Points in Chapter Five          | 509 |

## Part VI. Describing Relationships and Structures

|                                           |     |
|-------------------------------------------|-----|
| 36. Introduction (VI)                     | 517 |
| 37. Describing Relationships: An Overview | 521 |
| 38. The Semantic Perspective              | 524 |
| 39. The Lexical Perspective               | 545 |
| 40. The Structural Perspective            | 558 |
| 41. The Architectural Perspective         | 583 |
| 42. The Implementation Perspective        | 587 |
| 43. Relationships in Organizing Systems   | 592 |
| 44. Key Points in Chapter Six             | 601 |

## Part VII. Categorization: Describing Resource Classes and Types

|                                             |     |
|---------------------------------------------|-----|
| 45. Introduction (VII)                      | 609 |
| 46. The What and Why of Categories          | 613 |
| 47. Principles for Creating Categories      | 641 |
| 48. Category Design Issues and Implications | 677 |
| 49. Implementing Categories                 | 687 |
| 50. Key Points in Chapter Seven             | 718 |

Part VIII. Classification: Assigning Resources to  
Categories

|                                          |     |
|------------------------------------------|-----|
| 51. Introduction (VIII)                  | 727 |
| 52. Understanding Classification         | 749 |
| 53. Bibliographic Classification         | 773 |
| 54. Faceted Classification               | 785 |
| 55. Classification by Activity Structure | 803 |
| 56. Computational Classification         | 806 |
| 57. Key Points in Chapter Eight          | 809 |

Part IX. The Forms of Resource Descriptions

|                                |     |
|--------------------------------|-----|
| 58. Introduction (IX)          | 817 |
| 59. Structuring Descriptions   | 820 |
| 60. Writing Descriptions       | 865 |
| 61. Worlds of Description      | 884 |
| 62. Key Points in Chapter Nine | 896 |

Part X. Interactions with Resources

|                                             |     |
|---------------------------------------------|-----|
| 63. Introduction (X)                        | 903 |
| 64. Determining Interactions                | 913 |
| 65. Reorganizing Resources for Interactions | 926 |
| 66. Implementing Interactions               | 947 |
| 67. Evaluating Interactions                 | 966 |
| 68. Key Points in Chapter Ten               | 974 |

Part XI. The Organizing System Roadmap

|                       |     |
|-----------------------|-----|
| 69. Introduction (XI) | 981 |
|-----------------------|-----|

|                                                       |      |
|-------------------------------------------------------|------|
| 70. The Organizing System Lifecycle                   | 984  |
| 71. Defining and Scoping the Organizing System Domain | 986  |
| 72. Identifying Requirements for an Organizing System | 999  |
| 73. Designing and Implementing an Organizing System   | 1012 |
| 74. Operating and Maintaining an Organizing System    | 1021 |
| 75. Key Points in Chapter Eleven                      | 1029 |

## Part XII. Case Studies

|                                                      |      |
|------------------------------------------------------|------|
| 76. A Multi-generational Photo Collection            | 1039 |
| 77. Knowledge Management for a Small Consulting Firm | 1043 |
| 78. Smarter Farming in Japan                         | 1047 |
| 79. Single-Source Textbook Publishing                | 1051 |
| 80. Organizing a Kitchen                             | 1056 |
| 81. Earth Orbiting Satellites                        | 1061 |
| 82. CalBug and its Search Interface Redesign         | 1068 |
| 83. Weekly Newspaper                                 | 1074 |
| 84. The CODIS DNA Database                           | 1077 |
| 85. Honolulu Rail Transit                            | 1081 |
| 86. The Antikythera Mechanism                        | 1087 |
| 87. Autonomous Cars                                  | 1098 |
| 88. IP Addressing in the Global Internet             | 1102 |
| 89. The Art Genome Project                           | 1105 |
| 90. Making a Documentary Film                        | 1109 |
| 91. The Dabbawalas of Mumbai                         | 1113 |
| 92. Managing Information About Data Center Resources | 1119 |

|                                      |      |
|--------------------------------------|------|
| 93. Neuroscience Lab                 | 1123 |
| 94. A Nonprofit Book Publisher       | 1128 |
| Afterword                            | 1133 |
| Robert J. Glushko                    |      |
| Dedication                           | 1134 |
| Robert J. Glushko                    |      |
| Acknowledgments                      | 1135 |
| <i>Acknowledgments</i>               |      |
| Robert J. Glushko                    |      |
| Credits and Notices                  | 1141 |
| Robert J. Glushko and Murray Maloney |      |
| Dedication                           | 1143 |
| <i>Dedication</i>                    |      |
| Murray Maloney                       |      |
| Bibliography                         | 1144 |
| Murray Maloney                       |      |
| Glossary                             | 1272 |
| Murray Maloney                       |      |
| Index                                | 1369 |
| Murray Maloney                       |      |
| Colophon                             | 1604 |
| Murray Maloney                       |      |

# Foreword to the First Edition

## *Foreword*

This wonderful book arrives at the right time. It is more than a textbook—it defines and creates the field for which it is a text. Befitting a book that lays out a discipline of organization that spans print and digital media, this volume is carefully organized, with a focus on future print and digital editions.

[\*The Discipline of Organizing\*](#) has a broad scope. Even more valuable is its depth, the result of years of examining and thinking through related concepts—often overlapping but not identical—from the fields of library science, information science, business, and computer science. The rare combination of breadth and depth empowers readers by providing a new perspective and framework for organizing subsequent experiences. The organization is comprehensive and systematic, but it is not simple. A lot of concepts must be assimilated. Yet thanks to the authors' thoroughness, you can proceed confident that investing the time to master novel concepts will pay off, that a coherent structure is being assembled, without inconsistencies or confusions. Into this framework you can fit your own examples, alongside the many provided by the authors. You can identify extensions and form new associations, building on a strong foundation.

The authors ask us to step back and adopt a general, multidisciplinary perspective. This is unusual for a textbook. For good reason, the world is marked by increasing specialization, the division of labor on which complex civilization depends. First we master a discipline; then we are encouraged to be multidisciplinary, interdisciplinary, and transdisciplinary—to balance our specialized pursuits. But scholars thrive within single disciplines, and even in fields created as multidisciplinary efforts, such as the neurosciences



or cognitive science, most researchers soon become highly specialized. So why should a student of information undertake to master this broad perspective?

By growing from insignificance to centrality in the century-old field of information management in a few decades, digital technology has forced a disciplinary merger. Library science, information science, computer science or informatics, and information systems have developed different terminologies and sets of abstractions. Rather than asking each camp to learn the others' languages, the authors ask each of us to engage with a new terminology and set of abstractions.

The analogy of the artificial language Esperanto may come to mind, but we are in a better position. Esperanto is only useful if you are in a community of Esperantists. The abstractions in this book will be useful if others share them, but any reader will benefit by understanding the correspondences across the approaches to information organization that we encounter today. Unlike Esperanto, which is just another language, the concepts in this book reveal linkages and dependencies that we would not otherwise appreciate. The book provides a deep foundation for understanding changes that affect our lives and will do so more in the years ahead, a foundation that you will carry even if much of the time you converse in the language of one or another professional tribe.

Why do I say the timing is perfect, that this effort is worthwhile today? Have people not gotten by without it until now? The answer has two parts. One, which is important even if you have heard it before, is that this is a time of extraordinary change in our uses of information. The other is that people have not always "gotten by" very well; years have been wasted and careers damaged by not understanding the principles in this book. The likelihood of such wreckage is growing, as the waves of change are larger and come at us faster. On the positive side, the waves offer tremendous opportunity for accomplishment. The coming era of monster waves

may be risky if we surf with a narrow focus, but thrilling for those whose view extends up and down the shoreline. I believe that if you read this book, you will see this point and be glad you read it. Let me know.

We are used to hearing about Moore's law and related legislation, but familiarity lowers our guard. Human beings do not reason well about exponential growth: our experience is linear, not exponential. What we overlook is that exponential growth can proceed for a long time under the radar—one grain of rice, two grains, four grains, etc., not adding up to much, but when it reaches the point of having an impact, the impact comes so fast that we are unprepared for it. Decades passed before accessible digital technology could support high-quality photography, but when the time came, film photography disappeared so quickly that most major companies went bankrupt. Digital audio and video were a long time coming, then panicked and shuttered major industries. The expanding capacity and diminishing cost of information storage alter the balances described in this book. Bandwidth, increasing more slowly, is also reaching disruptive levels. This book provides the best tools available for understanding the disruptions of today and tomorrow in information management.

This perspective is invaluable now. It would have been useful earlier, but it was not considered imperative for the disciplines of library science, information science, informatics, and information systems. Historically they prospered despite interacting less than one might have expected. Library and information science, rooted in the humanities, focused technology efforts primarily on administrators and specialized users. Delivering services to the public was secondary. In contrast, academic computer science and human-computer interaction focused on widespread applications.

This book consciously connects fields that have focused on aspects of information organization and management such as archiving, records management, and curation, to information retrieval and

related aspects of informatics. It explores how related issues play out in different contexts. The authors are admirably positive. They do not drag us through the myriad disasters that resulted when library and information science did not understand the potential contributions of digital technology and the equally unfortunate disasters that resulted when technologists ignored a century of work on information organization.

However, I will sound a cautionary note about what might go wrong if you do not understand the principles laid out in this book. First, for computer scientists and engineers: Major system-building efforts foundered due to a lack of insight into the principles of information organization. I will describe an early one, whose protagonists, good or bad, right or wrong, are all equal now.

Although not a computer scientist or computer engineer, Vannevar Bush had as much influence on the field as anyone through his work on shaping government support for research after the Second World War and his 1945 Atlantic essay [\*As We May Think\*](#). Discussed in [\*Activities in Organizing Systems\*](#) [and [\*Describing Relationships and Structures\*](#)], this essay describes a hypothetical machine called the Memex that would enable information retrieval through a complex “associative memory” that supports links much like those found in the World Wide Web today. Although Bush’s design was based on microfilm and optical scanning rather than silicon, his vision has inspired countless researchers to this day.

Less well known are Bush’s classified efforts from the 1930s through early 1950s to build machines for the military and information agencies with Memex capabilities. Meticulously detailed by historian Colin Burke in the book [\*Information and Secrecy: Vannevar Bush, Ultra, and the other Memex\*](#), these projects consumed massive funding, occupied many brilliant MIT scientists for years, and produced nothing useful. A working machine was finally produced, but Bush never consulted with library science scholars who understood information organization from decades of work and

made naïve assumptions about how information could be organized for retrieval—the extraordinarily expensive machine was not usable in the real world.

Computer scientists interested in information should adopt a broad perspective, and this book is a place to begin. Several fields of computer science garner attention today, such as machine learning, data mining, information visualization, and design. Those centered on information, which is most directly affected by Moore's law, are likely to have the greatest impact.

The library and information science side of the bridge was also severely disrupted. Pride in a century of disciplinary accomplishment led to complacency and inertia. When silicon could no longer be avoided, there was not enough time to react gracefully. Major library schools closed. Today there are schools of information and a range of “library and information science” schools, some more forward-looking than others. Curriculum change has been relatively *ad hoc*, shaped by local personnel and context. Consider Information to be a large new volcanic cone pushing up in the midst of other peaks. No consistent approach has emerged to navigate the range. This book provides bridges where before there were slippery trails.

What can you do by virtue of reading and studying this book? Most importantly, perhaps, you can avoid confusion—when reading something or talking with someone from a different discipline, when asked a question in a job interview or by a colleague with a different background. Knowing that differences in terminology and abstractions are possible, you can ask questions and home quickly in on understanding. I have written elsewhere that through my career, such confusions frequently arose when I interacted with people in diverse disciplines, such as management information systems, software engineering, human factors, and so on. Because there was no book like this to clarify, it took me years to comprehend the source of many communication problems.

Beyond that, this book provides a foundation and framework for organizing and thinking about your experiences. This is a textbook, pointing to areas for research, providing ways of looking at new developments, and revealing to the perceptive reader yet unexplored territory in the spaces between disciplines. This is a book to read and put on the bookshelf—or in a folder in your digital reader—to reread in a few years' time.

Jonathan Grudin, 17 December 2012

# Preface to the Fourth Edition

## *Preface*

- [Preface to the First Edition](#)
- [Preface to the Second Edition](#)
- [Preface to the Third Edition](#)

A year ago I wrote “for a book to be republished in a third edition barely two years after its first is highly unusual, but we were compelled to update the book by the extremely positive reception it has received.” I also wrote that each new course that adopts the book “ratifies the idea that multiple perspectives can reinforce a shared focus on organizing, while at the same time highlighting the concepts, technologies, and methods that distinguish those points of view.”

It is happening again. In the last year the idea of [data science](#) as a new career field has led many universities to add new courses, modify existing ones, to hire new faculty, and even to change the names of schools or departments. Many people teaching with or studying [The Discipline of Organizing](#) have suggested that the book incorporate more discussion of data science concepts, and we’ve done that in this 4<sup>th</sup> edition.

The new methods and tools of [data science](#) and [machine learning](#) let us organize more information, to do it faster, and to make predictions based on what people have clicked on, bought, or said. Data science introduces new considerations of scale and speed when massive computational power and new statistical techniques are harnessed to organize and act on information.

But this is not the first time that new ideas and technologies have

challenged how people organized and interacted with resources, and it won't be the last. Data science will not replace human organizers, any more than any other science has replaced humans.

A data scientist needs to learn statistics, machine learning, and other new methods and technologies, and this book briefly sketches them, but does not try to teach them in any detail. However, data scientists need to understand the fundamental concepts of information organization, resource description, category design, and classification that are at the heart of this book. Data scientists need to select resources wisely and decide how best to describe them, they need to understand that resource description and categorization can be biased, they need to understand tradeoffs and complements between people and computers, and they need to understand when interpretability of features and organizing principles are more important than a bit more classification accuracy in a machine learning model.

The 4<sup>th</sup> edition builds a bridge between organizing and data science. It reframes descriptive statistics as organizing techniques, expands the treatment of classification to include computational methods, and incorporates many new examples of data-driven resource selection, organization, maintenance, and personalization. It introduces a new “data science” category of discipline-specific content, both in the chapter text and in endnotes, marked with [DS] in editions that contain endnotes.

New sections and sidebars include:

- New sidebar: [The Distinction between Data and Information](#)
- New sidebar: [Data Science and the Discipline of Organizing](#)
- New section: [“The Concept of “Interaction Resource””](#)
- New section: [“Organizing With Descriptive Statistics”](#)

- New section: [“Exploratory Analysis to Understand Data”](#)
- New section: [“Resource Description for Sensemaking and Science”](#)
- New sidebar: [Sensemaking and Organizing](#)
- New sidebar: [Geometric Distance Functions](#)
- New section: [“Implementing Categories Defined by Probability and Similarity”](#)
- New figure: [Figure: Rule-based Decision Tree](#)
- New figure: [Figure: Probabilistic Decision Tree](#)
- New sidebar: [Finding Friends and Dates: Lessons for Learning Categories](#)
- New sidebar: [Statistical Bias and Variance](#)
- New sidebar: [Bias and Variance on Dartboards](#)

Just as with the 2nd and 3rd editions, we are publishing the 4<sup>th</sup> edition in a “Professional Edition” that contains all of the discipline-tagged supplemental content and endnotes, and in a simplified “Core Concepts Edition” that omits all supplemental content. In addition, the 4<sup>th</sup> edition is being published in an “Informatics Edition” that includes all the new content related to data science, but omits the discipline-specific content about library science, museums, and document archives.

Many instructors, students, and readers identified content in previous editions that was inaccurate, confusing, redundant, or missing, and I thank them as a group. We have worked hard to resolve every concern, but as prefaces often say, any remaining flaws are our responsibility.



However, there are some whose contributions to this 4<sup>th</sup> edition have been so substantial that it would be thoughtless not to thank them by name. Most of them were participants in a “book club” seminar at Berkeley during the 2015-16 academic year that thoroughly deconstructed a number of books to help us design, build, and cross the bridge between organizing and data science. These books included [\*The Signal and the Noise\*](#) by Nate Silver, [\*Predictive Analytics\*](#) by Eric Siegel, and [\*The Master Algorithm\*](#) by Pedro Domingos. The book club participants were Pascual Arrechea, Stacey Baradit, Dina Bseiso, Phil Braddock, Bill Chambers, Jason Danker, Laura Desmond-Black, Paul Glenn, Daniel Griffin, Rob Kuvinka, Molly Mahar, Emily Paul, Robyn Perry, Keshav Potluri, Shom Sarkar, Jordan Shedlock, Vijay Velagapudi, and Emily Witt. We learned a lot together, and the important things we learned are now in this 4<sup>th</sup> edition.

Other Berkeley students and alums who reviewed the 4<sup>th</sup> edition include Andy Brooks, Lisa Jervis, Ian MacFarland, Jason Ost, and Richa Prajapati.

Instructors who teach with *The Discipline of Organizing* can easily see places to improve it, but David Bamman, John King, Vivien Petras, Isabelle Sperano, Mikael Gunnarsson, Yasar Tonta, and Nina Wacholder reviewed the book carefully and proposed new content.

Without Robyn Perry’s contributions as an author, reviewer, and graphic artist, we would not have finished this edition in time for the 2016-2017 academic year. The 4<sup>th</sup> edition was produced by Murray Maloney, as markup and production editor.

Robert J. Glushko, 5 August 2016

# Typographic Conventions

## *Typographic Conventions*

MURRAY MALONEY

Our typographic conventions are a bit less traditional than the overall composition of the book. Our goal is to make the reading experience easier for you by subtly emphasizing more important material, while de-emphasizing content that is less central, yet still relevant. For example, we emphasize the definition of terms and discussion of organizing principles while we de-emphasize editorial asides and parenthetical references.

These formatting choices are intended to facilitate cognitive uptake. Where colored text is used, we chose dark colors to maintain contrast, while still offering subtle visual cues.

**Definition of term.** A sentence or paragraph in which a term is formally defined. The subtle underline color is Silver, a CSS built-in. Hover effect is to black underline the definition and embolden the term.

**Editorial content:** In which the authors and editors discuss the book as a book, cross-references and way-finding, and the editorial process; or, ruminations upon matters quite apart from the topic of the book itself; material that won't be on the quiz; such as, for example, this sentence, or this entire section for that matter. Hover effect restores text opacity.

**Organizing principle.** Alphabetic ordering is an organizing principle. The text color is #004400, a dark green. Hover effect is bold text.

**Parenthetical content:** Typically used for asides, such as references to related material. (For a broader discussion of this

topic, see: [\*Activities in Organizing Systems\*](#).) Hover effect restores text opacity.

Statement. A topic sentence, or a richly semantic sentence or paragraph. Often, these are pithy sentences that might otherwise be overlooked, or hard to re-discover. The text color is Indigo, a CSS built-in. Hover effect is bold text.

## Hypertext Links

To reduce the strain on our readers' eyes, we have chosen to tone down all the glaring links we normally encounter. Rather than employing the typical blue underlined links that scream for attention, our links are presented in a darker, maroon color, without underline. Again, we hope that this will reduce distraction and facilitate cognitive uptake. The link color is #840024, a maroon. The hover effect restores the link to blue, underlined text to assure the gentle reader that this is, in fact, a hypertext link. Hover over a link to see the title of the target. Following are examples of some link types; other link types follow the same pattern.

- Chapter reference: See [Case Studies](#)
- Section reference: See [“What Is Being Organized?”](#)
- Sidebar reference: See the sidebar, [Perspectives on Hypertext Links](#)
- Glossary term reference: Links to terms in the [Glossary](#), such as [discipline](#) and [organizing principles](#)
- Footnotes: This is a footnote marker.<sup>1</sup> Follow the footnote to see the return arrow. Hover over the footnote marker to see

1. This is a footnote. The arrow below returns to the footnote marker.

the contents of the footnote in a text-box popup.

- Citations: ([Nunberg 1996, 2011](#)). ([Buckland 1991](#)). See also ([Bates 2005](#)). Hover over to see title of the work associated with the citation.
- Bibliography backlinks: See Note at the beginning of the [Bibliography](#).
- Glossary links and backlinks: See Note at the beginning of the [Glossary](#).
- Index referents: See the [Index](#).
- External links: [Discipline of Organizing](#) website.

## Disciplines

This book attempts to represent all the disciplines that contribute to it, without compromising the need for depth, to treat each contributing discipline in a substantive way. Our design solution had been a core text with disciplinary and domain-specific content in hundreds of supplemental endnotes tagged by discipline. Eventually, we also started tagging core paragraphs and sentences. This design allows the book to emphasize concepts that bridge the different organizing disciplines while satisfying the additional topical needs of different academic programs.

Hundreds of paragraphs, sentences, and endnotes tagged with a lozenge-like prefix indicate that the content may be especially relevant within the realm of that discipline. The most tagged disciplinary paragraphs and endnotes are computing, library and information science, business, web, and cognitive science. Data science is closing the gap.

- Business: Intellectual capital, human resources, access control, branding, decision support and strategic planning, economics.
- Computing: Computer science, software engineering, and computing technology.

- - DS: Data Science, machine learning, statistical analysis, decision trees, scale and speed considerations, and new techniques for resource selection and classification.
  - Information Architecture: Data and document modeling. Information architecture is about the design of information models and their systematic manifestation in user experiences, for people using web sites, or in other contexts that are information-intensive.
  - Web: Web architecture, web standards, and particular web sites and web services.
- LIS: Library and information science; these are not the same, but this is a conventional disciplinary category. Issues that apply broadly to memory institutions (i.e., libraries, museums, and archives).
- - Archives: Issues that apply more narrowly to archives.
  - Museums: Issues that apply more narrowly to museums and cultural collections.
- CogSci: Cognitive science at the intersection of psychology, linguistics, philosophy, anthropology, and computer science. Broadly discusses human perception, decision making, problem solving, and other activities that affect organizing systems and actions, especially personal ones.
  - Linguistics: Issues that apply more narrowly to language construction and use.
  - Philosophy: Apply more narrowly to philosophy.
- Law: Copyright law, license or contract agreements, “cultural property,” terms of use, and so on, and so forth.

And so on, and so forth and scooby dooby dooby. Oooh cha cha. —  
Everyday People, Sly Stone

# Abstract

## *Abstract*

We organize things, we organize information, we organize information about things, and we organize information about information. But even though “organizing” is a fundamental and ubiquitous challenge, when we compare these activities their contrasts are more apparent than their commonalities. We propose to unify many perspectives about organizing with the concept of an Organizing System, defined as an intentionally arranged collection of resources and the interactions they support. Every Organizing System involves a collection of resources, a choice of properties or principles used to describe and arrange resources, and ways of supporting interactions with resources. By comparing and contrasting how these activities take place in different contexts and domains, we can identify patterns of organizing. We can create a discipline of organizing in a disciplined way.

Robert J. Glushko

PART I

# FOUNDATIONS FOR ORGANIZING SYSTEMS

**Robert J. Glushko**





# I. The Discipline of Organizing

To *organize* is to create capabilities by intentionally imposing order and structure.

Organizing is such a common [activity](#) that we often do it without thinking much about it. We organize shoes in our closet, books on our book shelves, spices in our kitchen, receipts and records in tax preparation folders, and people on business projects and sports teams. Quite a few of us have jobs that involve specific types of organizing tasks. We might even have been explicitly trained to perform them by following specialized disciplinary practices. We might learn to do these tasks very well, but even then we often do not reflect on the similarity of the organizing tasks we do and those done by others, or on the similarity of those we do at work and those we do at home. We take for granted and as givens the concepts and methods used in the Organizing System we work with most often.

The goal of this book is to help readers become more self-conscious about what it means to organize resources of any type and about the principles by which the resources are organized. In particular, this book introduces the concept of an *Organizing System*: an intentionally arranged collection of resources and the interactions they support. The book analyzes the design decisions that go into any systematic organization of resources and the design patterns for the interactions that make use of the resources, as follows:

**We organize physical things.** Each of us organizes many kinds of things in our lives—our books on bookshelves; printed financial records in folders and filing cabinets; clothes in dressers and closets; cooking and eating utensils in kitchen drawers and cabinets. Public libraries organize printed books, periodicals, maps, CDs, DVDs, and maybe some old record albums. Research libraries also

organize rare manuscripts, pamphlets, musical scores, and many other kinds of printed information. Museums organize paintings, sculptures, and other artifacts of cultural, historical, or scientific value. Stores and suppliers organize their goods for sale to consumers and to each other. Sports leagues organize players into teams, and the teams organize players by position or role.

**We organize information about physical things.** Each of us organizes information about things: when we inventory the contents of our house for insurance purposes, when we sell our unwanted stuff on eBay, or when we rate a restaurant on Yelp. Library card catalogs, and their online replacements, tell us what books a library's collection contains and where to find them. Sensors and RFID tags track the movement of goods—even library books—through supply chains, and the movement (or lack of movement) of cars on highways.

**We organize digital things.** Each of us organizes personal digital information—email, documents, ebooks, MP3 and video files, appointments, contacts—on our computers, smartphone, ebook readers, or in “the cloud,” —through information services that use Internet protocols. Large research libraries organize digital journals and books, computer programs, government and scientific datasets, databases, and many other kinds of digital information. Companies organize their digital business records and customer information in enterprise applications, content repositories, and databases. Hospitals and medical clinics maintain and exchange electronic health records and digital X-rays and scans.

**We organize information about digital things.** Digital library catalogs, web portals, and aggregation websites organize links to other digital resources. Web search engines use content and link analysis along with relevance ratings, to organize the billions of web pages competing for our attention. Web-based services, data feeds and other information resources can be interconnected and choreographed to carry out information-intensive business processes, or aggregated and analyzed to enable prediction and personalization of information services.

Let us take a closer look at these four different types or contexts of organizing. We contrasted “organizing things” with “organizing information.” At first glance it might seem that organizing physical things like books, compact discs, machine parts, or cooking utensils has an entirely different character than organizing intangible digital things. We often arrange physical things according to their shapes, sizes, material of manufacture, or other intrinsic and visible properties: for example, we might arrange our shirts in the clothes closet by style and color, and we might organize our music collection by separating the old vinyl albums from the CDs. We might arrange books on bookshelves by their sizes, putting all the big, heavy picture books on the bottom shelf. Organization for clothes and information artifacts in tangible formats that is based on visible properties does not seem much like how you store and organize digital books on your Kindle or arrange digital music on your music player. Arranging, storing, and accessing X-rays printed on film might appear to have little in common with these activities when the X-rays are in digital form.

It is hardly surprising that organizing things and organizing information sometimes do not differ much when information is represented in a tangible way. The era of ubiquitous digital information of the last decade or two is just a blip in time compared with the more than ten thousand years of human experience with information carved in stone, etched in clay, or printed with ink on papyrus, parchment, or paper. These tangible information artifacts have deeply embedded the notion of information as a physical thing in culture, language, and methods of information design and organization. This perspective toward tangible information artifacts is especially prominent in rare book collections where books are revered as physical objects with a focus on their distinctive binding, calligraphy, and typesetting.

Nevertheless, at other times there are substantial differences in how we organize things and how we organize information, even when the latter is in physical form. We more often organize our

“information things” according to what they are about rather than on the basis of their visible properties. At home we sort our CDs by artist or genre; we keep cookbooks separate from travel books, and fiction books apart from reference books. Libraries employ subject-based classification schemes that have a few hundred thousand distinct categories.

Likewise, there are times when we pay little attention to the visible properties of tangible things when we organize them and instead arrange them according to functional or task properties. We keep screwdrivers, pliers, a hammer, a saw, a drill, and a level in a toolbox or together on a workbench, even though they have few visual properties in common. We are not organizing them because of what we see about them, but because of what we know about to use them. The task-based organization of the tools has some similarity to the subject-based organization of the library.

We also contrasted “organizing things” with “organizing information about things.” This difference seems clear if we consider the traditional library card catalog, whose printed cards describe the books on library shelves. When the things and the information about them are both in physical format, it is easy to see that the former is a primary resource and the latter a surrogate or associated resource that describes or relates to it.

#### What Is Information?

Most of the hundreds of definitions of information treat it as an idea that swirls around equally hard-to-define terms like “data,” “knowledge,” and “communication.” Moreover, these intellectual and ideological perspectives on information coexist with more mundane uses of the term, as when we ask a

station agent: “Can you give me some information about the train schedule?”

An abstract view of information as an intangible thing is the intellectual foundation for both modern information science and the information economy and society. Nevertheless, the abstract view of information often conflicts with the much older idea that information is a tangible thing that naturally arose when information was inextricably encoded in material formats. We often blur the sense of “information as content” with the sense of “information as container,” and we too easily treat the number of stored bits on a computer or in “the cloud” as a measure of information content or value.

Geoff Nunberg has eloquently explained in [Farewell to the Information Age](#) that information is “a collection of notions, rather than a single coherent concept.” Michael Buckland’s oft-cited essay [Information as Thing](#) argues against the notion that information is inherently intangible and instead defines it more broadly and provocatively based on function. A resource that can be learned from or serve as evidence is “information-as-thing,” a definition that treats the tangible objects in museum or personal collections as information.<sup>1</sup>

When it comes to “organizing information about digital things” the

1. [\(Nunberg 1996, 2011\)](#). [\(Buckland 1991\)](#). See also [\(Bates 2005\)](#).

contrast is much less clear. When you search for a book using a search engine, first you get the catalog description of the book, and often the book itself is just a click away. When the things and the information about them are both digital, the contrast we posed is not as sharp as when one or both of them is in a physical format. And while we used X-rays—on film or in digital format—as examples of things we might organize, when a physician studies an X-ray, is it not being used as information about the subject of the X-ray, namely, the patient? And when businesspeople make marketing and pricing decisions by analyzing digital information about what and when people buy, we can think of this as organizing customers into categories, or as organizing customer information.

These differences and relationships between “physical things” and “digital things” have long been discussed and debated by philosophers, linguists, psychologists, and others. (See the sidebars, [What Is Information?](#) and [The Distinction between Data and Information.](#))

The distinctions among organizing physical things, organizing digital things, or organizing information about physical or digital things are challenging to describe because many of the words we might use are as overloaded with multiple meanings as “information” itself. For example, the library science perspective often uses presentation or implementation properties in definitions of “document,” using the term to refer only to traditional physical forms. In contrast, the informatics or computer science perspective takes an abstract view of “document” to refer to any self-contained unit of information, separating a document’s content from its presentation or container.<sup>2</sup>

The most abstract definition of “document,” presented in [What is](#)

## 2. ([Glushko and McGrath 2005](#)).

[a Document?](#) follows from Buckland's "information as thing" idea. Because it can be studied to provide evidence, an antelope is both "information as thing" and also a "document" when it is in a zoo, even though it is just an animal when it is running wild on the plains of Africa. However, in 2015 the United States Supreme Court rejected this expansive definition in a case that hinged on whether a fish could be viewed as a document.<sup>3</sup>

3. ([Buckland 1997](#)). The idea that an antelope could be a document was first proposed in ([Briet 1951](#)).

A commercial fisherman in Florida was found with fish in his catch below the legal size limit. An inspector ordered him to return to port and hand the fish over to the authorities; when he dumped them overboard instead, he was charged with violating the Sarbanes-Oxley Act, a law drafted in response to high-profile white-collar crimes such as the Enron scandal. The law imposes harsh penalties for destroying "any record, document, or tangible object" to impede a federal investigation. The fisherman argued that the law should only apply to written documents, but the United States government contended that because the fish were "tangible objects" whose presence on the boat served as the only documentation of the allegedly illegal fishing, there was no practical difference between a fish and a document in this case. The Supreme Court ruled in favor of the fisherman, finding that "tangible object" must be interpreted in the context of "record" and "document" and, as such, only applies to an object "used to record



### The Distinction between Data and Information

Astute readers might have noticed that we included sensor data as “information about physical things” and data feeds as “information about digital things.” Many textbooks in the information science and knowledge management fields distinguish data and information in a more precise way. To them, data sits at the bottom of an Information Hierarchy, Knowledge Pyramid, or DIKW Hierarchy in which Data is transformed into Information, which is transformed into Knowledge, which is then transformed into Wisdom.

or preserve information.” The fact that a fish is tangible evidence in this case does not make it a document.

[\(Buckland 1991\).](#)

[\(Liptak 2014\).](#) Brief for the United States in Opposition, *Yates v. United States*. SCOTUSblog, March 14, 2014.

<http://www.scotusblog.com/case-files/cases/yates-v-united-states/>

For the complete history of the case, see:

<http://www.scotusblog.com/case-files/cases/yates-v-united-states/>.

See also the related Sarbanes-Oxley Act endnote.<sup>4</sup>

4. [link to footnote]

In this framework, data are raw or elementary observations about properties of objects, events, and their environment. Data becomes information when it is aggregated, processed, analyzed, formatted, and organized to add meaning and context so it can be used to answer questions. This processing can include calculation, inference, or refinement operations on the data. For example, measurements of temperature, precipitation, and wind speed are data. When combined and summarized, a set of data becomes statistical information about the weather on a particular day. When collected over a period of months or years, these datasets become information about the climate of the location where they were collected.

The Discipline of Organizing does not make this sharp contrast between data and information in the Hierarchy/Pyramid. People who read this book are likely to be aspiring or practicing professionals in information-intensive industries where information and data are often treated as synonyms to mean the content of a database or data-managing application. A distinction between data and information might be useful in theory, but not in these applied settings.

The distinction between data and information is also being blurred by the expansion in the scope of the definition of data in the emerging career field of [data science](#). Indeed, a popular introductory text eliminates information entirely from the Hierarchy/Pyramid with

its title, [\*Discovering knowledge in data: an introduction to data mining.\*](#)<sup>5</sup>

Similar definitional variation occurs with “author” or “creator.” When we say that “Herman Melville is the author of [\*Moby Dick\*](#)” ([Melville 1851](#)) the meaning of “author” does not depend on whether we have a printed copy or an ebook in mind, but what counts as authorship varies a great deal across academic disciplines. Furthermore, different standards for describing resources disagree in the precision with which they identify the person(s) or organization(s) primarily responsible for creating the intellectual

5. The DIKW hierarchy seems to have been inspired by [\*The Rock, A Pageant Play\*](#) ([Eliot 1934](#)) by the poet T S Eliot, whose opening chorus contains these lines:

Where is the wisdom we have lost in knowledge?  
Where is the knowledge we have lost in information?

Most people credit Ackoff's [\*From Data to Wisdom\*](#) ([Ackoff 1989](#)) as the first articulation of the hierarchy in an information science and systems context. The hierarchy is mentioned in nearly twenty textbooks, but their close analysis by ([Rowley 2007](#)) reveals only partial agreement on the definitions and relationships among the four key concepts. The hierarchy has been criticized as lacking in philosophical rigor ([Fricke 2009](#)) and for ignoring the context-specificity of how knowledge is learned and applied ([Jennex 2009](#)). ([Larose 2014](#))

content of the resource, People who are serious about music description rightly criticize streaming services and online stores that have only a single “artist” field because this fails to distinguish the composer, conductor, orchestra, and other people with distinct roles in creating the music.

If we allow the concept of information to be anything we can study—to be “anything that informs”—the concept becomes unbounded. Our goal in this book is to bridge the intellectual gulf that separates the many disciplines that share the goal of organizing but differ in what they organize. This requires us to focus on situations where information exists because of intentional acts to create or organize. (See the sidebar, [The Discipline of Organizing](#))

#### The Discipline of Organizing

A *discipline* is an integrated field of study in which there is some level of agreement about the issues and problems that deserve study, how they are interrelated, how they should be studied, and how findings or theories about the issues and problems should be evaluated. A *framework* is a set of concepts that provide the basic structure for understanding a domain, enabling a common vocabulary for different explanatory theories.

[Organizing](#) is a fundamental issue in many disciplines, most notably library and information science, computer science, systems analysis, informatics, law, economics, and business. However, these disciplines have only limited agreement in how they approach problems of organizing and what they seek as their solutions. For example, library and information science has traditionally studied organizing from a public sector

bibliographic perspective, paying careful attention to user requirements for access and preservation, and offering prescriptive methods and solutions.<sup>6</sup>

In contrast, computer science and informatics tend to study organizing in the context of information-intensive business applications with a focus on process efficiency, system architecture, and implementation. The disciplines of management and industrial organization deal with the organization of human, material, and information resources in contexts shaped by commercial, competitive, and regulatory forces.

This book presents a more abstract framework for

6. We can continue the debate in the previous paragraphs and the sidebar, [What Is Information?](#) by pointing out that in both common and professional usage, “bibliographic” activities involve describing and organizing information resources of the kinds that might be found in a library. But noted information scientist Patrick Wilson argued for a much broader expanse of the bibliographic universe, suggesting that “it includes manuscripts as well as printed books, bills of lading and street signs as well as personal letters, inscriptions on stone as well as phonograph recordings of speeches, and most notably, memorized texts in human heads and texts stored up in the *memories* of machines” ([Wilson 1968, p. 12](#)).

issues and problems of organizing that emphasizes the common concepts and goals of the disciplines that study them. Our framework proposes that every system of organization involves a collection of resources, and we can treat physical things, digital things, and information about such things as resources. Every system of organization involves a choice of properties or principles used to describe and arrange the resources, and ways of supporting interactions with the resources. By comparing and contrasting how these activities take place in different contexts and domains, we can identify patterns of organizing and see that Organizing Systems often follow a common life cycle. We can create a discipline of organizing in a disciplined way.

Many of the foundational topics for a discipline of organizing have traditionally been presented from the perspective of the library sector and taught as “library and information science.” These include bibliographic description, classification, naming, authority control, curation, and information standards. In recent decades these foundations have been built on and extended by computer science, cognitive science, informatics, and other new fields to include more private sector and non-bibliographic contexts, multimedia and social media, and new information-intensive applications and service systems enabled by mobile, pervasive, and scientific computing. The latest additions to the discipline of organizing are coming from [data science](#) and [machine learning](#), introducing considerations of speed and scale that arise when massive computational power and new statistical techniques are harnessed to organize and act on information.

The new methods and tools of [data science](#) and [machine learning](#)

let us organize more information, to do it faster, and to make predictions based on what people have clicked on, bought, or said. But this is not the first time that new ideas and technologies have challenged how people organized and interacted with resources. Fifty years ago, searchable online catalogs radically changed how people used libraries. The web, invented less than thirty years ago so that scientists could share technical reports, is now an essential part of many human activities. It is important not to view the latest new thing as changing everything, because new things will continue to come, and these technology breakthroughs still depend on and complement the organizing work done by people. Data science will not replace human organizers, any more than any other science has replaced humans. (See sidebar, [Data Science and the Discipline of Organizing](#)).

This is why we need to take a transdisciplinary view that lets us emphasize what the different disciplines have in common and how they fit together rather than what distinguishes them. Resource selection, organizing, interaction design, and maintenance are taught in every discipline, but these concepts go by different names. A vocabulary for discussing common organizing challenges and issues that might be otherwise obscured by narrow disciplinary perspectives helps us understand existing systems of organizing better while also suggesting how to invent new ones by making different design choices.

Data Science and the Discipline of Organizing  
Advances in computing power and statistical techniques are making it possible to identify patterns in data and extract meaningful information at a scale never before possible. Many books and articles about [data science](#), machine learning, and predictive analytics make

bold predictions that these emerging fields will radically change the world. These claims are both provocative and promising, but at its core, data science is about how resources are selected, described, and organized; concepts with a long tradition in information and library science. Instead of organizing and describing the books in a library or the products in a warehouse, a data scientist might organize information about books or products into massive data tables, treating each resource as a row and its descriptive properties as the columns. After people might have organized books or products into categories, machine learning techniques might classify new books or products using those categories, or perhaps discover new categories based on access or purchasing behaviors. So while the techniques of data science are new, many of the challenges are not; data scientists need to select resources wisely and decide how best to describe them; they need to understand that resource description and categorization can be biased; they need to understand the tradeoffs and complements between people and computers; and, they need to test the discoveries that algorithms make with controlled experiments.

To make sense of the discussions around data science, one must understand the difference between kind and degree. A hundred years ago, a car's highway travel speed was about forty miles an hour. Today's cars travel twice as fast, but this is just a change in degree. However, an increase in speed to about 17,500 miles an hour achieves an "orbital velocity" that allows us to go into Earth orbit in space, travel that is different in kind.



What about data science? Some data science involves collections of data that are “tall,” containing many millions or even billions of records that each have a relatively small number of variables. Being able to analyze “tall” data more rapidly than ever before is primarily a change in degree compared with traditional database techniques. Nevertheless, for collections of data that are “wide,” where each record might contain hundreds or thousands of variables, data science techniques might allow us to see patterns that could not be seen at all, or could not be seen affordably and in quantity. Here, data science might be yielding changes in kind.<sup>7</sup>

7. Siegel’s [\*Predictive Analytics: The Power to Predict who will Click, Buy, Lie or Die\*](#) (Siegel 2013) is written for a non-technical audience and enthusiastically describes over 100 applications. [\*The Master Algorithm\*](#) (Domingos 2015) shares Siegel’s enthusiasm but is far more technical; the book attempts to explain and compare the five “tribes” of machine learning: the symbolists, connectionists, evolutionaries, Bayesians, and analogizers. The title of Chris Anderson’s provocative article in Wired Magazine ([Anderson, 2008](#)) is self-explanatory: “The end of theory: The data deluge makes the scientific method obsolete.”

“Difference in kind or difference in degree” is an

important issue in legal contexts and more generally arises whenever there is a disagreement about whether some difference or change is strict and categorical or whether it is incremental. We introduce it here so that readers can think critically about the socio-business-technical changes that might come about as a result of new methods and technologies for organizing and analyzing data. We believe that data science is on its way to becoming an important part of the organizing tool box. But everyone needs to remember that humans own the tool box, and that they design and build the tools.

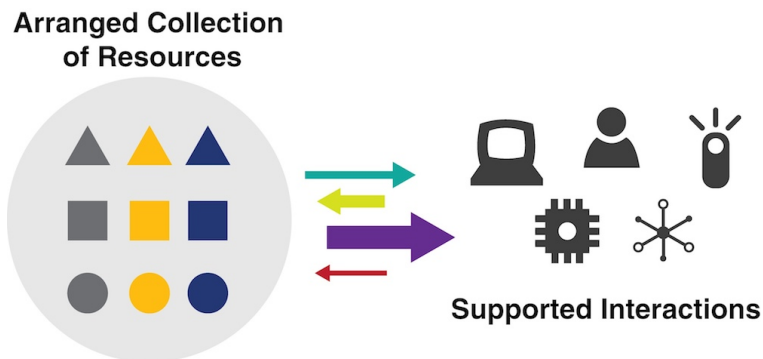


## 2. The “Organizing System” Concept

We propose to unify many perspectives about organizing and information with the concept of an [Organizing System](#), an *intentionally arranged collection of resources and the interactions they support*. This definition brings together several essential ideas that we will briefly introduce in this chapter and then develop in detail in subsequent chapters.

[Figure: An Organizing System](#) depicts a conceptual model of an Organizing System that shows intentionally arranged resources, interactions (distinguished by different types of arrows), and the human and computational agents interacting with the resources in different contexts.

An Organizing System



An Organizing System is a collection of resources arranged in ways that enable people or computational agents to interact with them.

An Organizing System is an abstract characterization of how some collection of resources is described and arranged to enable human or computational agents to interact with the resources. The Organizing System is an architectural and conceptual view that is distinct from the physical arrangement of resources that might embody it, and also distinct from the person, enterprise, or institution that implements and operates it. These distinctions are sometimes hard to maintain in ordinary language; for example, we might describe some set of resource descriptions, organizing principles, and supported interactions as a “library” Organizing System. However, we also need at times to refer to a “library” as the institution in which this Organizing System operates, and of course the idea of a “library” as a physical facility is deeply engrained in language and culture.

Our concept of the Organizing System was in part inspired by the concepts proposed in 2000 for bibliographic domains by Elaine Svenonius, in [The Intellectual Foundation of Information Organization](#). She recognized that the traditional [information organization](#) activities of bibliographic description and cataloging were complemented, and partly compensated for, by automated text processing and indexing that were usually treated as part of a separate discipline of [information retrieval](#). Svenonius proposed that decisions about organizing information and decisions about retrieving information were inherently linked by a tradeoff principle and thus needed to be viewed as an interconnected system: “*The effectiveness of a system for accessing information is a direct function of the intelligence put into organizing it*” (p.ix). We celebrate and

build upon her insights by beginning each of the sub-parts of [Design Decisions in Organizing Systems](#) with a quote from her book.<sup>1</sup>

A systems view of information organization and information retrieval captures and provides structure for the inherent tradeoffs obscured by the silos of traditional disciplinary and category perspectives: the more effort put into organizing information “on the way in” when it is created or added to a collection, the more effectively it can be retrieved, and the more effort put into retrieving information “on the way out,” the less it needs to be organized first. Sometimes a collection of resources is highly organized, but because it was organized by someone else for different purposes that have in mind, we need to reorganize it “on the way in.” This is especially common with digital text or datasets, where previously organized resources or their descriptions might be sorted, translated in format or language, combined, summarized, or otherwise transformed to fit into a new Organizing System. For example, to understand seasonal buying patterns, a retailer might combine shopping data with weather data and calendar data about commonly-watched sporting events (because bad weather and broadcast sports cause people to stay home), and all three datasets would need to describe “time” and “location” in the same way.

A systems view no longer contrasts information organization as a human activity and information retrieval as a machine activity, or information organization as a topic for library and information science and information retrieval as one for computer science. Instead, we readily see that computers now assist people in organizing and that people contribute much of the information used when computers analyze and organize resources. For example,

1. ([Svenonius 2000](#)).

many algorithms for computational classification use [supervised learning](#) approaches that start with items classified by people.

Finally, a systems view can be applied to Organizing Systems with any kind of resource, enabling more nuanced discussion of how economic, social, and cognitive costs and benefits of organizing are allocated among different stakeholders. Every Organizing System is biased by the perspectives and experiences of the people who create it. Some of these biases are inescapable, a kind of automatic organizing, because they reflect innate human perceptual and cognitive capabilities. Our minds impose structure and find patterns, even when there aren't any, and we are not capable of acting perfectly rationally, so we simplify without realizing it. People are also not very good at thinking about future possibilities and revising their expectations given new evidence, and this mental inertia makes us preserve resources and interactions in Organizing Systems that are no longer needed. Other biases in Organizing Systems reflect more intentional choices that implicitly or explicitly create winners or losers, treat some interactions as preferred while deprecating others, or otherwise impose or overlay a set of values on the stakeholders of the system. For example, many Organizing Systems arrange people in groups or queues to make interactions more efficient, but when an airline gives boarding priority to customers who paid more for their tickets it might not seem fair to you if you are in the last boarding group.

### 3. The Concept of “Resource”

*Resource* has an ordinary sense of anything of value that can support goal-oriented activity. This definition means that a resource can be a physical thing, a non-physical thing, information about physical things, information about non-physical things, or anything you want to organize. Other words that aim for this broad scope are [entity](#), [object](#), [item](#), and [instance](#). [Document](#) is often used for an information resource in either digital or physical format; [artifact](#) refers to resources created by people, and [asset](#) for resources with economic value.

[Resource](#) has specialized meaning in Internet architecture. It is conventional to describe web pages, images, videos, and so on as [resources](#), and the protocol for accessing them, Hypertext Transfer Protocol(HTTP), uses the Uniform Resource Identifier(URI).<sup>1</sup>

1. The [URI](#) identifies a resource as an abstract entity that can have “multiple representations,” which are the “things” that are actually exposed through applications or user interfaces. The [HTTP](#) protocol can transfer the representation that best satisfies the content properties specified by a web client, most often a browser. This means that interactions with web resources are always with their representations rather than directly with the resource *per se*. The representation of the resource might seem to be implied by the [URI](#) (as when it ends in `.htm` or `.html` to suggest text in Hypertext Markup Language(HTML) format), but the [URI](#) is not required to



## Concert Tickets

indicate anything about the “representation.” A web resource can be a static web page, but it can also be dynamic content generated at the time of access by a program or service associated with the [URI](#). Some resources like geolocations have “no representations at all;” the resource is simply some point or space and the interaction is “show me how to get there.” The browser and web server can engage in “content negotiation” to determine which “representation” to retrieve, and this is particularly important when that format further requires an external application or “plug-in” in order for it to be rendered properly, as it does when the server returns a Power Point file or an other file format that is not built into the browser.

Internet architecture’s definition of [resource](#) as a conceptual entity that is never directly interacted with is difficult for most people to apply when those resources are physical or tangible objects, because then it surely seems like we are interacting with something real. So we will most often talk about interactions with resources, and will mention “resource representations” only when it is necessary to align precisely with the narrower Internet architecture sense.



Tickets are physical artifacts that convey event-related [metadata](#): including time, place, and seat number; price and terms of admission; and featured performers. For concert goers, tickets offer the promise of all that, and a memory of the ineffable quality of more.

(Photo by Murray Maloney.)

What IS a Concert Ticket?

A concert ticket is a vehicle for conveying a package of assertions about an event, so it is a [description resource](#), like a card in a library card catalog. A concert ticket is also a [resource](#) in its own right, with intrinsic value; it can be bought and sold, sometimes for a greater price than its [resource description](#) specifies. A ticket is a license to use a seat in a venue for a specified purpose at a specified time; after the event, the ticket loses its intrinsic value, but might acquire extrinsic value as an artifact in a collection like this one.

Treating as a *primary resource* anything that can be identified is an important generalization of the concept because it enables web-based services, data feeds, objects with RFID tags, sensors or other [smart devices](#), or computational agents to be part of Organizing Systems.

Instead of emphasizing the differences between tangible and intangible resources, we consider it essential to determine whether the tangible resource has information content—whether it needs to be treated as being “about” or representing some other resource rather than being treated as a thing in itself. Whether a book is printed or digital, we focus on its information content, what it is about; its tangible properties become secondary. In contrast, the hangars in our closet and the measuring cups in our kitchen are not about anything more than their obvious utilitarian features, which

makes their tangible properties most important. (Of course, there is no sharp boundary here; you can buy “fashion hangers” that make a style statement, and the old measuring cup could be a family memento because it belonged to Grandma).

Many of the resources in Organizing Systems are *description resources* or [surrogate resources](#) that describe the [primary resources](#); library catalog entries or the list of results in web search engines are familiar examples. In museums, information about the production, discovery, or history of ownership of a resource can be more important than the resource; a few shards of pottery are of little value without these associated information resources. Similarly, business or scientific data often cannot be understood or analyzed without additional information about the manner in which they were collected. Most web-based businesses exploit data about how users interact with resources, such as the log files that record every web search you make, every link you click, and every web page you visit.

Resources that describe, or are associated with other resources are sometimes called [metadata](#). However, when we look more broadly at Organizing Systems, it is often difficult to distinguish between the resource being described and any description of it or associated with it. One challenge is that when descriptions are embedded in resources, as metadata often is—in the title page of a book, the masthead of a newspaper, or the source of web pages—deciding which resources are primary is often arbitrary.

A second challenge is that what serves as metadata for one person or process can function as a primary resource or data for another one. Rather than being an inherent distinction, the difference between primary and associated resources is often just a decision about which resource we are focusing on in some situation. An animal specimen in a natural history museum might be a [primary resource](#) for museum visitors and scientists interested in anatomy,

but information about where the specimen was collected is the [primary resource](#) for scientists interested in ecology or migration.

Organizing Systems can refer to people as resources, and we often use that term to avoid specifying the gender or specific role of an employee or worker, as in the management concept of the “human resources” department in a workplace. A business is defined by its intentional arrangement of human resources, and there is both variety and regularity in these arrangements (see the sidebar, [Business Structures](#) in [“The Structural Perspective”](#)).<sup>2</sup>

Human resources in Organizing Systems can be understood much the same way as inanimate physical or digital resources: they are selected, organized, and managed, and can create value individually or through their interactions with others inside and outside of the system.<sup>3</sup>

2. In addition, groups of people have come together to form “intentional communities” for thousands of years in monasteries, communes, artist colonies, cooperative houses, and religious or ethnic enclaves so they can live with people who share their values and beliefs. A directory of intentional communities organized by type and location is managed by the [Fellowship of Intentional Communities](#).
3. The shift from a manufacturing to an information and services economy in the last few decades has resulted in greater emphasis on intellectual resources represented in skills and knowledge rather than on the natural resources of production materials and physical goods.

However, human beings are uniquely complicated resources, and any Organizing System that uses them must take into account their rights, motivations, and relationships. (See the sidebar, [People as Resources](#).)

The intellectual resources of a firm are embodied in a firm's people, systems, management techniques, history of strategy and design decisions, customer relationships, and intellectual property like patents, copyrights, trademarks, and brands. Some of this knowledge is explicit, tangible, and traceable in the form of documents, databases, organization charts, and policy and procedure manuals. But much of it is tacit: informal and not systematized in tangible form because it is held in the minds and experiences of people; a synonym is "know-how." A more modern term is [Intellectual Capital](#), a concept originated in a 1997 book with that title ([Stewart 1997](#)).

## 4. The Concept of “Collection”

A *collection* is a group of resources that have been selected for some purpose. Similar terms are *set* (mathematics), *aggregation* (data modeling), *dataset* (science and business), and *corpus* (linguistics and literary analysis).

We prefer [collection](#) because it has fewer specialized meanings. *Collection* is typically used to describe personal sets of physical resources (my stamp or record album collection) as well as digital ones (my collection of digital music). We distinguish law libraries from software libraries, knowledge management systems from data warehouses, and personal stamp collections from coin collections primarily because they contain different kinds of resources. Similarly, we distinguish document collections by resource type, contrasting narrative document types like novels and biographies with transactional ones like catalogs and invoices, with hybrid forms like textbooks and encyclopedias in between.

A collection can contain identifiers for resources along with or instead of the resources themselves, which enables a resource to be part of more than one collection, like songs in playlists.

A [collection](#) itself is also a [resource](#). Like other resources, a collection can have description resources associated with it. An *index* is a [description resource](#) that contains information about the locations and frequencies of terms in a document [collection](#) to enable it to be searched efficiently.

Because [collections](#) are an important and frequently used kind of [resource](#), it is important to distinguish them as a separate concept. In particular, the concept of [collection](#) has deep roots in libraries, museums and other institutions that select, assemble, arrange, and

maintain resources. Organizing Systems in these domains can often be described as collections of collections that are variously organized according to resource type, author, creator, or collector of the resources in the collection, or any number of other principles or properties. In business contexts, the use of “collection” to describe a set of resources is much less common, but businesses organize many types of resources, including their employees, suppliers, customers, products, and the tangible and intangible assets used to create the products and run the business. Indeed, a business itself can sometimes be abstractly described as a collection of resources, especially when the resources are software components or services. (See endnote<sup>1</sup>.)

A type of resource and its conventional Organizing System are often the focal point of a discipline. Category labels such as library, museum, zoo, and data repository have core meanings and many associated experiences and practices. Specialized concepts and vocabularies often evolve to describe these. The richness that follows from this complex social and cultural construction makes it difficult to define category boundaries precisely.

Libraries can be defined as institutions that “select, collect, organize, conserve, preserve, and provide access to information on behalf of a community of users.” Many Organizing Systems are described as libraries, although they differ from traditional libraries in important respects. (See the sidebar, [What Is a Library?](#))

What Is a Library?  
Most birds fly, but not all of them do. What

1. [link to footnote]



characteristics are most important to us when we classify something as a bird? What characteristics are most important when we think of something as a library?

We might treat *circulation*, borrowing and returning the same item, as one of the interactions with resources that defines a library. In that case, an institution that lends items in its collection with the hope that the borrowers return something else that is better hardly seems like a library. But if the resources are the seeds of heirloom plants and the borrowers are expected to return seeds from the plants they grew from the borrowed seeds, perhaps “seed library” is an apt name for this novel Organizing System. Similarly, even though the resources in its collection are encyclopedia articles rather than living species, the Wikipedia open-source encyclopedia resembles the Seed Library by encouraging its users to “return” articles that are improvements of the current ones.

The photo-sharing website Flickr functions for most of its users as a personal photo archiving site. Flickr’s billions of user-uploaded photos and the choice of many users to share them publicly transform it into a searchable shared collection, and many people also think of Flickr as a photo library. But Flickr lacks the authoritative description and standard classification that typify a library.

A similar categorization challenge arises with the Google Books digitization project.<sup>2</sup>

2. In 2004, Google began digitizing millions of books from several major research libraries with the goal of making them available through its search engine ([Brin 2009](#)). But many millions of these books are still in copyright, and in 2005 Google was sued for copyright infringement by several publishers and an author's organization. In 2011 a US District Court judge rejected the proposed settlement the parties had negotiated in 2008 because many others objected to it, including the US Justice Department, several foreign governments, and numerous individuals ([Samuelson 2011](#)).

The major reason for the rejection was that the settlement was a “bridge too far” that went beyond the claims made against Google to address issues that were not in litigation. In particular, the judge objected to the treatment of the so-called “orphan works” that were still under copyright but out of print because money they generated went to the parties in the settlement and not to the rights holders who could not be located (why the books are “orphans”) or to defray the costs of subscriptions to the digital book collection. The judge also was concerned that the settlement did not

We can always create new categories by stretching the conventional definitions of “library” or other familiar Organizing Systems and adding modifiers, as when Flickr is described as a web-based photo-sharing library. But whenever we define an Organizing System with respect to a familiar category, the typical or mainstream instances and characteristics of that category that are deeply embedded in

adequately address the concerns of academic authors—who wrote most of the books scanned from research libraries—who might prefer to make their books freely available rather than seek to maximize profits from them. Other concerns were that the settlement would have entrenched Google’s monopoly in the search market and that there were inadequate controls for protecting the privacy of readers.

Google’s plan would have dramatically increased access to out of print books, and the rejection of the proposed settlement has heightened calls for an open public digital library ([Darnton 2011](#)). A good start toward such a library was the digital copies that the research libraries received in return for giving Google books to scan, which were collected and organized by the Hathi Trust (See the sidebar, [The Hathi Trust Digital Library](#)). In 2010, the Alfred P. Sloan Foundation provided funding to launch the *Digital Public Library of America*(DPLA): <http://dp.la/>. This non-proprietary goal might induce the US Congress and other governments to pass legislation that fixes the copyright problems for orphan works.

language and culture are reinforced, and those that are atypical are marginalized. In the Flickr case, this means we suggest features that are not there (like authoritative classification) or omit the features that are distinctive (like tagging by users).

More generally, a categorical view of Organizing Systems makes it matter greatly which category is used to anchor definitions or comparisons. The Google Books project makes out-of-print and scholarly works vastly more accessible, but when Google co-founder Sergei Brin described it as “a library to last forever” it upset many people with a more traditional sense of what the library category implies. We can readily identify design choices in Google Books that are more characteristic of the Organizing Systems in business domains, and the project might have been perceived more favorably had it been described as an online bookstore that offered many beneficial services for free.

## 5. The Concept of “Intentional Arrangement”

*Intentional arrangement* emphasizes explicit or implicit acts of organization by people, or by computational processes acting as proxies for, or as implementations of, human intentionality. Intentional arrangement is easiest to see in Organizing Systems created by individual people who can make all the necessary decisions about organizing their own resources. It is also easy to see in Organizing Systems created by institutions like libraries, museums, businesses, and governments where the responsibility and authority to organize is centralized and explicit in policies, laws, or regulations.

However, top-down intentionality is not always necessary to create an Organizing system. Organization can emerge over time via collective behavior in situations without central control when decisions made by individuals, each acting intentionally, create traces, records, or other information that accumulates over time. Organizing systems that use bottom-up rather than top-down mechanisms are sometimes called self-organizing, because they emerge from the aggregated interactions of actors with resources or with each other. *Self-organizing systems* can change their internal structure or their function in response to feedback or changed circumstances.

This definition is broad enough to include business and biological ecosystems, traffic patterns, and open-source software projects. Another good example of emergent organization involves path systems, where people (as well as ants and other animals) can follow and thereby reinforce the paths taken by their predecessors. When highly orderly and optimal arrangements emerge from local interactions among ants, bees, birds, fish, and other animal species,

it is often called “swarm intelligence.” When this happens with human ratings for news stories, YouTube videos, restaurants, and other types of digital and physical resources we call it “crowdsourcing.” What the animal and human situations have in common is that information is being communicated between individuals. Sometimes this communication is direct, as when Amazon shows you the average rating for a book or what books have been bought by people like you. At other times the communication is indirect, achieved when the agents modify their environment (as they do when they create paths) and others can respond to these modifications. Adam Smith’s “invisible hand” is another example where individuals collectively generate an outcome they did not directly intend but that arose from their separate self-interested actions as they respond to price signals in the marketplace. Likewise, even though there is no top-down organization, the web as a whole, with its more than a trillion unique pages, is a self-

organizing system that at its core follows clear organizing

principles.<sup>123</sup>

1. Self-organizing is also used to describe phenomena like climate, neural networks, and phase transitions and equilibrium states in physics and chemistry. But when systems involve collections of *inanimate* resources that are very large and open, with complex interactions among the resources, it seems less sensible to attribute intentional arrangement to the outcomes. The resource arrangements that emerge cannot always be interpreted as the result of intentional or deterministic principles and instead are more often described in probabilistic or statistical terms. And even though it involves animate resources, Charles Darwin's "natural selection" in evolutionary biology is a self-organizing mechanism where intentionality is hard to pinpoint or absent entirely.

The rules governing these local interactions can be simple and yet produce highly complex structures. For example, in flocks of birds or schools of fish the rules are: (1) follow things like you, (2) do not bump into each other, but stay close, and (3) move in the same direction as the rest of the group. With just these three rules computer models can create complex three-dimensional arrangements that can make abrupt changes in shape and density while moving rapidly, just as live things do. ([Friederici 2009](#))



The term “Crowdsourcing” was invented by Jeff Howe in a June 2006 article in Wired magazine, and the concept was developed further in a book published two years later ([Howe 2006, 2008](#)). “Folksonomy” was coined by Thomas Van der Wal at about the same time in 2004; see <http://vanderwal.net/folksonomy.html> and ([Trant 2009](#)).

([Goldstone and Gureckis 2009](#)) present a cognitive science perspective on collective behavior, analyze important themes and controversies, and suggest areas for future research. ([Moussaid et al. 2009](#)) analyze self-organizing phenomena in animal swarms and human crowds in terms of information exchange among individuals.

Self-organizing behaviors in ants, bees, bats, cuckoos, fireflies and other animals have been analyzed to identify heuristics that can be applied to difficult optimization problems in network design, cryptography, and other domains where deterministic algorithms are infeasible. ([Yang 2010](#))

([Smith 1776](#))

2. ([Banzhaf 2009](#)).
3. The concept of a web page is imprecise because many web pages, especially home pages designed as navigation gateways to an organized collection of pages, are

### The Web as an Organizing System

Today's web barely resembles the system for distributing scientific and technical reports it was designed to be when physicist and computer scientist Tim Berners-Lee devised it in 1990 at the European Organization for Nuclear Research (CERN) lab near Geneva. However, as an [Organizing System](#) the web still follows the principles that Berners-Lee defined at its creation. These include standard data formats and interaction protocols; no need for centralized control of page creation or linking; remote access over the network from anywhere; and the ability to run on a large variety of computers and operating systems. This architecture makes the web open and extensible, but gives it no built-in mechanisms for authority or trust.<sup>4</sup>

constructed from heterogeneous blocks of content that could have been organized as separate pages.

4. The “plain web” ([Wilde 2008a](#)), whose evolution is managed by the *World Wide Web Consortium* (W3C), is rigorously standardized, but unfortunately the larger ecosystem of technologies and formats in which the web exists is becoming less so. Web-based Organizing Systems often contain proprietary media formats and players (like Flash) or are implemented as closed environments that are intentionally isolated from the

Because the web works without any central authority or authorship control, any person or organization can add to it. As a result, even though the web as a whole does not exhibit the centralized intentional arrangement of resources that characterizes many Organizing Systems, we can view it as consisting of millions of Organizing Systems that each embody a separate intentional arrangement of web pages. In addition, we most often interact with the web indirectly by using a search engine, which meets the definition of [Organizing System](#) because its indexing and retrieval algorithms are principled.

A great many Organizing Systems are implemented as collections of web pages. Some of these collections are created on the web as new pages, some are created by transforming existing collections of resources, and some combine new and existing resources.

The requirement for intentional arrangement excludes naturally occurring patterns created by physical or geological processes from being thought of as Organizing Systems. There is information in the piles of debris left after a tornado or tsunami and the strata of the Grand Canyon. But they are not Organizing Systems because the patterns of arrangement were created by deterministic natural forces rather than by agents following one or more organizing principles. On the other hand, collections of geological data like the measurements of chemical composition from different strata and

rest of the web (like Facebook or Apple's iTunes and other smart phone "app stores").

locations in the Grand Canyon **are** Organizing Systems. Decisions about what to measure, how to combine and analyze the measurements, and any theories that are tested or created, reflect intentional arrangement of the data by the geologist.

Other patterns of resource arrangements are illusions or perceptions that require a particular vantage point. The best examples are patterns of stars as they appear to an observer on Earth. The three precisely aligned stars, often described as “Orion’s belt,” are hundreds of light years from Earth, and also from each other. The perceived arrangement of the stars is undeniable, but the stars are not aligned in the universe. Astronomical constellations like Orion are [\*intentional arrangements\*](#) imposed on our perceived locations of the stars, and these perceived arrangements and the explanations for them that constellations provide, form an Organizing System that is deeply embedded in human culture and in the practice of celestial navigation over the seas.

Not an Intentional Arrangement



The composition and arrangement of the rock layers (“strata”) in the Grand Canyon in the Southwest United States have been studied extensively by geologists. The composition of rock suggests the environment in which it was formed, and the absolute and relative arrangement of the rock layers reveals the timing of important geological events.

([Photo by B. Rosen](#), Creative Commons CC BY-ND 2.0 license.)

Taken together, the intentional arrangements of resources in an Organizing System are the result of decisions about what is organized, why it is organized, how much it is organized, when it is organized, and how or by whom it is organized (each of these will be discussed in greater detail in [Design Decisions in Organizing Systems](#)). An Organizing System is defined by the composite impact of the choices made on these design dimensions. Because these questions are interrelated their answers come together in an integrated way to define an Organizing System.

## 6. The Concept of “Organizing Principle”

The arrangements of resources in an Organizing System follow or embody one or more organizing principles that enable the Organizing System to achieve its purposes. *Organizing principles* are directives for the design or arrangement of a [collection](#) of resources that are ideally expressed in a way that does not assume any particular implementation or realization. We call this design philosophy “Architectural Thinking” ([“Architectural Thinking”](#).)

Organizing Spices By Cuisine



An alternative to organizing spices alphabetically is to organize them according to cuisines or flavor profiles,

which can be defined in terms of ingredients and spices that tend to be used together. Patricia Glushko organizes her spices into three groups: Indian (includes cayenne pepper, coriander, cumin, turmeric), Mediterranean / Middle Eastern (includes basil, dill, oregano, paprika, thyme), and seeds. Each group of spices is in a separate large container, which makes it convenient when cooking.

(Photo by R. Glushko.)

When we organize a bookshelf, home office, kitchen, or the MP3 files on our music player, the resources themselves might be new and modern but many of the principles that govern their organization are those that have influenced the design of Organizing Systems for thousands of years. For example, we organize many collections of resources using the properties that are easiest to perceive, or whose values vary the most among the items in the collection, because these principles make it easy to locate a particular resource. We also group together resources that we often use together, we make resources that we use often more accessible than those we use infrequently, and we put rare or unique resources where we can protect them. Very general and abstract organizing principles are sometimes called design heuristics (e.g., “make things easier to find”). More specific and commonly used organizing principles include *alphabetical ordering* (arranging resources according to their names) and *chronological ordering* (arranging resources according to the date of their creation or other important event in the lifetime of the resource). Some organizing principles sort resources into pre-defined categories and other organizing principles rely on novel combinations of resource properties to create new categories.

Because this book was motivated by the goal of broadening the study of information organization beyond its roots in library and information science, it emphasizes organizing principles with a specific functional purpose like identifying, selecting, retrieving, or preserving resources. However, for thousands of years people have systematically collected things, information about those things, and observations of all kinds, organizing them in an effort to understand how their world works; the Babylonians created inventories and star charts; ancient Egyptians tracked the annual Nile floods; and, Mesoamericans created astronomical calendars. The term [sensemaking](#) is often used to describe this generic and less specific purpose of organizing to derive meaning from experience by fitting new events or observations into what they already know.<sup>1</sup>

Expressing organizing principles in a way that separates design and implementation aligns well with the three-tier architecture familiar to software architects and designers: user interface (implementation of interactions), business logic (intentional arrangement), and data (resources). (See the sidebar, [The Three Tiers of Organizing Systems](#).)

The logical separation between organizing principles and their implementation is easy to see with digital resources. In a digital library it does not matter to a user if the resources are stored locally or retrieved over a network. The essence of a library Organizing System emerges from the resources that it organizes and the interactions with the resources that it enables. Users typically care a lot about the interactions they can perform, like the kinds of searching and sorting allowed by the online library catalog. How the resources and interactions are implemented are typically of little concern.

1. ([Weick et al , 2005 p. 410](#)).



The separation of organizing principles and their implementation is harder to recognize in an Organizing System that only contains physical resources, such as your kitchen or clothes closet, where you appear to have unmediated interactions with resources rather than accessing them through some kind of user interface or “presentation tier” that supports the principles specified in the “middle tier” and realized in the “storage tier.” As a result, people can easily get distracted by presentation-tier concerns. Too often we waste time color-coding file folders and putting labels on storage containers, when it would have better to think more carefully about the logical organization of the folder and container contents. It does not help to use colors and labels to make the logical organization more salient if that is not well designed first.

One place where you can easily appreciate these different tiers for physical resources is in the organization of spices in a kitchen. Different kitchens might all embody an [alphabetic order](#) organizing principle for arranging a collection of spices, but the exact locations and arrangement of the spices in any particular kitchen depends on the configuration of shelves and drawers, whether a spice rack or rotating tray is used, and other storage-tier considerations. Similarly, spices could be logically organized by cuisine, with Indian spices separated from Mexican spices, but this organizing principle does not imply anything about where they can be found in the kitchen.

#### The Three Tiers of Organizing Systems

Software architects and designers agree that it is desirable to build applications that separate the storage of data, the business logic or functions that use the data, and the user interface or presentation components through which users or other applications interact with

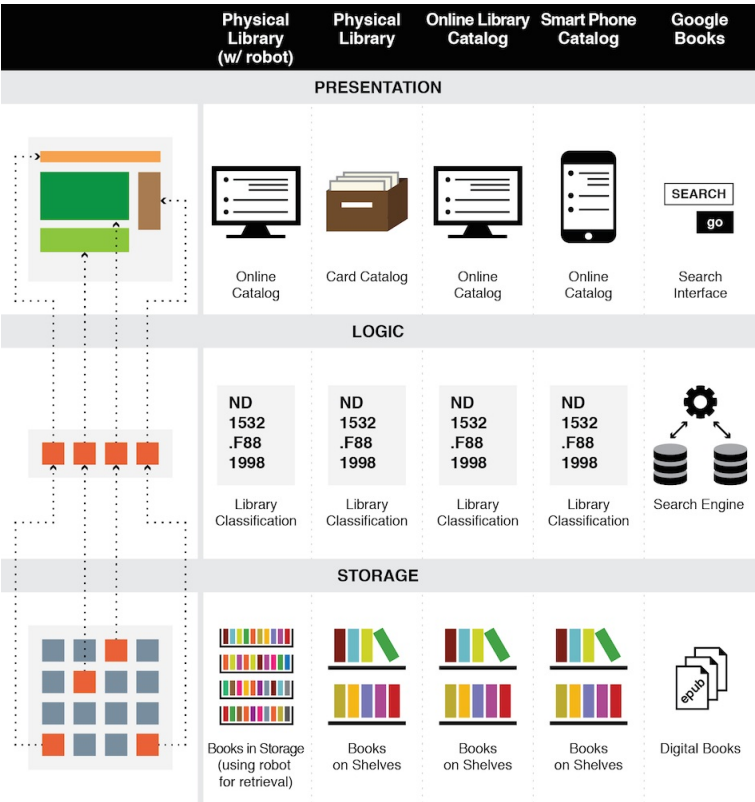
the data. This modular architecture allows each of the three tiers to be upgraded or reimplemented independently to satisfy changed requirements or to take advantage of new technologies. An analogous distinction is that between an algorithm as a logical description of a method for solving a computational problem and its implementation in a particular programming language like Java or Python.

These architectural distinctions are equally important to librarians and information scientists. Our new way of looking at Organizing Systems emphasizes the importance of identifying the desired interactions with resources, determining which organizing principles can enable the interactions, and then deciding how to store and manage the resources according to those principles. Applying architectural thinking to Organizing Systems makes it easier to compare and contrast existing ones and design new ones. Separating the organizing principles in the “middle tier” from their implications in the “data” and “presentation” tiers often makes it possible to implement the same logical Organizing System in different environments that support the same or equivalent interactions with the resources. For example, a new requirement to support searching through a library catalog on a smart phone would only affect the presentation tier.

[Figure: Presentation, Logic and Storage Tiers.](#) illustrates the separation of the presentation, logic, and storage tiers for four different types of library Organizing Systems and for Google Books. No two of them are the same in every tier. Note how a library that

uses inventory robots to manage the storage of books does not reveal this in its higher tiers. (See the sidebar, [Library Robot](#).)

Presentation, Logic and Storage Tiers



It is highly desirable when the design and implementation of an Organizing System separates the storage of the resources from the logic of their arrangement and the methods for interacting with them. This three-tier architecture is familiar to designers of computerized Organizing Systems but it is also useful to think about

Organizing Systems in this way even when it involves physical resources.

Because tangible things can only be in one place at a time, many Organizing Systems, like those in the modern library with online catalogs and physical collections, resolve this constraint by creating digital proxies or surrogates to organize their tangible resources, or create parallel digital resources (e.g., digitized books).<sup>2</sup>

The implications for arranging, finding, using and reusing resources in any Organizing System directly reflect the mix of these two embodiments of information; in this way we can think of the modern library as a digital Organizing System that primarily relies on digital resources to organize a mixture of physical and digital ones.

The Organizing System for a small collection can sometimes use only the minimal or default organizing principle of *colocation*—putting all the resources in the same location: in the same container, on the same shelf, or in the same email in-box. If you do not cook much and have only a small number of spices in your kitchen, you do not need to alphabetize them because it is easy to find the one you want.<sup>3</sup>

2. Instead of thinking of a digital book as a “parallel resource” to a printed book, we could consider both of them as alternate representations of the same abstract resource that are linked together by an “alternative” relationship, just as we can use the `HTML ALT` tag to associate text with an image so its content and function can be understood by text-only readers.

3. For collections of non-trivial size the choice of searching

## Separation Of Organizing Principle From Implementation



or sorting algorithm in computer programs is a critical design decision because they differ greatly in the time they take to complete and the storage space they require. For example, if the collection is arranged in an unorganized or random manner (as a “pile”) and every resource must be examined, the time to find a particular item increases linearly with the collection size. If the collection is maintained in an ordered manner, a binary search algorithm can locate any item in a time proportional to the logarithm of the number of items. Analysis of algorithms is a fundamental topic in computer science; a popular textbook is [Introduction to Algorithms](#) by [\(Cormen et al. 2009\)](#).

Whether spices are organized alphabetically by their names, by cuisines, by season, by frequency of use, or any other principle, this decision is logically distinct from the physical arrangement of the spices. There are many types of spice racks, shelves, circular “lazy susans,” and other devices designed for arranging spices.

(Photo collage created by R. Glushko from various web catalogs.)

Some organization emerges implicitly through a *frequency of use* principle. In your kitchen or clothes closet, the resources you use most often migrate to the front because that is the easiest place to return them after using them. But as a collection grows in size, the time to arrange, locate, and retrieve a particular resource becomes more important. The collection must be explicitly organized to make these interactions efficient, and the organization must be preserved after the interaction takes place; i.e., resources are put back in the place they were found. As a result, most Organizing Systems employ organizing principles that make use of properties of the resources being organized (e.g., name, color, shape, date of creation, semantic or biological category), and multiple properties are often used simultaneously. For example, in your kitchen you might arrange your cooking pots and pans by size and shape so you can nest them and store them compactly, but you might also arrange things by cuisine or style and separate your grilling equipment from the wok and other items you use for making Chinese food.

Unlike those for physical resources, the most useful organizing properties for information resources are those that reflect their

content and meaning, and these are not directly apparent when you look at a book, document, or collection of data. Significant intellectual effort or statistical computation is necessary to reveal these properties when assigning subject terms, creating an index, or using them as input features for machine learning and data analysis programs.

The most effective Organizing Systems for information resources often are based on statistical properties that emerge from analyzing the collection as a whole. For example, the relevance of documents to a search query is higher when they contain a higher than average frequency of the query terms compared to other documents in the collection, or when they are linked to relevant documents. Likewise, algorithms for classifying email messages continuously recalculate the probability that words like “beneficiary” or “Viagra” indicate whether a message is “spam” or “not spam” in the collection of messages processed.

## 7. The Concept of “Agent”

Many disciplines have specialized job titles to distinguish among the people who organize resources (for example: cataloger, archivist, indexer, curator, collections manager...).<sup>1</sup>

We use the more general word, *agent*, for any entity capable of autonomous and intentional organizing effort, because it treats organizing work done by people and organizing work done by computers as having common goals, despite obvious differences in methods.

We can analyze agents in Organizing Systems to understand how human and computational efforts to arrange resources complement and substitute for each other. We can determine the economic, social, and technological contexts in which each type of agent can best be employed. We can determine how the Organizing System allocates effort and costs among its creators, users, maintainers and other stakeholders.

A group of people can be an organizing *agent*, as when a group of people come together in a service club or standards body technical committee in which the members of the group subordinate their own individual agency to achieve a collective good.

We also use the term *agent* when we discuss interactions with

1. For precise distinctions, see the US Department of Labor, Bureau of Labor Statistics occupational outlook handbooks at <http://www.bls.gov/oco/ocos065.htm> and <http://www.bls.gov/oco/ocos068.htm> and <http://www.michellemach.com/jobtitles/realjobs.html>.



Organizing Systems. The entities that most typically access the contents of libraries, museums, or other collections of physical resources are human agents—that is, people. In other Organizing Systems, such as business information systems or data repositories, interactions with resources are carried out by computational processes, robotic devices, or other entities that act autonomously on behalf of a person or group.

In some Organizing Systems, the resources themselves are capable of initiating interactions with other resources or with external agents. This is most obvious with human or other living resources, where a critical part of the design of any Organizing System with them is determining what kinds of interactions they should be encouraged or allowed to initiate. We will return to this issue after we discuss the design of interactions with ordinary resources that are passive, the situation in most Organizing Systems that involve physical resources.

Other resources that can initiate interactions are resources augmented with sensory, computational or communication capabilities that enable them to obtain information from their environment and then do something useful with it. You are probably familiar with RFID tags, which enable the precise identification and location of physical resources as they move through supply chains and stores, and with “smart” devices like Nest thermostats that learn how to program themselves.

## 8. The Concept of “Interactions”

An *interaction* is an action, function, service, or capability that makes use of the resources in a collection or the collection as a whole. The interaction of *access* is fundamental in any collection of resources, but many Organizing Systems provide additional functions to make access more efficient and to support additional interactions with the accessed resources. For example, libraries and similar Organizing Systems implement catalogs to enable interactions for *finding* a known resource, *identifying* any resource in the collection, and discriminating or *selecting* among similar resources.<sup>1</sup>

Some of the interactions with resources in an Organizing System are inherently determined by the characteristics of the resource. Because many museum resources are unique or extremely valuable, visitors are allowed to view them but cannot borrow them, in contrast with most of the resources in libraries. A library might have multiple printed copies of [Moby Dick](#) but can never lend more

1. The four objectives listed in this paragraph as those proposed in 1997 by the *International Federation of Library Associations and Institutions*(IFLA). The first statement of the objectives for a bibliographic system was made by ([Cutter 1876](#)), which ([Svenonius 2000](#)) says it is likely the most cited text in the bibliographic literature. Cutter called his three objectives “finding,” “co-locating,” and “choice.”

of them than it possesses. After a printed book is checked out from the library, there are many types of interactions that might take place—reading, translating, summarizing, annotating, and so on—but these are not directly supported by the library Organizing System and are invisible to it.

For works not in the public domain, copyright law gives the copyright holder the right to prevent some uses, but at the same time “fair use” and similar copyright doctrines enable certain limited uses even for copyrighted works.<sup>2</sup>

Digital resources enable a greater range of interactions than physical ones. Any number of people or processes can request a weather forecast from a web-based weather service because the forecast is not used up by the request and the marginal cost of allowing another access is nearly zero. Furthermore, with digital resources many new kinds of interactions can be enabled through application software, web services, or *application program interfaces* (APIs) in the Organizing System. In particular, translation, summarization, annotation, and keyword suggestion are highly useful services that are commonly supported by web search engines and other web applications. Similarly, an Organizing System with digital resources can implement a “keep everything up to date” interaction that automatically pushes current content to your browser.

2. Copyright law, license or contract agreements, terms of use and so on that shape interactions with resources are part of the Organizing System, but compliance with them might not be directly implemented as part of the system. With digital resources, digital rights management (DRM), passwords, and other security mechanisms can be built into the Organizing System to enforce compliance.

But just as technology can enable interactions, it can prevent or constrain them. If your collection of digital resources (ebooks or music, for example) is not stored on your own computer or device, a continuous Internet connection is a requirement for access. In addition, access control policies and *digital rights management* (DRM) technology can limit the devices that can access the collection and prevent copying, annotation and other actions that might otherwise be enabled by the fair use doctrine.

Interaction design is especially crucial for managing resources that have the capability to initiate interactions with each other or with external agents. Consider the vast differences in how workers behave in businesses organized according to principles of scientific management and those that embody the Kaizen principles of continuous improvement. In the former, work is highly standardized and bureaucratic, giving workers little autonomy. In the latter, work is also standardized, but workers are motivated to analyze and improve work processes whenever possible, and they are given great discretion in how to do that.<sup>3</sup>

3. Frederick Taylor developed “scientific management” to improve industrial efficiency and conducted detailed time and motion studies to devise what he thought were optimal ways to perform work tasks ([Taylor 1914](#)). The Kaizen principles of continuous improvement were introduced to Western audiences by Imai Masaaki and by numerous books about their application in the Toyota production system ([Masaaki 1986](#)).

Scientific management views a business as a machine, while Kaizen principles treat it as a brain that learns. These metaphors for business organization are among

Just as with organizing principles, it is useful to think of interactions in an abstract or logical way that does not assume an implementation because it can encourage innovative designs for Organizing Systems.

those described by [\(Morgan 1997\)](#) in a classic business textbook. Other metaphors discussed include organisms, cultures, political systems, and psychic prisons.

## 9. The Concept of “Interaction Resource”

Interactions with physical resources sometimes leave traces or other evidence. Many of these traces are unintentional, like fingerprints, a coffee cup stain on a newspaper, or the erosion on a shortcut path across a lawn. Fans of Sherlock Holmes and CSI know that clever forensic investigators can use these residues of interactions to identify or vindicate suspects. Other interaction traces are intentional, like a student’s yellow highlighting or notes in a textbook or spray-painted graffiti on a building. But not every interaction leaves a trace, traces fade over time, and different traces associated with the same resource lack consistency. This means that most traces are not of much use.

However, when Organizing Systems contain digital resources, or physical resources that have sensing, recording, or communication capabilities, interaction traces can be made predictable, persistent, and consistent. Each record of a user choice in accessing, browsing, buying, highlighting, linking, and other interactions then becomes an “interaction resource” that can be analyzed to reorganize the resource collection or otherwise influence subsequent interactions with the primary resources.

Interaction resources are often essential pieces of information that make Organizing Systems function. Most human toll-takers have been replaced by smart “toll tags” that broadcast their identity when the car they are in passes a radio receiver at a tolling location. Each interaction resource created identifies an account and credit card with which to pay the toll; taken together, the collection of these interaction resources can be used as the primary resources in other Organizing Systems that manage traffic congestion, or that support road design. Similarly, interaction resources created by search

engines can be used to adjust the order of search hits, select ads, or personalize the content of web pages.

# 10. Organizing This Book

Devising concepts, methods, and technologies for describing and organizing resources have been essential human activities for millennia, evolving both in response to human needs and to enable new ones. Organizing Systems enabled the development of civilization, from agriculture and commerce to government and warfare. Today Organizing Systems are embedded in every domain of purposeful activity, including research, education, law, medicine, business, science, institutional memory, sociocultural memory, governance, public accountability, as well as in the ordinary acts of daily living.

With the World Wide Web and ubiquitous digital information, along with effectively unlimited processing, storage and communication capability, millions of people create and browse websites, blog, tag, tweet, and upload and download content of all media types without thinking “I am organizing now” or “I am retrieving now.” Writing a book used to mean a long period of isolated work by an author followed by the publishing of a completed artifact, but today some books are continuously and iteratively written and published through the online interactions of authors and readers. When people use their smart phones to search the web or run applications, location information transmitted from their phone is used to filter and reorganize the information they retrieve. Arranging results to make them fit the user’s location is a kind of computational curation, but because it takes place quickly and automatically we hardly notice it.

Likewise, almost every application that once seemed predominantly about information retrieval is now increasingly combined with activities and functions that most would consider to be information organization. Google, Microsoft, and other search engine operators have deployed millions of computers to analyze billions of web



pages and millions of books and documents to enable the almost instantaneous retrieval of published or archival information. However, these firms increasingly augment this retrieval capability with information services that organize information in close to real-time. Further, the selection and presentation of search results, advertisements, and other information can be tailored for the person searching for information using his implicit or explicit preferences, location, or other [contextual information](#).

Taken together, these innovations in technology and its application mean that the distinction between *information organization* and *information retrieval* that is often manifested in academic disciplines and curricula is much less important than it once was.

This book has few sharp divisions between *information organization* (IO) and *information retrieval* (IR) topics. Instead, it explains the key concepts and challenges in the design and deployment of Organizing Systems in a way that continuously emphasizes the relationships and tradeoffs between IO and IR. The concept of the Organizing System highlights the design dimensions and decisions that collectively determine the extent and nature of resource organization and the capabilities of the processes that compare, combine, transform and interact with the organized resources.

## Navigating The Discipline of Organizing

### [Design Decisions in Organizing Systems](#)

This chapter introduces six broad design questions or dimensions whose intertwined answers define an Organizing System: What, why, how much, when, how, and where. This framework for describing and comparing Organizing Systems overcomes the biases and conservatism built into familiar categories like libraries and museums while enabling us to

describe them as design patterns. We can then use these patterns to support inter-disciplinary work that cuts across categories and applies knowledge about familiar domains to unfamiliar ones.

### **Activities in Organizing Systems**

Developing a view that brings together how we organize as individuals with how libraries, museums, governments, research institutions, and businesses create Organizing Systems requires that we generalize the organizing concepts and methods from these different domains. [Activities in Organizing Systems](#) surveys a wide variety of Organizing Systems and describes four activities or functions shared by all of them: selecting resources, organizing resources, designing resource-based interactions and services, and maintaining resources over time.

### **Resources in Organizing Systems**

The design of an Organizing System is strongly shaped by what is being organized, the first of the six design decisions we introduced earlier in [“What Is Being Organized?”](#). To enable a broad perspective on this fundamental issue we use [resource](#) to refer to anything being organized, an abstraction that we can apply to physical things, digital things, information about either of them, or web-based services or objects. [Resources in Organizing Systems](#) discusses the challenges and methods for identifying the resources in an Organizing System in great detail and emphasizes how these decisions reflect the goals and interactions that must be supported—the “why” design decisions introduced in [“Why Is It Being Organized?”](#).

### **Resource Description and Metadata**

The principles by which resources are organized and the kinds of services and interactions that can be supported for them

largely depend on the nature and explicitness of the resource descriptions. This “how much description” design question was introduced in [“How Much Is It Being Organized?”](#); [Resource Description and Metadata](#) presents a systematic process for creating effective descriptions and analyzes how this general approach can be adapted for different types of Organizing Systems.

### **[Describing Relationships and Structures](#)**

An important aspect of organizing a collection of resources is describing the relationships between them. [Describing Relationships and Structures](#) introduces the specialized vocabulary used to describe semantic relationships between resources and between the concepts and words used in resource descriptions. It also discusses the structural relationships within multipart resources and between resources, like those expressed as citations or hypertext links.

### **[Categorization: Describing Resource Classes and Types](#)**

Groups or sets of resources with similar or identical descriptions can be treated as equivalent, making them members of an *equivalence class* or *category*. Identifying and using categories are essential human activities that take place automatically for perceptual categories like “red things” or “round things.” Categorization is deeply ingrained in language and culture, and we use linguistic and cultural categories without realizing it, but categorization can also be a deeply analytic and cognitive process. [Categorization: Describing Resource Classes and Types](#) reviews theories of categorization from the point of view of how categories are created and used in Organizing Systems.

### **[Classification: Assigning Resources to Categories](#)**

The terms *categorization* and *classification* are often used

interchangeably but they are not the same. Classification is applied categorization—the assignment of resources to a system of categories, called classes, using a predetermined set of principles. [Classification: Assigning Resources to Categories](#) discusses the broad range of how classifications are used in Organizing Systems. These include enumerative classification, faceted classification, activity-based classification, and computational classification. Because classification and standardization are closely related, the chapter also analyzes standards and standards-making as they apply to Organizing Systems.

### [The Forms of Resource Descriptions](#)

[The Forms of Resource Descriptions](#) complements the conceptual and methodological perspective on the creation of resource descriptions with an implementation perspective. [The Forms of Resource Descriptions](#) reviews a range of metamodels for structuring descriptions, with particular emphasis on [XML](#), [JSON](#), and [RDF](#). It concludes by comparing and contrasting three “worlds of description”—document processing, the web, and the [Semantic Web](#)—where each of these three metamodels is most appropriate.

### [Interactions with Resources](#)

When Organizing Systems overlap, intersect, or are combined (temporarily or permanently), differences in resource descriptions can make it difficult or impossible to locate resources, access them, or otherwise impair their use. [Interactions with Resources](#) reviews some of the great variety of concepts and techniques that different domains use when interacting with resources in Organizing Systems—integration, interoperability, data mapping, crosswalks, mash-ups, and so on. Interactions are characterized by the layers of resource properties they use: instance, collection-based, derived, or

properties combined from different resources. [Interactions with Resources](#) extends the idea of an information organization—information retrieval continuum, and describes information retrieval interactions (and others) in terms of information organization (i.e., resource description) requirements.

### [The Organizing System Roadmap](#)

[The Organizing System Roadmap](#) complements the descriptive perspective of chapters 2-10 with a more prescriptive one that analyzes the design choices and tradeoffs that must be made in different phases in an Organizing System's life cycle. System life cycle models exhibit great variety, but we use a generic four-phase model that distinguishes a domain identification and scoping phase, a requirements phase, a design and implementation phase, and an operational phase.

### [Case Studies](#)

In [Case Studies](#) we use the model described in [The Organizing System Roadmap](#) to guide the analysis of studies that span the range of Organizing Systems, and make reference to the principles, guidelines, vocabulary, and models discussed in the preceding chapters.

PART II

# DESIGN DECISIONS IN ORGANIZING SYSTEMS

**Robert J. Glushko**



## II. Introduction (II)

A set of resources is transformed by an organizing system when the resources are described or arranged to enable interactions with them. Explicitly or by default, this requires many interdependent decisions about the identities of resources; their names, descriptions and other properties; the classes, relations, structures and collections in which they participate; and the people or technologies interacting with them.

One important contribution of the idea of the organizing system is that it moves beyond the debate about the definitions of “things,” “documents,” and “information,” with the unifying concept of [“resource”](#) while acknowledging that “what is being organized” is just one of the questions or dimensions that need to be considered. These decisions are deeply intertwined, but it is easier to introduce them as if they were independent.

We introduce six groups of design questions, itemizing the most important dimensions in each group:

- **What is being organized?** What is the scope and scale of the domain? What is the mixture of physical things, digital things, and information about things in the organizing system? Is the organizing system being designed to create a new resource collection, catalog an existing and closed resource collection, or manage a collection in which resources are continually added or deleted? Are the resources unique, or are they interchangeable members of a category? Do they follow a predictable “life cycle” with a “useful life”? Does the organizing system use the interaction resources created through its use, or are these interaction resources extracted and aggregated for use by another organizing system? ([“What Is Being Organized?”](#))
- **Why is it being organized?** What interactions or services will



be supported, and for whom? Are the uses and users known or unknown? Are the users primarily people or computational processes? Does the organizing system need to satisfy personal, social, or institutional goals? ([“Why Is It Being Organized?”](#))

- **How much is it being organized?** What is the extent, granularity, or explicitness of description, classification, or relational structure being imposed? What organizing principles guide the organization? Are all resources organized to the same degree, or is the organization sparse and non-uniform? ([“How Much Is It Being Organized?”](#))
- **When is it being organized?** Is the organization imposed on resources when they are created, when they become part of the collection, when interactions occur with them, just in case, just in time, all the time? Is any of this organizing mandated by law or shaped by industry practices or cultural tradition? ([“When Is It Being Organized?”](#))
- **How or by whom, or by what computational processes, is it being organized?** Is the organization being performed by individuals, by informal groups, by formal groups, by professionals, by automated methods? Are the organizers also the users? Are there rules or roles that govern the organizing activities of different individuals or groups? ([“How \(or by Whom\) Is It Organized?”](#))
- **Where is it being organized?** Is the resource location constrained by design or by regulation? Are the resources positioned in a static location? Are the resources in transit or in motion? Does their location depend on other parameters, such as time? ([“Where is it being Organized?”](#))

How well these decisions coalesce in an organizing system depends on the requirements and goals of its human and computational users, and on understanding the constraints and tradeoffs that any

set of requirements and goals impose. How and when these constraints and tradeoffs are handled can depend on the legal, business, and technological contexts in which the organizing system is designed and deployed; on the relationship between the designers and users of the organizing system (who may be the same people or different ones); on the economic or emotional or societal purpose of the organizing system; and on numerous other design, deployment, and use factors.

Classifying organizing systems according to the kind of resources they contain is the most obvious and traditional approach. We can also classify organizing systems by their dominant purposes, by their intended user community, or other ways. No single fixed set of categories is sufficient by itself to capture the commonalities and contrasts between organizing systems.

We can augment the categorical view of organizing systems by thinking of them as existing in a multi-faceted or multi-dimensional design space in which we can consider many types of collections at the same time.

This [framework](#) for describing and comparing organizing systems overcomes some of the biases and conservatism built into familiar categories like libraries, museums, and archives, while enabling us to describe them as design patterns that embody characteristic configurations of design choices. We can then use these patterns to support inter-disciplinary work that cuts across categories and applies knowledge about familiar domains to unfamiliar ones. A dimensional perspective makes it easier to translate between category- and discipline-specific vocabularies so that people from different disciplines can have mutually intelligible discussions about their organizing activities. They might realize that they have much in common, and they might be working on similar or even the same problems.

A faceted or dimensional perspective acknowledges the diversity of instances of collection types and provides a generative, forward-

looking framework for describing hybrid types that do not cleanly fit into the familiar categories. Even though it might differ from the conventional categories on some dimensions, an organizing system can be designed and understood by its family resemblance on the basis of its similarities on other dimensions to a familiar type of resource collection.

Thinking of organizing systems as points or regions in a design space makes it easier to invent new or more specialized types of collections and their associated interactions. If we think metaphorically of this design space as a map of organizing systems, the empty regions or “white space” between the densely-populated centers of the traditional categories represent organizing systems that do not yet exist. We can consider the properties of an organizing system that could occupy that white space and analyze the technology, process, or policy innovations that might be required to let us build it there. We can reason by analogy to identify and apply the principles used in one organizing system to understand or design others.<sup>1</sup>

1. Depending on which characteristics of Google Books and libraries you think about, you might complete this analogy with an animal theme park like Sea World (<http://www.seaworld.com/>) or a private hunting reserve that creates personalized “big game” hunts. Or maybe you can invent something completely new.

## 12. What Is Being Organized?

*“What is difficult to identify is difficult to describe  
and therefore difficult to organize.”*

[\*\(Svenonius 2000, p. 13\)\*](#)

Before we can begin to organize any resource we often need to identify it. It might seem straightforward to devise an organizing system around tangible resources, but we must be careful not to assume what a resource is. In different situations, the same “thing” can be treated as a unique item, one of many equivalent members of a broad category, or a component of an item rather than as an item on its own. For example, in a museum collection, a handmade, carved chess piece might be a separately identified item, identified as part of a set of carved chess pieces, or treated as one of the 33 unidentified components of an item identified as a chess set (including the board). When merchants assign a stock-keeping unit (SKU) to identify the things they sell, that SKU can be associated with a unique item, sets of items treated as equivalent for inventory or billing purposes, or intangible things like warranties.

You probably do not have explicit labels on the cabinets and drawers in your kitchen or clothes closet, but department stores and warehouses have signs in the aisles and on the shelves because of the larger number of things a store needs to organize. As a collection of resources grows, it often becomes necessary to identify each one explicitly; to create surrogates like bibliographic records or descriptions that distinguish one resource from another; and to create additional organizational mechanisms like shelf labels, store directories, library card catalogs and indexes that facilitate understanding the collection and locating the resources it contains. These organizational mechanisms often suggest or parallel the organizing principles used to organize the collection itself.

Organization mechanisms like aisle signs, store directories and

library card catalogs are embedded in the same physical environment as the resources being organized. But when these mechanisms or surrogates are digitized, the new capabilities that they enable create design challenges. This is because a digital organizing system can be designed and operated according to more abstract and less constraining principles than an organizing system that only contains physical resources. A single physical resource can only be in one place at a time, and interactions with it are constrained by its size, location, and other properties. In contrast, digital copies and surrogates can exist in many places at once and enable searching, sorting, and other interactions with an efficiency and scale impossible for tangible things.

When the resources being organized consist of information content, deciding on the unit of organization is challenging because it might be necessary to look beyond physical properties and consider conceptual or intellectual equivalence. A high school student told to study Shakespeare's play [Macbeth](#) might treat any printed copy or web version as equivalent, and might even try to outwit the teacher by watching a film adaptation of the play. To the student, all versions of [Macbeth](#) seem to be the same resource, but librarians and scholars make much finer distinctions.<sup>1</sup>

1. Organizing systems that follow the rules set forth in the *Functional Requirements for Bibliographic Records* (FRBR) ([Tillett 2005](#)) treat all instances of [Macbeth](#) as the same “work.” However, they also enforce a hierarchical set of distinctions for finer-grained organization. FRBR views books and movies as different “expressions,” different print editions as “manifestations,” and each distinct physical thing in a collection as an “item.” This organizing system thus encodes the degree of intellectual

Archival organizing systems implement a distinctive answer to the question of what is being organized. Archives are a type of collection that focuses on resources created by a particular person, organization, or institution, often during a particular time period. This means that archives have themselves been previously organized as a result of the processes that created and used them. The “original order” of the resources in an archive embodies the implicit or explicit organizing system of the person or entity that created the documents; it is treated as an essential part of the meaning of the collection. As a result, the unit of organization for archival collections is the *fonds*—the original arrangement or grouping, preserving any hierarchy of boxes, folders, envelopes, and individual documents—and thus they are not re-organized according to other (perhaps more systematic) classifications.<sup>2</sup>

equivalence while enabling separate identities where the physical form is important, which is often the case for scholars.

2. Typical examples of archives might be national or government document collections or the specialized Julia Morgan archive at the University of California, Berkeley (<http://www.oac.cdlib.org/findaid/ark:/13030/tf7b69n9k9/>), which houses documents by the famous architect who designed many of the university’s most notable buildings as well as the famous Hearst Castle along the central California coast. The “original order” organizing principle of archival organizing systems was first defined by 19th-century French archivists and is often described as “*respect pour les fonds*.”

Some organizing systems contain legal, business or scientific documents or data that are the digital descendants of paper reports or records of transactions or observations. These organizing systems might need to deal with legacy information that still exists in paper form or in electronic formats like image scans that are different from the structural digital format in which more recent information is likely to be preserved. When legacy conversions from printed information artifacts are complete or unnecessary, an organizing system no longer deals with any of the traditional tangible artifacts. Digital libraries dispense with these artifacts, replacing them with the capability to print copies if needed. This enables libraries of digital documents or data collections to be vastly larger and more accessible across space and time than any library that stores tangible, physical items could ever be.

An increasing number of organizing systems handle resources that are born digital. Ideally, digital texts can be encoded with explicit markup that captures structural boundaries and content distinctions, which can be used to facilitate organization, retrieval, or both. In practice the digital representations of texts are often just image scans that do not support much processing or interaction.

The [William Ashburner collection](#) of historical photos from an 1867-1869 surveying expedition in the Western United States is kept in the University of California, Berkeley's Bancroft Library in the order in which Ashburner, a member of the survey party, had arranged it when he donated it to the library decades later. The arrangement roughly follows a chronological and geographical progression, with some photos obviously out of order and some whose locations cannot be determined.

A similar situation exists for the digital representations of music, photographs, videos, and other non-text content like sensor data, where the digital formats are structurally and semantically opaque.

### Computational Descriptions of People

Each of us is associated with a great many computational descriptions, some of which are used almost every day to make predictions about our behavior using a variety of statistical techniques that are collectively called “predictive analytics.” Whenever you use a credit card, fraud detection algorithms use a model derived from your purchase history to decide, in fractions of a second, whether the transaction is being initiated by you, or by someone who has stolen your card. When you want to buy something expensive on credit, the seller consults your credit score—based on what you owe, your payment history, how long you have had credit, the kinds of credit you have, and other factors—to predict whether you are a good credit risk, and your credit score then gets adjusted if the seller decides to give you credit. Then, after you have bought that expensive item, the seller’s predictive model can use that information to suggest other things you might want to buy.

Philosophers have long debated the extent to which observations of a person’s behavior can yield an understanding of their true and unobservable nature. But whether or not computational descriptions capture a person’s essence, there is no escaping them. If you want to get life or car insurance or a mortgage, models determine what you have to pay. Predictive models are being used to admit people to college, to hire them, to



draft or trade them in professional sports, and to decide whether to monitor them closely because they might be planning a terrorist act. Some companies use “people analytics” software that analyzes every email, calendar item, and document created by employees to build a model of what they know, what they do, when they do it, and who they work with—the goal being to improve communication and collaboration within the firm and with customers.

This book does not emphasize systems that organize people, but it would be remiss not to mention them. Businesses organize their employees, schools organize their faculties and students, sports leagues and teams organize their players, and governments organize their citizens and residents to enable them to vote, drive, attend schools, and receive medical care and ancillary benefits.

Data scientists in all of these fields increasingly predict how employees, students, athletes, voters, drivers – and other categories of people defined by intrinsic or derived characteristics – will behave, decide, live, or die. Once people die, it is no longer necessary to predict anything about them, but nonetheless cemeteries are highly organized.

We often think and talk about time as a resource, and time fits the definition of “anything of value that supports goal-oriented activity” from [“The Concept of ‘Resource’”](#). Furthermore, we could think of the calendar and clock as organizing systems that define time at different levels of granularity to support different kinds of interactions. However, it is probably more useful to think of time as a constraint that influences how and how much to organize.

If you’re sorting your own mail, you can question whether the time you spend on sorting is worth the time you save on searching. But at

scale—imagine 10 million books in a library—the considerable effort required to organize resources saves vastly more time for the many users of the system over its lifetime. Note the inherent tradeoff between time spent on organizing versus retrieval; this will be a recurring theme throughout this book. In a personal context the tradeoff is a matter of individual need or preference, but in social or institutional contexts organization and retrieval are generally done by different people, and their time is likely valued in different ways by the system owner.

# 13. Why Is It Being Organized?

*“The central purpose of systems for organizing information [is] bringing like things together and differentiating among them.”*

[\(Svenonius 2000 p. xi\)](#)

Almost by definition, the essential purpose of any organizing system is to describe or arrange resources so they can be located and accessed later. The organizing principles needed to achieve this goal depend on the types of resources or domains being organized, and in the personal, social, or institutional setting in which organization takes place.

Organizing systems can be distinguished by their dominant purposes or the priority of their common purposes. Libraries, museums, and archives are often classified as [memory institutions](#) to emphasize their primary emphasis on resource preservation. In contrast, “management information systems” or “business systems” are categories that include the great variety of software applications that implement the organizing systems needed to carry out day-to-day business operations.

“Bringing like things together” is an informal organizing principle for many organizing systems. Almost as soon as libraries were invented over two thousand years ago, the earliest librarians saw the need to develop systematic methods for arranging and inventorying their collections.<sup>1</sup> The invention of mechanized

1. [\(Casson 2002\)](#).

printing in the fifteenth century, which radically increased the number of books and periodicals, forced libraries to begin progressively more refined efforts to state the functional requirements for their organizing systems and to be explicit about how they met those requirements.

Today, any information-driven enterprise must have systematic processes and technologies in place that govern information creation or capture and then manage its entire life cycle. Commercial firms need processes for transacting with customers or other firms to carry out business operations, to support research and innovation, marketing, and to develop business strategy and tactics in compliance with laws and regulations for accounting, taxes, human resources, data retention, and so on. In large firms these functions are so highly specialized and complex that the different types of organizing systems have distinct names: *Enterprise Resource Planning (ERP)*, *Enterprise Content Management (ECM)*, *Enterprise Data Management (EDM)*, *Supply Chain Management (SCM)*, *Records Management*, *Customer Relationship Management (CRM)*, *Business Intelligence (BI)*, *Knowledge Management (KM)*, and so on. And even though the most important functions in the organizing systems of large enterprises are those that manage the information resources needed for its business operation, these firms might also need to maintain corporate libraries and archives.

Preserving documents in their physical or original form is the primary purpose of archives and similar organizing systems that contain culturally, historically, or economically significant documents that have value as long-term evidence. Preservation is also an important motivation for the organizing systems of information- and knowledge-intensive firms, where information is primarily in digital formats. Businesses and governmental agencies are usually required by law to keep records of financial transactions, decision-making, personnel matters, and other information essential to business continuity, compliance with regulations and

legal procedures, and transparency. As with archives, it is sometimes critical that these business knowledge or records management systems can retrieve the original documents, although digital copies that can be authenticated are increasingly being accepted as legally equivalent.

This discussion of the requirements for organizing resources in memory institutions and businesses might convey the impression that storing and retrieving resources efficiently are paramount goals, and indeed they are in many contexts. But there are many other reasons for organizing resources, as is easily seen when we look at personal organizing systems. And there are many other ways to compare organizing systems than just how efficiently they enable storing and retrieval functions.

An overarching goal when people are organizing their personal resources is to minimize the effort needed to find the resources. But unlike the finding task in institutional organizing systems, which is generally facilitated with external resource descriptions, finding aids, classifications, search engines, and orientation and navigation mechanisms, the finding task in personal organizing systems is primarily a cognitive one: you need to remember where the resources are and how they are arranged. Because each person has unique experiences and preferences, it is not surprising that people often organize the same types of resources in different ways to make the organization easier to perceive and remember. The resulting resource arrangements often emphasize aesthetic or emotional goals, as when books or clothes are arranged by color or preference, or behavioral goals, as when most frequently used condiments and spices are kept on the kitchen counter rather than stored in a pantry.

When individuals manage their papers, books, documents, record albums, compact discs, DVDs, and other information resources, their organizing systems can vary greatly. This is in part because the content of the resources being organized becomes a consideration.

Furthermore, many of the organizing systems used by individuals are implemented by web applications, and this makes them more accessible than physical resources.<sup>2</sup>

Put another way, an information resource inherently has more potential uses than resources like forks or frying pans, so it is not surprising that the organizing systems in offices are even more diverse than those in kitchens.

When the scale of the collection or the number of intended users increases, two things can happen. The first is that if the system can turn its interaction traces into interaction resources, additional value can be created by analyzing these resources to enhance the interactions, to suggest new ones, or make predictions about how individual users or groups of them will behave. Every business that

2. For example, many people manage their digital photos with Flickr, their home libraries with Library Thing, and their preferences for dining and shopping with Yelp. It is possible to use these “tagging” sites solely in support of individual goals, as tags like “my family,” “to read,” or “buy this” clearly demonstrate. But maintaining a personal organizing system with these web applications potentially augments the individual’s purpose with social goals like conveying information to others, developing a community, or promoting a reputation. Furthermore, because these community or collaborative applications aggregate and share the tags applied by individuals, they shape the individual organizing systems embedded within them when they suggest the most frequent tags for a particular resource.

has a high volume of customer transactions does this; for example, a fast-food restaurant would analyze time-stamped sales data, and might introduce a quick pickup line for items that sell the most, or create product bundles that increase sales while optimizing kitchen and counter work. Amazon.com and other retailers that can capture detailed browsing traces can augment the sales data they collect by treating items that were looked at but not purchased as potential transactions, making them additional inputs to their sophisticated pricing and recommendation systems.

A second likely outcome of increased scale or use is that not everyone is likely to share the same goals and design preferences for the organizing system. If you share a kitchen with housemates, you might have to negotiate and compromise on some of the decisions about how the kitchen is organized so you can all get along. In more formal or institutional organizing systems conflicts between stakeholders can be much more severe, and the organizing principles and policies or permissions for the kinds of interactions available to different users might even be specified in commercial contracts or governed by laws or standards. For example, Bowker and Star note that physicians view the creation of patient records as central to diagnosis and treatment, insurance companies think of them as evidence needed for payment and reimbursement, and researchers think of them as primary data. These groups do not agree on the priority and quality requirements they assign to different information in the patient record, and physicians understandably resist doing work that has no direct benefit for them. Not surprisingly, policy making and regulations about patient records are highly contentious.<sup>3</sup>

Once we acknowledge that stakeholders might not share the same

### 3. [\(Bowker and Star 2000\)](#).

goals, it is clear that efficiency is too narrow a measure for evaluating organizing systems. The ways that resources are organized and interacted with embody the priorities and values of those designing the organizing system, yielding arrangements and interactions designed to control or change the behaviors of the users. Put more bluntly, resources are always organized in ways that are designed to allocate value for some people (e.g., the owners of the resources, or the most frequent users of them) and not for others. From the perspective of the other types of user trying to interact with the system, this organization will likely seem unfair. In this way, organizing resources can often be seen as creating winners and losers, providing benefits to the former and imposing costs or constraints on the latter. For example, search engines analyze interaction resources to adjust search results and choose an ad that is related to your latest query. These are considered improved interactions from the perspective of the search engine, but you might consider it a violation of your privacy and a bit creepy to have the targeted ad follow you around the web until you click on it.

The emerging field of applied behavioral economics, popularized in books like [Freakonomics](#) and [Nudge](#), explains how subtle differences in resource arrangement, the number and framing of choices, and default values can have substantial effects on the decisions people make. Consider the arrangement of salads, pasta dishes, bread, fish, meat, desserts and other types of food in a self-serve cafeteria buffet. In a school setting, the food might be organized and presented to encourage healthier eating, perhaps by making the fatty french fries and high-calorie desserts hard to reach or by providing smaller trays and plates. The same foods would likely be organized differently in an all-you-can-eat restaurant, where the goal is to minimize food costs, with less expensive items like salads at the front of the line to ensure that trays and plates will already be



full when the customer gets to the more expensive items at the end of the line.<sup>4</sup>

The organization of cafeteria buffets to shape user behavior might not seem sinister. However, organizing systems can control behavior in ways that create or perpetuate inequities among their users. This unfairness is a matter of degree: a person who does not own a computer who goes to the public library to check out a popular book loses out when the library enables patrons with computer access to check out books online and assumes that everyone has an equal shot at accessing books via the Internet.

Looking to a much more insidious organizing system, when the South African government adopted Apartheid policies to classify and segregate people by race, it systematized economic and political discrimination and great suffering for the nonwhite population. (See the sidebar, [Power and Politics in Organizing](#).)

#### Power and Politics in Organizing

It is tempting to think of organizing systems and the technologies used to implement them as neutral or objective in their goals and impacts, but it is impossible to argue that the use of racial classification in apartheid South Africa was not a conscious manifestation of prejudice. And even if making it hard for school kids to find the junk food in the cafeteria buffet has health benefits, it nevertheless reflects a paternalistic point of view that restricts individual choices.

#### 4. [\(Levitt 2005\)](#) and [\(Thaler 2008\)](#)

Organizing systems and technology are not developed in a vacuum, unencumbered by politics or social context. As Langdon Winner underscores in [Do Artifacts Have Politics?](#), systems and technologies can be conscious manifestations of the personal (and often political) biases of their creators. Because all people have different experiences and biases, even when they are not conscious of them they influence the design and implementation of organizing systems in ways that can create or perpetuate inequalities.<sup>5</sup>

Technology innovators whose expressed goals are to make something faster, smaller, or cheaper are ignoring the potential for their innovations and automation to render certain types of work less viable and discriminate against people who lack the technology or skills to use it. For example, Winner describes the inadvertent social and political consequences of the introduction of mechanical tomato harvesters in California agriculture in the 1960s. Their industry-wide adoption favored larger farms with more resources to buy the expensive machines, resulting in the disappearance of small tomato farms and large-scale changes to many rural communities whose economies had relied on them.

Some may argue that the mechanical tomato harvester created massive benefits by increasing productivity, but the determination that more efficient tomato production is worth its consequences could be

5. [\(Winner 1980 p 121-136\)](#)

debated. In any case, the debate cannot be answered with a definite yes or no, just as it cannot be with whether the Internet is bad because it has eroded the need for librarians, or whether Uber's clever technologies for matching drivers and riders unfairly avoid the regulations imposed on the taxi industry. Affirming the introduction of the mechanical tomato harvester, search engines, and Uber in the name of productivity, progress, and efficiency is a political point of view.

(See also [“Classification Is Biased”](#) and [The Organizing System Roadmap](#).)

[Classification: Assigning Resources to Categories](#) more fully explains the different purposes for organizing systems, the organizing principles they embody, and the methods for assigning resources to categories.

# 14. How Much Is It Being Organized?

*“It is a general bibliographic truth that not all documents should be accorded the same degree of organization.”*

[\(Svenonius 2000, p. 24\)](#)

Not all resources should be accorded the same degree of organization. In this section we will briefly unpack this notion of degree of organization into three important and related dimensions: the amount of description detail or organization applied to each resource, the amount of organization of resources into classes or categories, and the overall extent to which interactions in and between organizing systems are shaped by resource description and arrangement.

It is important to note that this section is **not** asking the question “how much stuff is being organized?” but rather to what degree is the stuff being organized. Another way to ask the same question is “how many organizing principles are at work?” in this organizing system. Your closet might be arranged only by body part covered and season; an online music store will organize resources by genre, artist name, band name, album name, popularity, date released, and maybe others. So we would say that the online music store is organized much more than the closet, because more organizing principles are at work.

([Resource Description and Metadata](#) and [Categorization: Describing Resource Classes and Types](#), more thoroughly address these questions about the nature and extent of description in organizing systems.)

Not all resources in a collection require the same degree of

description for the simple reason we discussed in [“Why Is It Being Organized?”](#): Organizing systems exist for different purposes and to support different kinds of interactions or functions. Let us contrast two ends of the “degree of description” continuum. Many people use “current events awareness” or “news feed” applications that select news stories whose titles or abstracts contain one or more keywords (Google Alert is a good example). This exact match algorithm is easy to implement, but its all-or-none and one-item-at-a-time comparison misses any stories that use synonyms of the keyword, that are written in languages different from that of the keyword, or that are otherwise relevant but do not contain the exact keyword in the limited part of the document that is scanned. However, users with current events awareness goals do not need to see every news story about some event, and this limited amount of description for each story and the simple method of comparing descriptions are sufficient.

On the other hand, this simple organizing system is inadequate for the purpose of comprehensive retrieval of all documents that relate to some concept, event, or problem. This is a critical task for scholars, scientists, inventors, physicians, attorneys and similar professionals who might need to discover every relevant document in some domain. Instead, this type of organizing system needs rich bibliographic and semantic description of each document, most likely assigned by professional catalogers, and probably using terms from a [controlled vocabulary](#) to enforce consistency in what descriptions mean.

Similarly, different merchants or firms might make different decisions about the extent or granularity of description when they assign SKUs because of differences in suppliers, targeted customers, or other business strategies. If you take your car to the repair shop because windshield wiper fluid is leaking, you might be dismayed to find that the broken rubber seal that is causing the leak cannot be ordered separately and you have to pay to replace the “wiper fluid reservoir” for which the seal is a minor but vital

part. Likewise, when two business applications try to exchange and merge customer information, integration problems arise if one describes a customer as a single “NAME” component while the other separates the customer’s name into “TITLE,” “FIRSTNAME,” and “LASTNAME.”

Even when faced with the same collection of resources, people differ in how much organization they prefer or how much disorganization they can tolerate. A classic study by Tom Malone of how people organize their office workspaces and desks contrasted the strategies and methods of “filers” and “pilers.” Filers maintain clean desktops and systematically organize their papers into categories, while pilers have messy work areas and make few attempts at organization. This contrast has analogues in other organizing systems and we can easily imagine what happens if a “neat freak” and “slob” become roommates.<sup>1</sup>

An equally wide range, from a little organization to a lot, can be seen in the organizing systems for businesses, armies, governments, or any other institutional organizing systems for people. Organizations with broad scope and many people usually have deep hierarchies and explicit reporting relationships with the CEO, general, or president at the top with numerous layers of vice presidents, directors, department heads, and managers (or colonels, majors, captains, lieutenants, and sergeants). Smaller organizations are more varied, with some embodying multi-layered management, and some embracing a flatter arrangement with fewer management levels, wider spans of authority, and more autonomy for individual

1. ([Malone 1983](#)) is the seminal research study, but individual differences in organizing preferences were the basis of Neil Simon’s Broadway play *The Odd Couple* in 1965, which then spawned numerous films and TV series.

workers. Many start-up firms try to grow without any management structure at all in the belief that it makes them more innovative and nimble, but evidence suggests that when no one is responsible for making decisions, the lack of accountability results in poor decisions, or in no decisions at all even when some were sorely needed.<sup>2</sup>

In any case, when people have to do it, describing and organizing resources is work. Stakeholders in an organizing system often have disagreements among about how much organization is necessary because of the implications for who performs the work and who derives the benefits, especially the economic ones. Physicians prefer narrative descriptions and broad classification systems because they make it easier to create patient notes. In contrast, insurance companies and researchers want fine-grained “form-filling” descriptions and detailed classifications that would make the physician’s work more onerous.<sup>3</sup>

The cost-effectiveness of creating systematic and comprehensive descriptions of the resources in an information collection has been debated for nearly two centuries, beginning in 1841 when Sir Anthony Panizzi proposed rules for cataloguing the British Library. In the last half century, the scope of the debate grew to consider the role of computer-generated resource descriptions.<sup>4</sup>

2. ([Silverman 2013](#))

3. See Grudin’s classic work on non-technological barriers to the successful adoption of collaboration technology ([Grudin 1994](#)).

4. Panizzi is most often associated with the origins of modern library cataloging. He ([Panizzi 1841](#)) published

The amount of resource description is always shaped by the currently available technology for capturing, storing, and making use of it. Nineteenth century geologists and paleontologists typically recorded only general information about the depth and surrounding geological features when they found fossils because they had no technology for making more precise measurements and everything they noted they had to record by hand. Today, vastly more detailed information is recorded by instruments and exploited by sophisticated techniques for carbon dating and 3D reconstruction.<sup>5</sup>

Automatically generated descriptions are increasingly an alternative or complement to those created by people. “Smart” resources use sensors to capture information about themselves and their environments (see [“Identity and Active Resources”](#)). Our own computers and phones record information about our keystrokes, clicks, communications, and locations. Business and government computers analyze and index most of the text and speech content that flows through and between our personal phones and

91 cataloging rules for the British Library that defined authoritative forms for titles and author names, but the complexity of the rules and the resulting resource descriptions were widely criticized. For example, the famous author and historian Thomas Carlyle argued that a library catalog should be nothing more than a list of the names of the books in it. Standards for bibliographic description are essential if resources are to be shared between libraries. See ([Denton 2007](#)), ([Anderson and Perez-Carballo 2001a](#), [2001b](#)).

5. ([Bowker and Star 2000 p. 69.](#))



computers. These indexes typically assign weights to the terms according to calculations that consider the frequency and distribution of the terms in both individual documents and in the collection as a whole to create a description of what the documents are about. These descriptions of the documents in the collection are more consistent than those created by human organizers. They allow for more complex query processing and comparison operations by the retrieval functions in the organizing system. For example, query expansion mechanisms can automatically add synonyms and related terms to the search. Additionally, retrieved documents can be arranged by relevance, while “citing” and “cited-by” links can be analyzed to find related relevant documents.

It is important to recognize the potential downside to automated resource description. A detailed description produced by sensors or computers can seem more accurate or authoritative than a simpler one created by a human observer, even if the latter would be more useful for the intended purposes. Moreover, the more detailed the description, the greater the opportunity to use it for new purposes. This might be desirable, as when a company realizes that it can cross- and up-sell because it has been tracking every click in a web store to create a collection of interaction resources. But it could be undesirable, because detailed transaction data can be used to violate privacy and civil rights. It depends on who controls the collected information and their incentives for using it or not using it.

A second constraint on the degree of organization comes from the size of the collection within the scope of the organizing system. Organizing more resources requires more descriptions to distinguish any particular resource from the rest, and more constraining organizing principles. Similar resources need to be grouped or classified to emphasize the most important distinctions among the complete set of resources in the collection. A small neighborhood restaurant might have a short wine list with just ten wines, arranged in two categories for “red” and “white” and

described only by the wine's name and price. In contrast, a gourmet restaurant might have hundreds of wines in its wine list, which would subdivide its "red" and "white" high-level categories into subcategories for country, region of origin, and grape varietal. The description for each wine might in addition include a specific vineyard from which the grapes were sourced, the vintage year, ratings of the wine, and tasting notes.

#### Using "Information Theory" to Quantify Organization

We often hear news stories hyping "how much information" there is in the information society with breathless exuberance about the creation of peta-, exa-, whatever-bytes of content. A much more important and intellectually deeper question than absolute size in bytes is measuring how much information is encoded in the structure or organization of a system. For this we can turn to "Information Theory," a formal approach to understanding the theoretical maximum amount of information that can be carried by a communications system by using efficient coding, data compression, and error correction. It was developed by Claude Shannon, a researcher at Bell Laboratories, and first published as "a mathematical theory of communication" in 1948. We can apply it in the discipline of organizing to compare the amount of structure in different ways of organizing the same resources.<sup>6</sup>

6. Information theory was developed to attack the technical problem of packing the maximum amount of

Information theory quantifies the amount of organization in terms of the number of bits, binary decisions, or rules needed to describe some structure or pattern: the more complex or arbitrary a structure is, the more information it takes to describe it. For example, the organization of a company with a four-level hierarchy and a highly regular reporting structure where everyone supervises five people, can be described quite succinctly. In contrast, a company in which the number of direct reports at any management level is highly variable requires many more rules to describe.

Using measures from information theory to assess the amount of organization yields the somewhat counter-intuitive result that there is less information in the organization of a highly structured system than in a less structured one. It might help to flip this around and describe the amount of organization in terms of the reciprocal of the information measure. A system that is “highly organized” can be modeled or codified with

data into the signal carrying telephone calls, but it quickly provided an essential statistical foundation in language analysis and computational linguistics. ([Shannon 1948](#)). Company organization and other examples applying information theory to the analysis of organizing systems can be found in ([Levitin 2014, Chapter 7](#)).

relatively few rules or organizing principles, compared to a less organized system with many exceptions, corner cases, or one-off rules.

The “entropy” measure is often used to create predictive models of the “decision tree” variety, which is an algorithm that classifies or predicts by making a sequence of logical tests. Each test divides a collection of data into sets with less entropy (more predictability). (See [“Implementing Categories”](#))

At some point a collection grows so large that it is not economically feasible for people to create bibliographic descriptions or to classify each separate resource, unless there are so many users of the collection that their aggregated effort is comparably large; this is organizing by “crowdsourcing.” This leaves two approaches that can be done separately or in tandem.

- The simpler approach is to describe sets of resources or documents as a set or group, which is especially sensible for archives with its emphasis on the *fonds* (see [“What Is Being Organized?”](#)).
- The second approach is to rely on automated and more general-purpose organizing technologies that organize resources through computational means. Search engines are familiar examples of computational organizing technology, and [“Computational Classification”](#) describes other common techniques in machine learning, clustering, and discriminant analysis that can be used to create a system of categories and to assign resources to them.

Finally, we must acknowledge the ways in which information

processing and telecommunications technologies have transformed and will continue to transform organizing systems in every sphere of economic and intellectual activity. A century ago, when the telegraph and telephone enabled rapid communication and business coordination across large distances, these new technologies enabled the creation of massive vertically integrated industrial firms. In the 1920s, the Ford Motor Company owned coal and iron mines, rubber plantations, railroads, and steel mills so it could manage every resource needed in automobile production and reduce the costs and uncertainties of finding suppliers, negotiating with them, and ensuring their contractual compliance. Adam's Smith's invisible hand of the market as an organizing mechanism had been replaced by the visible hand of hierarchical management to control what Ronald Coase in 1937 termed "transaction costs" in [\*The Nature of the Firm\*](#).

In recent decades, a new set of information and computing technologies enabled by Moore's law—unlimited computing power, effectively free bandwidth, and the Internet—have turned Coase upside down, leading to entirely new forms of industrial organization made possible as transaction costs plummet. When computation and coordination costs drop dramatically, it becomes possible for small firms and networks of services (provided by people or by computational processes) to out-compete large corporations through more efficient use of information resources and services, and through more effective information exchange with suppliers and customers, much of it automated. Herbert Simon, a pioneer in artificial intelligence, decision making, and human-computer interaction, recognized the similarities between the design of computing systems and human organizations and developed principles and mechanisms applicable to both.<sup>7</sup>

7. Coase won the 1991 Nobel Prize in economics for his

[\*The Forms of Resource Descriptions\*](#), focuses on the representation of resource descriptions, taking a more technological or implementation perspective. [\*Interactions with Resources\*](#), discusses how the nature and extent of descriptions determines the capabilities of the interactions that locate, compare, combine, or otherwise use resources in information-intensive domains.

work on transaction costs, which he first published as a graduate student ([Coase 1937](#)). Berkeley business professor Oliver Williamson received the prize in 2009 for work that extended Coase's framework to explain the shift from the hierarchical firm to the network firm ([Williamson 1975, 1998](#)). The notion of the "visible hand" comes from ([Chandler 1977](#)). Simon won the Nobel Prize in economics in 1978, but if there were Nobel Prizes in computer science or management theory he surely would have won them as well. Simon was the author or co-author of four books that have each been cited over 10,000 times, including ([Simon 1997, 1996](#)) and ([Newell and Simon 1972](#)).

# 15. When Is It Being Organized?

*“Because bibliographic description, when manually performed, is expensive, it seems likely that the ‘pre’ organizing of information will continue to shift incrementally toward ‘post’ organizing.”*

[\(Svenonius 2000, p. 194-195\)](#)

The organizing system framework recasts the traditional tradeoff between information organization and information retrieval as the decision about *when* the organization is imposed. We can contrast organization imposed on resources “on the way in” when they are created or made part of a collection with “on the way out” organization imposed when an interaction with resources takes place.

When an author writes a document, he or she gives it some internal organization via title, section headings, typographic conventions, page numbers, and other mechanisms that identify its parts and relationship to each other. The document could also have some external organization implied by the context of its publication, such as the name of its author and publisher, its web address, and citations or links to other documents or web pages.

Digital photos, videos, and documents are generally organized to some minimal degree when they are created because some descriptions, notably time and location, are assigned automatically to these types of resources by the technology used to create them. At a minimum, these descriptions include the resource’s creation time, storage format, and chronologically ordered, auto-assigned

filename (IMG00001.JPG, IMG00002.JPG, etc.), but often are much more detailed.<sup>1</sup>

Digital resources created by automated processes generally exhibit a high degree of organization and structure because they are generated automatically in conformance with data or document schemas. These schemas implement the business rules and information models for the orders, invoices, payments, and the numerous other document types created and managed in business organizing systems.

Before a resource becomes part of a library collection, its author-created organization is often supplemented by additional information supplied by the publisher or other human intermediaries, such as an *International Standard Book Number*(ISBN) or *Library of Congress Call Number*(LOC-CN) or *Library of Congress Subject Headings*(LOC-SH).

In contrast, Google and other search engines apply massive computational power to analyze the contents and associated structures (like links between web pages) to impose organization on resources that have already been published or made available so that they can be retrieved in response to a user's query "on the way out." Google makes use of existing organization within and between information resources when it can, but its unparalleled technological capabilities and scale yield competitive advantage in imposing organization on information that was not previously

1. Most digital cameras annotate each photo with detailed information about the camera and its settings in the *Exchangeable Image File Format*(EXIF), and many mobile phones can associate their location along with any digital object they create.



organized digitally.<sup>2</sup> One reaction to the poor quality of some computational description has been the call for libraries to put their authoritative bibliographic resources on the open web, which would enable reuse of reliable information about books, authors, publishers, places, and subject classifications. This “linked data” movement is slowly gathering momentum.<sup>3</sup>

Google makes almost all of its money through personalized ad placement, so much of the selection and ranking of search results is determined “on the way out” in the fraction of a second after the user submits a query by using information about the user’s search history and current context. Of course, this “on the way out” organization is only possible because of the more generic organization that Google’s algorithms have imposed “on the way in.”

2. Indeed, Geoff Nunberg criticized Google for ignoring or undervaluing the descriptive metadata and classifications previously assigned by people and replacing them with algorithmically assigned descriptors, many of which are incorrect or inappropriate. Calling Google’s Book Search a “disaster for scholars” and a “metadata train wreck,” he lists scores of errors in titles, publication dates, and classifications. For example, he reports that a search on “Internet” in books published before 1950 yields 527 results. The first 10 hits for Whitman’s *Leaves of Grass* are variously classified as Poetry, Juvenile Nonfiction, Fiction, Literary Criticism, Biography & Autobiography, and Counterfeits and Counterfeiting. ([Nunberg 2009](#))

3. ([Byrne and Goddard 2010](#)).

In many organizing systems the nature and extent of organization changes over time as the resources are used. The arrangement of resources in a kitchen or office changes incrementally as frequently used things end up in the front of the pantry, drawer, shelf or filing cabinet or on the top of a pile of papers. Printed books or documents acquire margin notes, underlining, turned down pages or coffee cup stains that differentiate the most important or most frequently used parts. Digital documents do not take on coffee cup stains, but when they are edited, their new revision dates put them at the top of directory listings.

The scale of emergent organization of websites, photos on Flickr, blog posts, and other resources that can be accessed and used online dwarfs the incremental evolution of individual organizing systems. This organization is clearly visible in the pattern of links, tags, or ratings that are explicitly associated with these resources, but search engines and advertisers also exploit the less visible organization created over time by analyzing interaction resources, the recorded information about which resources were viewed and which links were followed.

The sort of organic or emergent change in organizing systems that takes place over time contrasts with the planned and systematic maintenance of organizing systems described as [curation](#) or [governance](#), two related but distinct activities. *Curation* usually refers to the methods or systems that add value to and preserve resources, while the concept of *governance* more often emphasizes the institutions or organizations that carry out those activities. The former is most often used for libraries, museums, or archives and the latter for enterprise or inter-enterprise contexts. (For more discussion, see [“Governance”](#))

The organizing systems for businesses and industries often change because of the development of *de facto* or *de jure* standards, or because of regulations, court decisions, or other events or mandates.

We should always consider the extent to which people or technology in an organizing system are able to adapt when new resources, data, or people enter the picture. When and how much an organizing system can be changed depends on the extent of architectural thinking that went into its design (see [The Three Tiers of Organizing Systems](#)), because it should be possible to make a change to a component without having to rethink the system entirely.

Sometimes what prevents adaptation are physical or technological constraints in the implementation of an organizing system, as with a desk or closet with fixed “pigeon holes,” unmovable shelves, or with a music player with limited allowable formats and/or fixed storage capacity.

Machine learning algorithms use different techniques from those of human organizers; one of the important differences is that they’re designed to adapt to new inputs—which is why they’re known to be “learning.” In contrast, humans differ in how willing we are to re-organize to accommodate a different number or a different mix of resources. Without procedures in place to support or trigger adaptation, it may be quite difficult for us to change how we think or how we organize when our world changes, or even to realize that it has changed.

# 16. How (or by Whom) Is It Organized?

*“The rise of the Internet is affecting the actual work of organizing information by shifting it from a relatively few professional indexers and catalogers to the populace at large. An important question today is whether the bibliographic universe can be organized both intelligently (that is, to meet the traditional bibliographic objectives) and automatically.”*

[\*\(Svenonius 2000, p. 26\)\*](#)

In the preceding quote, Svenonius identifies three different ways for the “work of organizing information” to be performed: by professional indexers and catalogers, by the populace at large, and by automated (computerized) processes. Our notion of the organizing system is broader than her “bibliographic universe,” making it necessary to extend her taxonomy. Authors are increasingly organizing the content they create, and it is important to distinguish users in informal and formal or institutional contexts. We have also introduced the concept of an organizing [\*agent\*](#) ([\*“The Concept of ‘Organizing Principle’”\*](#)) to unify organizing done by people and by computer algorithms.

Professional indexers and catalogers undergo extensive training to learn the concepts, controlled descriptive vocabularies, and standard classifications in the particular domains in which they work. Their goal is not only to describe individual resources, but to position them in the larger collection in which they reside.<sup>1</sup>

1. This is an important distinction in library science

They can create and maintain organizing systems with consistent high quality, but their work often requires additional research, which is costly.

The class of professional organizers also includes the employees of commercial information services like Westlaw and LexisNexis, who add controlled and, often, proprietary metadata to legal and government documents and other news sources. Scientists and scholars with deep expertise in a domain often function as the professional organizers for data collections, scholarly publications and proceedings, and other specialized information resources in their respective disciplines. The *National Association of Professional Organizers*(NAPO) claims several thousand members who will organize your media collection, kitchen, closet, garage or entire house or help you downsize to a smaller living space.<sup>2</sup>

education and library practice. Individual resources are described (“formal” cataloging) using “bibliographic languages” and their classification in the larger collection is done using “subject languages” ([Svenonius 2000, Ch. 4 and Ch. 8, respectively](#)). These two practices are generally taught in different library school courses because they use different languages, methods and rules and are generally carried out by different people in the library. In other organizations, the resource description (both formal and subject) is created in the same step and by the same person.

2. NAPO: <http://www.napo.net> The name and scope of this organization seems a bit odd given how much professional organizing takes place in business, science,

Many of today's content creators are unlikely to be professional organizers, but presumably the author best understands why something was created and the purposes for which it can be used. To the extent that authors want to help others find a resource, they will assign descriptions or classifications that they expect will be useful to those users. But unlike professional organizers, most authors are unfamiliar with controlled vocabularies and standard classifications, and as a result their descriptions will be more subjective and less consistent.

Similarly, most of us do not hire professionals to organize the resources we collect and use in our personal lives, and thus our organizing systems reflect our individual preferences and idiosyncrasies.

Non-author users in the “populace at large” are most often creating organization for their own benefit. These ordinary users are unlikely to use standard descriptors and classifications, and the organization they impose sometimes so closely reflects their own perspective and goals that it is not useful for others. Fortunately most users of “Web 2.0” or “community content” applications at least partly recognize that the organization of resources emerges from the aggregated contributions of all users, which provides incentive to use less egocentric [descriptors](#) and [classifications](#). The staggering number of users and resources on the most popular applications inevitably leads to “tag convergence” simply because of the statistics of large sample sizes.

Finally, the vast size of the web and the even greater size of the “deep” or invisible web, composed of the information stores of

government, medicine, education, and other domains where closets and garages are not the most important focus.

business and proprietary information services, makes it impossible to imagine today that it could be organized by anything other than the massive computational power of search engine providers like Google and Microsoft. Likewise, data mining, predictive analytics, recommendation systems, and many other application areas that involve computational modeling and classification simply could not be done any other way.<sup>3</sup>

Nevertheless, in the earliest days of the web, significant human effort was applied to organize it. Most notable is Yahoo!, founded by Jerry Yang and David Filo in 1994 as a directory of favorite websites. For many years the Yahoo! homepage was the best way to find relevant websites by browsing the extensive system of classification. Today's Yahoo! homepage emphasizes a search engine that makes it appear more like Google or Microsoft Bing, but the Yahoo! directory can still be found if you search for it.

3. ([He et al. 2007](#)) estimate that there are hundreds of thousands of websites and databases whose content is accessible only through query forms and web services, and there are over a million of those. The amount of content in this hidden web is many hundreds of times larger than that accessible in the surface or visible web.

See <http://www.worldwidewebsite.com/> for estimates of the size of the visible web calculated from comparisons of results from search engines.

# 17. Where is it being Organized?

*“Bibliographic control requires fixing a document in the bibliographic universe by its space-time coordinates.”*  
[\(Svenonius 2000, p. 120\)](#)

Having identified the resources, reasoned about our motivations, limited the scope and scale, and determined when and by whom the organization will occur, we come finally to the question of where the resources are being organized.

In ordinary use, “Where” refers to a physical location. But the answer to “where?” often depends on whether we are asking about the current location, a past location, or an intended destination for resources that are in transit or in process. The answer to the question “where?” can take a lot of different forms. We can talk about an abstract space like “a library shelf” or we can talk about “the hidden compartment in Section XY at the Library of Congress,” as depicted in the 2004 movie “National Treasure.” We can answer “where?” with a description of a set of environmental conditions that best suit a class of wildlife, or a tire, or a sleeping bag. We can answer “where?” with “Renaissance Europe” or “Colonial Williamsburg.” “Where?” can be a place in a mental construct, or even a place in an imagined location.

In the architectural design of an organizing system, its physical location is usually not a primary concern. In most organizing systems, the matter of where the organizing system and the resources are located can be abstracted away. So, in practice, resource location often is not as important as the other questions here. Physical constraints of the storage location should generally



be relegated to an implementation concern rather than an architectural one. The construction of a special display structure for a valuable resource is not an independent design dimension; it is just the implementation of the user interface. (See [“The Implementation Perspective”](#))

Physical resources are often stored where it is convenient and efficient to do so, whether in ordinary warehouses, offices, storerooms, shelves, cabinets, and closets. It can be necessary to adapt an organizing system to characteristics of its physical environment, but this could undermine architectural thinking and make it harder to maintain the organization over time, as the collection evolves in scope and scale. (See [“Organizing Physical Resources”](#))

Digital resources, on the other hand, are increasingly organized and stored “in the cloud” and their actual locations are invisible, indeterminate, and generally irrelevant, except in situations where the servers and the information they hold may be subject to laws or practices of their physical location. For example, a controversy arose in Canada in 2013 when researchers discovered that Internet service providers were, for various technical and business reasons, routinely routing trans-Canada web traffic through the United States. Because Canada has no jurisdiction over data traveling through cables and servers in another country, there was considerable outcry among Canadians who were concerned that their personal information was being subjected to the privacy laws and practices of another country without their knowledge or consent.

Sometimes location functions as an organizing principle in its own right, which in practice essentially collapses many of these architectural distinctions. This is frequently the case in our personal organizing systems, where we may exploit the innate human capability for spatial memory by always putting specific things like keys, eyeglasses, and cell phones in the same place, which makes

them easy to find. But we can also see this happening in systems as complex and varied as: real estate information systems; wayfinding systems, such as road signage or mile markers; standardized international customs forms with position-specific data fields; geographic information systems; air, ground, sea, and space traffic control systems; and historic landmark preservation.

In [“Organizing Places”](#) we consider the organization of the land, built environments, and wayfinding systems. [“The Structural Perspective”](#) discusses the structural perspective on resource relationships, and in some systems, it may be very significant where resources are located in relation to one another. In [The Barnes Collection](#), for example, works of art are physically grouped to enunciate common characteristics. Conversely, zoos do not mix the kangaroos with the wild dogs, and the military does not mix the ingredients for chemical weapons (at least, not until they plan to use them). There are also circumstances where resources can only exist in (or are particularly suited to) particular environments, such as the conditions required to grow wine grapes or mushrooms, or store spent nuclear fuel. UPS advises companies on where to put their warehouses and shipment centers. These are more substantial than questions of presentation, but it is debatable whether it falls under the storage or logic tier (you could have the principle of “keep the mushrooms somewhere moist” while not dictating where particularly).

Sometimes the location of an organizing system seems particularly salient, as in the design of cities where the street plan can be essential for orientation and navigation, and is embodied in zoning, voting, and other explicit organization, as well as in informal organization like neighborhood identity. But even here, it is really the people who live in the city who are being organized and whose interactions with the city and with each other are being encouraged or discouraged, not the physical location on which they live.

Indeed, in designing an organizing system you will often find that

questions about location tumble naturally out of the other five design dimensions. For instance, questions about “when,” “what,” and “where” are often inseparable, particularly when an organizing system is subject to outside regulations, which tend to have geographical jurisdictions. “Where” is also commonly bound up with “who” and “why,” when locational challenges or opportunities faced by a system’s creators or users necessitate special design consideration. (See [“Effectivity”](#))

Location can be critically important to an organizing system—too important, in fact, to be considered alone. The question of “where?” is best considered in context of the other five design dimensions as a whole; a narrow focus on where the resources are being organized too often privileges past convention over architectural thinking and perpetuates legacy issues and poorly organized systems.

# 18. Key Points in Chapter Two

- How does a “design questions” or “dimensional” perspective on the design of organizing systems complement the familiar use of categories like library and museum?

A dimensional perspective makes it easier to translate between category- and discipline-specific vocabularies so that people from different disciplines can have mutually intelligible discussions about their organizing activities.

(See [“Introduction”](#))

- Why is the question “What is a thing?” so fundamental and challenging?

In different situations, the same “thing” can be treated as a unique item, one of many equivalent members of a broad category, or a component of an item rather than as an item on its own.

(See [“What Is Being Organized?”](#))

- How are organizing systems for physical resources and those for digital resources fundamentally different?

A single physical resource can only be in one place at a time, and interactions with it are constrained by its size, location, and other properties. In contrast, digital copies and surrogates can exist in many places at once and enable searching, sorting, and other interactions with an efficiency and scale impossible for tangible things.

(See [“What Is Being Organized?”](#))

- Why is it challenging to decide on the unit of organization for information content?

When the resources being organized consist of information content, deciding on the unit of organization is challenging because it might be necessary to look beyond physical properties and consider conceptual or intellectual equivalence.

(See [“What Is Being Organized?”](#))

- What is the essential purpose of any organizing system?

Almost by definition, the essential purpose of any organizing system is to describe or arrange resources so they can be located and accessed later. The organizing principles needed to achieve this goal depend on the types of resources or domains being organized, and in the personal, social, or institutional setting in which organization takes place.

(See [“Why Is It Being Organized?”](#))

- What are the primary purposes for the organizing systems in libraries, museums, and archives?

Libraries, museums, and archives are often classified as [memory institutions](#) to emphasize their primary emphasis on resource preservation.

(See [“Why Is It Being Organized?”](#))

- What kinds of documents are businesses and governmental agencies required to keep?

Businesses and governmental agencies are usually required by law to keep records of financial transactions, decision-making, personnel matters, and other information essential to business continuity, compliance with regulations and legal procedures, and transparency.

(See [“Why Is It Being Organized?”](#))

- What is the value created if interaction traces can be turned into interaction resources?

If a system can turn its interaction traces into interaction resources, additional value can be created by analyzing these resources to enhance the interactions, to suggest new ones, or make predictions about how individual users or groups of them will behave.

(See [“Why Is It Being Organized?”](#))

- Why is efficiency too narrow a measure for evaluating organizing systems?

Resources are always organized in ways that are designed to allocate value for some people (e.g., the owners of the resources, or the most frequent users of them) and not for others.

(See [“Why Is It Being Organized?”](#))

- What lessons from applied behavioral economics about how people make decisions have implications for the design of organizing systems?

Subtle differences in resource arrangement, the number and framing of choices, and default values can have substantial effects on the decisions people make.

(See [“Why Is It Being Organized?”](#))

- Why might merchants or firms differ in the extent or granularity of their product descriptions?

Different merchants or firms might make different decisions about the extent or granularity of description when they assign

SKUs because of differences in suppliers, targeted customers, or other business strategies.

(See [“How Much Is It Being Organized?”](#))

- What are some of the potential downsides to automated resource description?

A detailed description produced by sensors or computers can seem more accurate or authoritative than a simpler one created by a human observer, even if the latter would be more useful for the intended purposes. Detailed transaction data can be used to violate privacy and civil rights.

(See [“How Much Is It Being Organized?”](#))

- How does the number of resources in a collection affect the amount of resource description and organization required?

Organizing more resources requires more descriptions to distinguish any particular resource from the rest, and more constraining organizing principles. Similar resources need to be grouped or classified to emphasize the most important distinctions among the complete set of resources in the collection.

(See [“How Much Is It Being Organized?”](#))

- How is organizing “on the way in” different from organizing “on the way out”?

We can contrast organization imposed on resources “on the way in” when they are created or made part of a collection with “on the way out” organization imposed when an interaction with resources takes place.

(See [“When Is It Being Organized?”](#))

- Why do digital resources created by automated processes exhibit a high degree of organization and structure?

Digital resources created by automated processes generally exhibit a high degree of organization and structure because they are generated automatically in conformance with data or document schemas.

(See [“When Is It Being Organized?”](#))

- What kinds of organizing systems would be impossible to create without the use of massive computational power?

The vast size of the web and the even greater size of the “deep” or invisible web makes it impossible to imagine today that it could be organized by anything other than the massive computational power of search engine providers like Google and Microsoft. Likewise, data mining, predictive analytics, recommendation systems, and many other application areas that involve computational modeling and classification simply could not be done any other way.

(See [“How \(or by Whom\) Is It Organized?”](#))





PART III

# ACTIVITIES IN ORGANIZING SYSTEMS

Robert J. Glushko

Erik Wilde

Jess Hemerly

Isabelle Sperano

Robyn Perry



## 19. Introduction (III)

There are four [activities](#) that occur naturally in every [organizing system](#); how explicit they are depend on the scope, the breadth or variety of the resources, and the scale, the number of resources that the organizing system encompasses. Consider the routine, everyday task of managing your wardrobe. When you organize your clothes closet, you are unlikely to write a formal [selection](#) policy that specifies what things go in the closet. You do not consciously itemize and prioritize the ways you expect to search for and locate things, and you are unlikely to consider explicitly the organizing principles that you use to arrange them. From time to time you will put things back in order and discard things you no longer wear, but you probably will not schedule this as a regular activity on your calendar.

Your clothes closet is an organizing system; defined as “an intentionally arranged collection of resources and the interactions they support.” As such, it exposes these four highly interrelated and iterative [activities](#):

### **Selecting**

Determining the scope of the organizing system by specifying which resources should be included. (*Should I hang up my sweaters in the clothes closet or put them in a dresser drawer in the bedroom?*)

### **Organizing**

Specifying the principles or rules that will be followed to arrange the resources. (*Should I sort my shirts by color, sleeve type, or season?*)

## **Designing resource-based interactions**

Designing and implementing the actions, functions or services that make use of the resources. (*Do I need storage places for clothes to be laundered? Should I have separate baskets for white and colors? Dry cleaning?*)

## **Maintaining**

Managing and adapting the resources and the organization imposed on them as needed to support the interactions. (*When is it time to straighten up the closet? What about mending? Should I toss out clothes based on wear and tear, how long I have owned them, or whether I am tired of them? What about excess hangers?*)

These activities are not entirely separable or sequential, and they can be informal for your clothes closet because its scope and scale are limited. In institutional organizing systems the activities and the inter-dependencies and iterations among them are more carefully managed and often highly formal.

For example, a data warehouse combines data from different sources like orders, sales, customers, inventory, and finance. Business analysts explore combinations and subsets of the data to find important patterns and relationships. The most important questions in the design and operation of the data warehouse can be arranged using the same activities as the clothes closet.

## **Selecting**

Which data sources should be included? How is their quality assessed? How much of the data is sampled? How are queries composed?

## **Organizing**

Which data formats and schemas will enable effective

processing? Are needed transformations made at load time or query time?

### **Designing resource-based interactions**

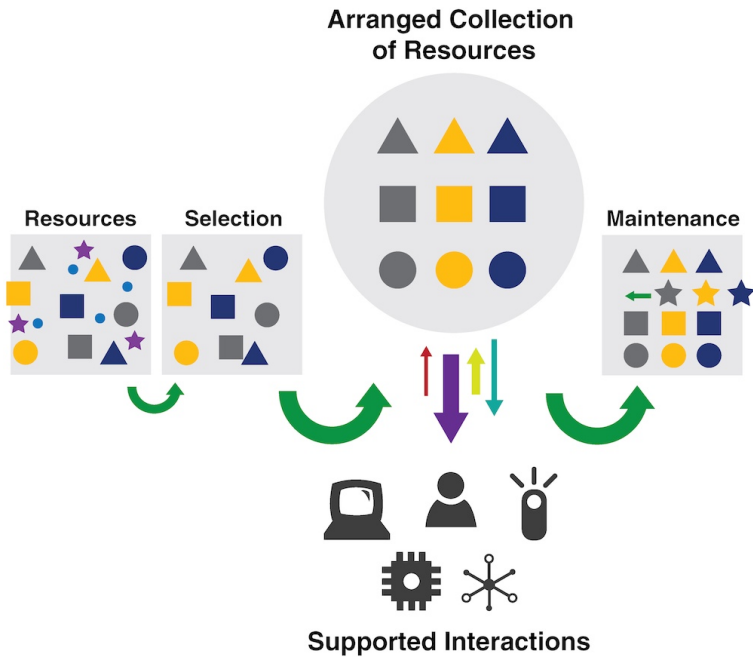
What are the most important and frequent queries that need to be pre-configured?

### **Maintaining**

What governance policies and procedures are needed to satisfy retention, compliance, security, and privacy requirements?

[Figure: Four Activities in all Organizing Systems.](#) illustrates these four [activities](#) in all organizing systems, framing the depiction of the organizing and interaction design activities shown in [Figure: An Organizing System.](#) with the [selection](#) and [maintenance](#) activities that necessarily precede and follow them.

Four Activities in all Organizing Systems



Four [activities](#) take place in all organizing systems: selection of resources for a collection; intentional organization of the resources; design and implementation of interactions with individual resources or with the collection, and; maintenance of the resources and the interactions over time.

These activities are deeply ingrained in academic curricula and professional practices, with domain-specific terms for their methods and results. Libraries and museums usually make their [selection](#) principles explicit in *collection development* policies. Adding a resource to a library collection is called *acquisition*, but adding to a museum collection is called *accessioning*. Documenting the contents of library and museum collections to organize them is called *cataloging*. [Circulation](#) is a central interaction in libraries, but because museum resources do not circulate the primary interactions for museum users are [viewing](#) or [visiting](#) the collection.

[Maintenance](#) activities are usually described as [preservation](#) or [curation](#).

In business information systems, [selection](#) of resources can involve *data generation, capture, sampling, or extraction*. Adding resources could involve *loading, integration, or insertion*. Schema development and *data transformation* are important organizing activities. Supported interactions could include *querying, reporting, [analysis](#), or visualization*. [Maintenance](#) activities are often described as [deletion, purging, data cleansing, governance](#), or [compliance](#).

#### What about “Creating” Resources?

Our definition of organizing system as an intentionally arranged collection of resources might seem to imply that resources must exist before they are organized. This is often the case when we organize physical resources because the need for principled organization only arises when the collection gets too big for us to see everything in the collection at once. Similarly, many data analytics projects begin by bringing together data collected by others.

However, organizing systems for digital resources are often put in place as a prerequisite for creating them. This is always necessary when the resources are created by automated processes or data entry in business systems, and usually the case with professional writers in a technical publications context. We can think of database or document schemas (at the implementation tier) or data entry forms or word processor templates (in the user interface tier) as embodiments of the organizing principles in the data records or documents that are then created in conformance with them.



Domain-specific methods and vocabularies evolve over time to capture the complex and distinctive sets of experiences and practices of their respective disciplines. We can identify correspondences and overlapping meanings, but they are not synonyms or substitutes for each other. We propose more general terms like [selection](#) and [maintenance](#), not as lowest common denominator replacements for these more specialized ones, but to facilitate communication and cooperation across the numerous disciplines that are concerned with organizing.

It might sound odd to describe the animals in a zoo as resources, to think of viewing a painting in a museum as an interaction, or to say that destroying information to comply with privacy regulations is maintenance. Taking a broader perspective on the activities in organizing systems so that we can identify best practices and patterns enables people with different backgrounds and working in different domains to understand and learn from each other.

Part of what a database administrator can learn from a museum curator follows from the rich associations the curator has accumulated around the concept of curation that are not available around the more general concept of maintenance. Without the shared concept of [maintenance](#) to bridge their disciplines, this learning could not take place.

# Navigating this chapter

In [“The Concept of “Resource””](#) and [“What Is Being Organized?”](#) we briefly discussed the fundamental concept of a [resource](#). In this chapter, we describe the four primary [activities](#) with resources, using examples from many different kinds of organizing systems.

[“Selecting Resources”](#)

[“Organizing Resources”](#)

[“Designing Resource-based Interactions”](#)

[“Maintaining Resources”](#)

We emphasize the activities of organizing and of designing resource-based interactions that make use of the organization imposed on the resources. We discuss [selection](#) and [maintenance](#) to create the context for the organizing [activities](#) and to highlight the interdependencies of [organizing](#) and these other [activities](#). This broad survey enables us to compare and contrast the activities in different resource domains, setting the stage for a more thorough discussion of resources and [resource description](#) in [Resources in](#)

Organizing Systems and Resource Description and Metadata.

## 20. Selecting Resources

When we talk about organizing systems, we often do so in terms of the contents of their collections. This implies that the most fundamental decision for an organizing system is determining its resource domain, the group or type of resources that are being organized. This decision is usually a constraint, not a choice; we acquire or encounter some resources that we need to interact with over time, and we need to organize them so we can do that effectively.

*Selecting* is the process by which resources are identified, evaluated, and then added to a collection in an organizing system. [Selection](#) is first shaped by the domain and then by the scope of the organizing system, which can be analyzed through six interrelated aspects:

1. the number and nature of users
2. the time span or lifetime over which the organizing system is expected to operate
3. the size of the collection
4. the expected changes to the collection
5. the physical or technological environment in which the organizing system is situated or implemented
6. the relationship of the organizing system to other ones that overlap with it in domain or scope.

(In [The Organizing System Roadmap](#), we discuss these six aspects in more detail.)

## Selection Criteria

Selection must be an intentional process because, by definition, an organizing system contains resources whose selection and arrangement was determined by human or computational agents, not by natural processes. And given the broad definition of resource as “anything of value that can support goal-oriented activity” it follows that resources should be selected by an implicit or explicit assessment to determine whether they can be used to achieve those goals. So even though particular selection methods and criteria vary across resource domains, their common purpose is to determine how well the resource satisfies the specifications for the properties or capabilities that enable a person or nonhuman agent to perform the intended activities. “Fitness for use” is a common and concise way to summarize this idea, and while it highlights the need to have activities in mind before resources are selected to enable them, it also explains why precise selection criteria are harder to define for organizing systems that have diverse sets of stakeholders or users with different goals, like those in public libraries.

Many resources are evaluated and selected one-at-a-time. This makes it impossible to specify in advance every property or criterion that might be considered in making a selection decision, especially for unique or rare resources like those being considered by a museum or private collector. In general, when resources are treated as instances, organizing activities typically occur after selection takes place, as in the closet organizing system with which we began this chapter.

When the resources being considered for a collection are more homogeneous and predictable, it is possible to treat them as a class or set, which enables selection criteria and organizing principles to be specified in advance. This makes selection and organizing into concurrent activities. This would be the case in the data warehouse organizing system, the other example at the beginning of this

chapter, because each data source can be described by a schema whose structure is reflected in the organization of the data warehouse. Put another way, as long as subsequent datasets from a specific source do not differ in structure, only in temporal attributes like their creation or acquisition dates, the organization imposed on the initial dataset can be replicated for each subsequent one.

Well-run companies and organizations in every industry are highly systematic in selecting the resources that must be managed and the information needed to manage them. “Selecting the right resource for the job” is a clichéd way of saying this, but this slogan nonetheless applies broadly to raw materials, functional equipment, information resources and datasets, and to people, who are often called “human resources” in corporate-speak.

For some types of resources, the specifications that guide selection can be precise and measurable. Precise specifications are especially important when an organizing system will contain or make use of all resources of a particular type, or if all the resources produced from a particular source become part of the organizing system on some regular schedule. Selection specifications can also be shaped by laws, regulations or policies that require or prohibit the collection of certain kinds of objects or types of information.<sup>1</sup>

1. Some governments attempt to preserve and prevent misappropriation of cultural property by enforcing import or export controls on antiquities that might be stolen from archaeological sites ([Merryman 2006](#)). For digital resources, privacy laws prohibit the collection or misuse of personally identifiable information about healthcare, education, telecommunications, video rental,

For example, when a manufacturer of physical goods selects the materials or components that are transformed into its products, it carefully evaluates the candidate resources and their suppliers before making them part of its supply chain. The manufacturer would test the resources against required values of measurable characteristics like chemical purity, strength, capacity, and reliability. A business looking for transactional or demographic data to guide a business expansion strategy would specify different measurable characteristics; data files must be valid with respect to a schema, must contain no duplicates or personally identifiable information, and must be less than one month old when they are delivered. Similarly, employee selection has become highly data-intensive; employers hire people after assessing the match between their competencies and capabilities (expressed verbally or in a resume, or demonstrated in some qualification test) and what is needed to do the required activities.<sup>2</sup>

and might soon restrict the information collected during web browsing.

2. The popular LinkedIn site, which has hundreds of millions of resumes that it data mines to find statistically superior job candidates, is literally a gold mine for the company because it makes money by referring those candidates to potential employers. Data-intensive hiring practices in baseball are entertainingly presented in the book entitled [Moneyball](#) book ([Lewis 2003](#)) or the 2011 movie starring Brad Pitt. Pro football teams have begun to assess college football players by comparing them statistically with the best pro players ([Robbins, 2016](#)).

*Selection* is an essential activity in creating organizing systems whose purpose is to combine separate web services or resources to create a composite service or application according to the business design philosophy of [Service Oriented Architecture\(SOA\)](#).<sup>3</sup> When an information-intensive enterprise or application combines its

Many examples of business strategies that required significant investment to acquire data assets with no current value are reported in [\(Provost and Fawcett 2013\)](#).

3. See [\(Cherbakov et al. 2005\)](#), [\(Erl 2005a\)](#). The essence of SOA is to treat business services or functions as components that can be combined as needed. An SOA enables a business to quickly and cost-effectively change how it does business and whom it does business with (suppliers, business partners, or customers). SOA is generally implemented using web services that exchange *Extensible Markup Language(XML)* documents in real-time information flows to interconnect the business service components. If the business service components are described abstractly it can be possible for one service provider to be transparently substituted for another—a kind of real-time resource selection—to maintain the desired quality of service. For example, a web retailer might send a Shipping Request to many delivery services, one of which is selected to provide the service. It probably does not matter to the customer which delivery service handles his package, and it might not even matter to the retailer.



internal services with ones provided by others via Application Programming Interfaces (APIs), the resources are selected to create a combined collection of services according to the “core competency” principle: resources are selected and combined to exploit the first party’s internal capabilities and those of its service partners better than any other combination of services could. For example, instead of writing millions of lines of code and collecting detailed maps to build an interactive map in an application, you can get access to the Google Maps organizing system with just a few lines of code.<sup>4</sup> (See the sidebar, [Selection of Web-based Resources](#))

#### Selection of Web-based Resources

The nature and scale of the web changes how we collect resources and fundamentally challenges how we think of resources in the first place. Web-based resources cannot be selected for a collection by

4. The idea that a firm’s long term success can depend on just a handful of critical capabilities that cut across current technologies and organizational boundaries makes a firm’s core competency a very abstract conceptual model of how it is organized. This concept was first proposed by ([Prahalad and Hamel 1990](#)), and since then there have been literally hundreds of business books that all say essentially the same thing: you cannot be good at everything; choose what you need to be good at and focus on getting better at them; let someone else do things that you do not need to be good at doing.

consulting a centralized authoritative directory, catalog, or index because one does not exist. ProgrammableWeb and other directories organize thousands of web-accessible APIs, and the dominant resource-organizing firms Amazon, Salesforce, Facebook, and Twitter offer hundreds of APIs to access massive amounts of information about products, people, and posts, but APIs enable access to only a fraction of the web's content. And although your favorite web search engine consults an index or directory of web resources when you enter a search query, you do not know where that index or directory came from or how it was assembled.<sup>56</sup>

5. See ([Borgman 2000](#)) on digitization and libraries. But while shared collections benefit users and reduce acquisition costs, if a library has defined itself as a physical place and emphasizes its *holdings*— the resources it directly controls—it might resist anything that reduces the importance of its physical reification, the size of its holdings, or the control it has over resources ([Sandler 2006](#)). A challenge facing conventional libraries today is to make the transition from emphasizing creation and preservation of physical collections to facilitating the use and creation of knowledge regardless of its medium and the location from which it is accessed.
6. ([Arasu et al. 2001](#)), ([Manning et al. 2008](#)). The web is a

However, the web has universal scope and global reach, making most of the web irrelevant to most people most of the time. Researchers have attacked this problem by treating the web as a combination of a very large number of topic-based or domain-specific collections of resources, and then developing techniques for extracting these collections as digital libraries targeted for particular users and uses.<sup>7</sup>

graph, so all web crawlers use graph traversal algorithms to find URIs of web resources and then add any hyperlink they find to the list of URIs they visit. The sheer size of the web makes crawling its pages a bandwidth- and computation intensive process, and since some pages change frequently and others not at all, an effective crawler must be smart at how it prioritizes the pages it collects and how it re-crawls pages. A web crawler for a search engine can determine the most relevant, popular, and credible pages from query logs and visit them more often. For other sites, a crawler adjusts its “revisit frequency” based on the “change frequency” ([Cho and Garcia-Molina 2000](#)).

7. Web resources are typically discovered by computerized “web crawlers” that find them by following links in a methodical automated manner. Web crawlers can be used to create topic-based or domain-specific

Scientific and business data are ideally selected after assessments of their quality and their relevance to answering specific questions. But this is easy to say and hard to do. It is essential to assess the quality of individual data items to find data entry problems such as misspellings and duplicate records, or data values that are illegal, statistical outliers, or otherwise suspicious. It is also essential to assess the quality of data as a collection to determine if there are problems in what data was collected, by whom or how it was collected and managed, the format and precision in which it is stored, whether the schema governing each instance is rigorous enough, and whether the collection is complete. In addition, copyright, licensing, consumer protection laws, competitive considerations, or simply the lack of incentives to share resources

collections of web resources by changing the “breadth-first” policy of generic crawlers to a “best-first” approach. Such “focused crawlers” only visit pages that have a high probability of being relevant to the topic or domain, which can be estimated by analyzing the similarity of the text of the linking and linked pages, terms in the linked page’s URI, or locating explicit semantic annotation that describes their content or their interfaces if they are invocable services ([Bergmark et al. 2002](#)), ([Ding et al. 2004](#)).

make it difficult to obtain the best or most appropriate resources.<sup>8</sup>  
(See the sidebar, [Assessing and Addressing Data Quality](#))

In some domains, the nature of the resources or the goals they are intended to satisfy imply selection criteria that are inherently less quantifiable and more subjective. This is easy to see in personal collections, where selection criteria can be unconventional, idiosyncratic, or otherwise biased by the subjective perspective and experience of the collector. Most of the clothes and shoes you own have a reason for being in your closet, but could anyone else explain the contents of your closet and its organizing system, and why you bought that crazy-looking dress or shirt?

Both libraries and museums typically formalize their selection principles in [collection development](#) policies that establish priorities for acquiring resources that reflect the people they serve and the services they provide to them. The diversity of user types in public libraries and many museums implies that narrowly-targeted criteria would produce a collection of resources that would fail to satisfy many of the users. As a result, libraries typically select resources on the basis of broader criteria like their utility and relevance to their user populations, and try to choose resources that add the most value to their existing collections, given the cost constraints that most libraries are currently facing. Museums often emphasize

8. FTC Fair Information Practice Principles say that consumer data collected for one purpose cannot be used for other purposes without the consumer's consent. Sometimes called the consumer privacy bill of rights.

See also [\(Zhu et al., 2014\)](#) and [\(Marchioni et al., 2012\)](#)

intrinsic value, scarcity, or uniqueness as selection criteria, even if the resources lack any contemporary use.<sup>9</sup>

Even when selection criteria can be measured and evaluated in isolation, they are often incompatible or difficult to satisfy in

9. Large research libraries have historically viewed their [collections](#) as their intellectual capital and have policies that specify the subjects and sources that they intend to emphasize as they build their collections. See [\(Evans 2000\)](#). Museums are often wary of accepting items that might not have been legally acquired or that have claims on them from donor heirs or descendant groups; in the USA, much controversy exists because museums contain many human skeletal remains and artifacts that Native American groups want to be repatriated.

Adding a resource to a museum implies an obligation to preserve it forever, so many museums follow rigorous accessioning procedures before accepting it. Likewise, archives usually perform an additional appraisal step to determine the quality and value of materials offered to them.

In archives, common appraisal criteria include uniqueness, the credibility of the source, the extent of documentation, and the rights and potential for reuse. To oversimplify: libraries decide what to keep, museums decide what to accept, and archives decide what to throw away.

combination. It would be desirable for data to be timely, accurate, complete, and consistent, but these criteria trade off against one other, and any prioritization that values one criterion over another is somewhat subjective. In addition, explicitly subjective perceptions of resource quality are hard to ignore; people are inclined to choose resources that come in attractive packages or that are sold and supported by attractive and friendly people.

Many of the examples in this section have involved selection principles whose purpose was to create a collection of desirable, rare, skilled, or otherwise distinctive resources. After all, no one would visit a museum whose artifacts were ordinary, and no one would watch a sports team made up of randomly chosen athletes because it could never win. However, choosing resources by randomly sampling from a large population is essential if your goal is to make inferences about it without having to study all its instances. Sampling is especially necessary with very large populations when timely decisions are required. A good sample for statistical purposes is one in which the selected resources are not different in any important way from the ones that were not selected.

Sampling is also important when large numbers of resources need to be selected to satisfy functional requirements. A manufacturer cannot test every part arriving at the factory, but might randomly test some of them from different shipments to ensure that parts satisfy their acceptance criteria.

## Looking “Upstream” and “Downstream” to Select Resources

As we have seen, selection principles and activities differ across resource domains, and there is another important difference in selection that considers resources from the perspective of their history or the future.

In [“Selection Criteria”](#) we discussed the activity of selecting resources by assessing their conformance with specifications for required properties or capabilities. However, if you can determine where the resources come from, you can make better selection decisions by evaluating the people, processes, and organizing systems that create them. Using the analogy of a river, we can follow a resource “upstream” from us until we find the “headwaters.” Physical resources might have their headwaters in a factory, farm, or artist’s studio. Digital resources might have headwaters in a government agency, a scientist’s laboratory, or a web-based commerce site.

When interaction resources ([“The Concept of “Interaction Resource””](#)) are incorporated into the organizing system that creates them, as when records of a person’s choices and behaviors are used to personalize subsequent information, the headwaters are obviously easy to find. However, even though finding the headwaters where resources come from is often not easy and sometimes not possible, that is where you are most likely to find the people best able to answer the questions, described in [Design Decisions in Organizing Systems](#), that define any organizing system. The resource creators or producers will know the assumptions and tradeoffs they made that influence whether the resources will satisfy your requirements, and you can assess what they (or their documents that describe the resources) tell you and the credibility they have in telling it. You should also try to evaluate the processes or algorithms that produce the resources, and then decide if they are capable of yielding resources of acceptable quality.

The best outcome is to find a credible supplier of good quality resources. However, if an otherwise desirable supplier does not currently produce resources of sufficient quality, it is worth trying to improve the quality by changing the process using instruction or incentives. Advocates for open government have succeeded in getting numerous US government entities to publish data for free in machine-readable formats, but it was partly as a result of somewhat



subversive demonstration projects and shaming that the government finally created data.gov in 2009. A clear lesson from the “quality movement” and statistical process control is that interventions that fix quality problems at their source are almost always a better investment than repeated work to fix problems that were preventable (see endnote<sup>10</sup>). But if you cannot find the headwaters or you are not able to address quality problems at their source, you can sometimes transform the resources to give them the characteristics or quality they need.<sup>11</sup> (See the sidebar, [Assessing and Addressing Data Quality](#), and [“Transforming Resources for Interactions”](#).)

#### Assessing and Addressing Data Quality

If an organizing system uses data acquired from some external source, it is essential to assess its quality as an “intake” process. Ideally, the data comes with a schema that explicitly specifies what is expected, including legal structures, data types, and values (See [“Structuring](#)

10. [link to footnote]

11. See [\(Tauberer 2014\)](#) for a history of the “civic hacking” and the open data movement.

The [Sunlight Foundation](#) (<http://sunlightfoundation.com/>) and [Code For America](#) (<https://www.codeforamerica.org/>) are good sources for keeping up with open government issues and initiatives.

[Descriptions](#)"). This intake process runs tests that find problems and then runs processes to fix the problems.

There are a great many techniques for finding problems in numeric and string data. Some problems like missing data, duplicate records, spelling mistakes, and extreme or “outlier” values are easy to detect. A credit card charge for \$10,000,000 is obviously a bad piece of data in a college student’s account. Detecting anomalous and duplicate data is especially important because including them can produce misleading statistics and predictions, as well as creating the nuisance for consumers of receiving multiple copies of product catalogs, each with a different misspelling in a name or address.

Sometimes a dataset is valid with respect to its own specification but becomes problematical when it is combined with another dataset that has a different specification. Entities in the two datasets might not be described using the same units at the same point in time. So instead of analyzing and repairing resource instances, data cleaning must now be applied to every resource in a dataset, as when every “Zip Code” in a United States mailing directory is given the more universal “Postal Code” label, or when datasets using DD/MM/YY and MM/DD/YY formats for dates are combined.

Other data quality problems are harder to detect because they are contextual; a data value might be valid in some contexts but the same value might be invalid in others. For example, if you live in San Francisco and your credit card is used for transactions in Barcelona or

Berlin, it could be fraud, or maybe you are on vacation. Similarly, high or low ratings for business establishments on sites like Yelp might be appropriate responses to excellent or poor service, but might also reflect “pay for rating” manipulation in the former case, and efforts by competitors to undermine rival businesses in the latter.<sup>12</sup>

When you cannot obtain resources directly from their source, even if you have confidence in their quality at that point, it is important to analyze any evidence or records of their use or interactions as they flow downstream. (See [“Resources over Time”](#)) Physical resources are often associated with printed or digital documents that make claims about their origin and authenticity, and often have bar codes, RFID tags, or other technological mechanisms that enable them to be tracked from their headwaters to the places where they are used. Tracking is very important for data resources because they can often be added to, derived from, or otherwise changed without

12. On data modeling: see [\(Kent 2012\)](#), [\(Silverston 2000\)](#), [\(Glushko and McGrath 2005\)](#). For data warehouses see [\(Turban et al. 2010\)](#).

For a classification and review of data cleaning problems and methods, see [\(Rahm and Do, 2000\)](#). A recent and popular analysis that describes data cleaning as “data wrangling, data munging, and data janitor work” is [\(Lohr 2014\)](#). For a survey of anomaly detection see [\(Chandola 2009\)](#).

leaving visible traces. Just as the water from melted mountain snow becomes less pure as it flows downstream, a data resource can become “dirty” or “noisy” over time, reducing its quality from the perspective of another person or computational agent further downstream. Data often gets dirty when it is combined with other datasets that contain duplicate or seemingly-duplicate information. Data can also become dirty when the hardware or software that stores it changes. Subtle differences in representation formats, transaction management, enforcement of integrity constraints, and calculations of derived values can change the original data.

In addition, a data resource can become inaccurate or obsolete simply because the world that the data describes has changed with the passage of time. People move, change jobs, get married or divorced, or die. Likewise, companies move, merge, get spun off, or go out of business. A poll taken a year before an election is often not a good predictor of the ultimate winner.<sup>13</sup>

Other selection processes look “downstream” to select resources on the basis of predicted rather than current properties, capability, or suitability. Sports teams often sign promising athletes for their minor league teams, and businesses hire interns, train their employees, and run executive development programs to prepare promising low-level managers for executive roles. Businesses sometimes conduct experiments with variable product offers and pricing to collect data they will need in the future to power predictive models that will repay the investment in data acquisition many times over.

13. [\(Kim et al, 2003\).](#)



## 2I. Organizing Resources

Organizing systems arrange resources according to many different principles. In libraries, museums, businesses, government agencies and other long-lived institutions, organizing principles are typically documented as cataloging rules, information management policies, or other explicit and systematic procedures so that different people can apply them consistently over time. In contrast, the principles for arranging resources in personal or small-scale organizing systems are usually informal and often inconsistent or conflicting.

### Property

In this book we use *property* in a generic and ordinary sense as a synonym for *feature* or “characteristic.” Many cognitive and computer scientists are more precise in defining these terms and reserve *property* for binary predicates (e.g., something is red or not, round or not). If multiple values are possible, the *property* is called an *attribute*, “dimension,” or “variable.” *Feature* is used in *data science* and *machine learning* contexts for both “raw” or observable variables and “latent” ones, extracted or constructed from the original set.<sup>1</sup>

For most types of resources, any number of principles could be

1. See ([Barsalou and Hale 1983](#)) for a rigorous contrast between feature lists and other representational formalisms in models of human categories.

used as the basis for their organization depending on the answers to the “why?” ([“Why Is It Being Organized?”](#)), “how much?” ([“How Much Is It Being Organized?”](#)), and “how?” ([“How \(or by Whom\) Is It Organized?”](#)) questions posed in [Design Decisions in Organizing Systems](#).

A simple principle for organizing resources is colocation —*putting all the resources in the same location: in the same container, on the same shelf, or in the same email in-box*. However, most organizing systems use principles that are based on specific resource properties or properties derived from the collection as a whole. What properties are significant and how to think about them depends on the number of resources being organized, the purposes for which they are being organized, and on the experiences and implicit or explicit biases of the intended users of the organizing system. The implementation of the organizing system also shapes the need for, and the nature of, the resource properties.<sup>2</sup>

Classification Biases

Libraries and bookstores use different classification

2. For example, a personal or small organizing system would typically use properties that are easy to identify and understand. In contrast, an organizing system for very large collections of resources, or data about them, would choose properties that are statistically optimal, even if they are not interpretable by people, because of the greater need for operational efficiency and predictive accuracy.

systems. The kitchen in a restaurant is not organized like a home kitchen because professional cooks think of cooking differently than ordinary people do. Scientists use the Latin or binomial (genus + species) scheme for identifying and classifying living things to avoid the ambiguities and inconsistencies of common names, which differ across languages and often within different regions in a single language community.

Many resource collections acquire resources one at a time or in sets of related resources that can initially be treated the same way. Therefore, it is natural to arrange resources based on properties that can be assessed and interpreted when the resource becomes part of the collection.

“Subject matter” organization involves the use of a classification system that provides categories and descriptive terms for indicating what a resource is about. Because they use *aboutness* properties that are not directly perceived, methods for assigning subject classifications are intellectually-intensive and in many cases require rigorous training to be performed consistently and appropriately.<sup>3</sup>

3. Libraries and bookstores use different classification systems. The kitchen in a restaurant is not organized like a home kitchen because professional cooks think of cooking differently than ordinary people do. Scientists use the Latin or binomial (genus + species) scheme for identifying and classifying living things to avoid the ambiguities and inconsistencies of common names,



Nevertheless, the cost and time required for this human effort motivates the use of computational techniques for organizing resources.

As computing power steadily increases, the bias toward computational organization gets even stronger. However, an important concern arises when computational methods for organizing resources use so-called “black box” methods that create resource descriptions and organizing principles that are not inspectable or interpretable by people. In some applications more efficient information retrieval or question answering, more accurate predictions, or more personalized recommendations justify making the tradeoff. But comprehensibility is critical in many medical, military, financial, or scientific applications, where trusting a prediction can have life or death implications or cause substantial time or money to be spent.<sup>4</sup>

## Organizing Physical Resources

When the resources being arranged are physical or tangible things—such as books, paintings, animals, or cooking pots—any resource can be in only one place at a time in libraries, museums, zoos, or kitchens. Similarly, when organizing involves recording information in a physical medium—carving in stone, imprinting in clay, applying ink to paper by hand or with a printing press—how

which differ across languages and often within different regions in a single language community.

4. [\(Freitas 2014\)](#) and [\(Burrell 2015\)](#).

this information can be organized is subject to the intrinsic properties and constraints of physical things.

The inescapable tangibility of physical resources means that their organizing systems are often strongly influenced by the material or medium in which the resources are presented or represented. For example, museums generally collect original artifacts and their collections are commonly organized according to the type of thing being collected. There are art museums, sculpture museums, craft museums, toy museums, science museums, and so on.

Similarly, because they have different material manifestations, we usually organize our printed books in a different location than our record albums, which might be near but remain separate from our CDs and DVDs. This is partly because the storage environments for physical resources (shelves, cabinets, closets, and so on) have co-evolved with the physical resources they store.<sup>5</sup>

The resource collections of organizing systems in physical environments often grow to fit the size of the environment or place in which they are maintained—the bookshelf, closet, warehouse, library or museum building. Their scale can be large: the Smithsonian Institute in Washington, D.C., the world's largest museum and research complex, consists of 19 museums, 9 research facilities, a zoo and a library with 1.5 million books. However, at some point, any physical space gets too crowded, and it is difficult and expensive to add new floors or galleries to an existing library or museum.

5. Many of the ancient libraries in Greece and Rome have been identified by archaeologists by characteristic architectural features ([Casson 2002](#)). See also ([Battles 2003](#)).

### Organizing People into Businesses

How people are organized into businesses is the essence of the discipline of management, and different aspects are taught in industrial organization and behavior, operations, entrepreneurship, and other courses. Organizing people in a business is often called “human resource management,” and many of the principles for organizing physical resources and information resources apply to organizing people.

In addition, economics, strategy, and business culture are important considerations. There are a huge number of ways to organize people that differ in the extent of hierarchical structure, the flow of information up and down the hierarchy, the span of control for managers, and the discretion people have to deviate or innovate with respect to the work they have been assigned to do. For example, we can contrast law firms with a hierarchy of partners, associates, and paralegals with the self-management “holacracy” that companies like Zappos have experimented with, in which authority and decision-making are highly distributed among the employees.

Regardless of how the firm is organized, we can analyze it using economist Ronald Coase’s idea of “transaction costs,” which a business incurs in searching for and negotiating with suppliers, business partners, and customers, and in particular we can consider how new information and computing technologies reduce

these costs to make the firm more efficient while remaining flexible.<sup>6</sup>

### *Organizing with Properties of Physical Resources*

Physical resources are often organized according to intrinsic physical properties like their size, color, or shape, because the human visual system automatically pays a lot of attention to them.

This inescapable aspect of visual perception was first formalized by German psychologists starting a century ago as the Gestalt principles (see the sidebar, [Gestalt Principles](#)). Likewise, because people have limited attentional capacity, we ignore a lot of the ongoing complexity of visual (and auditory) stimulation, making us perceive our sensory world as simpler than it really is. Taken together, these two ideas explain why we automatically or “pre-attentively” organize separate things we see as groups or patterns based on their proximity and similarity. They also explain why arranging physical resources using these quickly perceived attributes can seem more aesthetic or satisfying than organizing them using properties that take more time to understand. Look at the cover of this book; the most organized arrangement of the colors and shapes just jumps out at you more than the others.

6. ([Robertson 2015](#)) and ([Coase 1937](#)).

### Gestalt Principles

Psychologists Max Wertheimer, Wolfgang Kohler, and Kurt Koffka proposed several principles—proximity, similarity, continuity, connection, enclosure, and closure—that explain how our visual system imposes order on what it sees. There are always multiple interpretations of the sensory stimuli gathered by our visual system, but the mind imposes the simplest ones: things near each other are grouped, complex shapes are viewed as simple shapes that are overlapping, missing information needed to see separate visual patterns as continuous or whole is filled in, and ambiguous figure-ground illusions are given one interpretation at a time.

Koffka's pithy way of explaining the core idea of all the principles was that "The whole is other than the sum of the parts," which has been distorted over time to the cliché that "the whole is more than the sum of the parts."<sup>7</sup>

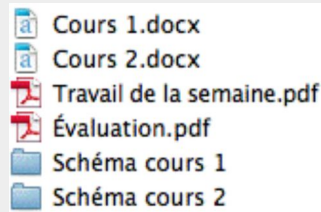
Designers of graphics and information visualizations rely on Gestalt rules because the automatic

7. The Gestalt principles are a staple in every introductory psychology textbook, but the classic text ([Koffka 1935](#)) has recently been reprinted. A group of distinguished contemporary researchers in visual perception ([Wagemans et al, 2012](#)) recently reviewed the history and impact of Gestalt psychology on their hundredth birthday.

interpretations created by the human visual system enable their designs to be understood more quickly. This of course implies that designs that violate the Gestalt rules will be harder to understand. Camouflage—the use of disruptive coloration, colors and patterns that resemble backgrounds, countershading, shadow elimination, and similar techniques that make it difficult for the visual system to detect objects and edges—proves the power of Gestalt processing.<sup>8</sup>

#### Gestalt Similarity in Action

This simple example of a list of files illustrates the Gestalt similarity principle that elements that look similar are seen as being in the same group. The shape and visual design for the icons



8. Texts that ground graphic design and information visualization in Gestalt principles include [\(Cairo 2012\)](#) and [\(Few 2004\)](#). [\(Johnson 2013\)](#) explains them within the broader scope of user interface design.

distinguish files from folders, and file types from each other. These interpretations do not require you to be able to read the labels, although the visual similarity of the names suggests some similarity of content.

Physical resources are also commonly organized using intrinsically associated properties such as the place and time they were created or discovered. The shirts in your clothes closet might be arranged by color, by fabric, or style. We can view dress shirts, T-shirts, Hawaiian shirts and other styles as configurations of shirt properties that are so frequent and familiar that they have become linguistic and cultural categories. Other people might think about these same properties or categories differently, using a greater or lesser number of colors or ordering them differently, sorting the shirts by style first and then by color, or vice versa.

In addition to, or instead of, physical properties of your shirts, you might employ behavioral or usage-based properties to arrange them. You might separate your party and Hawaiian shirts from those you wear to the office. You might put the shirts you wear most often in the front of the closet so they are easy to locate. Unlike intrinsic properties of resources, which do not change, behavioral or usage-based properties are dynamic. You might move to Hawaii, where you can wear Hawaiian shirts to the office, or you could get tired of what were once your favorite shirts and stop wearing them as often as you used to.

Some arrangements of physical resources are constrained or precluded by resource properties that might cause problems for other resources or for their users. Hazardous or flammable materials should not be stored where they might spill or ignite; lions and antelopes should not share the same zoo habitat or the

former will eat the latter; adult books and movies should not be kept in a library where children might accidentally find them; and people who are confrontational, passive aggressive, or arrogant do not make good team members when tough decisions need to be made. For almost any resource, it seems possible to imagine a combination with another resource that might have unfortunate consequences. We have no shortage of professional certifications, building codes, MPAA movie ratings, and other types of laws and regulations designed to keep us safe from potentially dangerous resources.

### *Organizing with Descriptions of Physical Resources*

To overcome the inherent constraints with organizing physical resources, organizing systems often use additional physical resources that describe the primary physical ones, with the library card catalog being the classic example. A specific physical resource might be in a particular place, but multiple description resources for it can be in many different places at the same time.



Card Catalog Cabinet





Library catalogs were managed as collections of printed cards for much of the 20th century, and the wooden cabinets that contained them were ubiquitous functional furniture in every library. Today such cabinets are often considered “retro” or antique treasures.

(Photo by R. Glushko.)

Card From Library Catalog

Z671                      Phillips, David Rhys, 1868-  
L6                        The monastic libraries of Wales, fifth to  
v.14                      sixteenth centuries (Celtic and mediaeval  
                              periods).  
  
                              (In: Library association record. London.  
                              v. 14 (1912), p. 288-316, 374-398)

A catalog card from the library of the School of Library and Information Studies at the University of California, Berkeley. The card describes a book about the monastic libraries of Wales, which like the library in which this card came from are no longer in existence.

(Photo by R. Glushko.)

When the description resources are themselves digital, as when a printed library card catalog is put online, the additional layer of abstraction created enables additional organizing possibilities that can ignore physical properties of resources and many of the details about how they are stored.

In organizing systems that use additional resources to identify or describe primary ones, “adding to a collection” is a logical act that need not require any actual movement, copying, or reorganization of the primary resources. This virtual addition allows the same resources to be part of many collections at the same time; the same book can be listed in many bibliographies, the same web page can be in many lists of web bookmarks and have incoming links from many different pages, and a publisher’s digital article repository can be licensed to any number of libraries.

## Organizing Places

Places are physical resources, but unlike the previous two subsections where we treat the environment as given (the library or museum building, the card catalog or bookshelf) and discuss how we organize resources like books in that environment, we can take an alternative perspective and discuss how we design that physical environment. These environments could be any of the following:

- The land itself, as when we lay out city plans when organizing how people live together and interact in cities.
- A “built environment,” a human-made space, particular building, or a set of connected spaces and buildings. A built environment could be a museum, airport, hospital, casino, department store, farm, road system, or any kind of building or space where resources are arranged and people interact with them.
- The orientation and navigation aids that enable users to understand and interact in built environments. These are resource descriptions that support the interaction requirements of the users.

These are not entirely separable contexts, but they are easier to discuss as if they are considered as such.

### *Organizing the Land*

Cities naturally emerge in places that can support life and commerce. Almost all major cities are built on coasts or rivers because water provides sustenance, transportation and commercial links, and power to enable industry. Many very old cities have crowded and convoluted street plans that do not seem intentionally

organized, but grid plans in cities also have a very long history. Cities in the Middle East were laid out in rough grids as far back as 2000+ BCE. Using long axes was a way to create an impression of importance and power.

Because the United States, and especially the American West, was not heavily settled until much more recently compared to most of Europe and Asia, it was a place for people to experiment with new ideas in urban design. The natural human tendency to impose order on habitation location had ample room to do just that. The easiest and most efficient way to organize space is using a coordinate grid, with streets intersecting at perpendicular angles. Salt Lake City, Albuquerque, Phoenix, and Seattle are notable examples of grid cities. An interesting hybrid structure exists in Washington DC, which has radiating diagonal avenues overlaid on a grid.<sup>9</sup>

9. Salt Lake City takes the use of a grid to an extreme because the central area is extremely flat. Streets are named by numbers and letters, so you might find yourself at the intersection of “North A Street” and “3rd Avenue N,” or at the intersection of “W 100 S” and “S 200 W.” It is a little creepy to think that your street address is a pinpoint location in the big grid.

In contrast, Seattle imposes the grid in an abstract way, ignoring the fact that there are many lakes, rivers, and hills that break up the grid. Streets keep the same names even though they are not connected, and the grid stretches for many miles out from its origin in Seattle. You can be up in the mountains at the corner of “294th Avenue SE” and “472nd Street SE,” giving you precise

## *Organizing Built Environments*

Built environments influence the expectations, behaviors, and experiences of everyone who enters the space—employees, visitors, customers, and inhabitants are all subject to the design of the spaces they occupy. These environments can be designed to encourage or discourage interactions between people, to create a sense of freedom or confinement, to reward exploration or enforce efficiency, and of course, much much more. The arrangement of

information about your location and nearly 50 mile distance from downtown Seattle.

(See also Pierre Charles L'Enfant's plan for DC at [http://en.wikipedia.org/wiki/Pierre\\_Charles\\_L%27Enfant](http://en.wikipedia.org/wiki/Pierre_Charles_L%27Enfant))

This is not to say that imposing arbitrary grids on top of a physical environment to create a simple and easily understood organization is always desirable. It is essential that any organization imposed on a region be sensitive to any social, cultural, linguistic, ethnic, or religious organizing systems already in place. Much of the recent conflict and instability in the Middle East can be attributed to the implausibly straight line borders drawn by the French and British to carve up the defeated Ottoman Empire a century ago. Because the newly-created countries of Syria and Iraq lacked ethnic and religious cohesion, they could only be held together by dictatorships. ([Trofimov 2015](#))

the resources in a built environment also encourages or discourages interactions, and sometimes the built environment is designed with a specific collection of resources in mind to enable and reinforce some particular interaction goals or policies.

If we contrast the built environments of museums, airports, and casinos, and the way in which each of them facilitates or constrains interactions are more obvious. Museums are often housed in buildings designed as architectural monuments that over time become symbols of national, civic, or cultural identity. Many old art museums mimic classical architecture, with grand stairs flanked by tall columns. They have large and dramatic entry halls that invite visitors inside. Modern museums are decidedly less traditional, and some people complain that the architecture of modern art museums can overshadow the art collection contained within because people are induced to pay more attention to the building than to its contents.

Some recently built airports have been designed with architectural flair, but airport design is more concerned with efficiency, walkability (maybe with the aid of moving walkways), navigability, and basic comfort for travelers getting in and out the airport. Wide walkways, multiple staircases, and people movers whose doors open in one direction at a time, all encourage people to move in certain directions, sometimes without the people even realizing they are being directed.

If you have ever been lost in a casino or had trouble finding the exit you can be sure you experienced a casino that achieved its main design goals: keeping people inside and making it easy for them to lose track of time because they lack both windows and clocks. As American architect Robert Venturi points out, “The intricate maze under the low ceiling never connects with outside light or outside

space...This disorients the occupant in space and time... He loses track of where he is and when it is.”<sup>10</sup>

If one accepts the premise that values and bias are at work in decisions about organizing systems, it is difficult not to see it in built environments. Consider queue design in banks, supermarkets, or boarding airplanes. Assuming that it is desirable to organize people efficiently to minimize wait times and crowding, how should the queue be designed? How many categories of people should there be? What is the basis for the categories?

It may be uncontroversial to include several express lanes in a supermarket checkout, because people can choose to buy fewer items if they do not want to wait. Similarly, it seems essential for hospital emergency rooms to have a triage policy that selects patients from the emergency room queue based on their likely benefit from immediate medical attention.

However, consider the dynamic created by queue design at Disneyland to give priority to people with physical limitations and disabilities. This seemingly socially respectful decision was exploited by a devious collaboration between disabled people and wealthy non-disabled people who hired them to pose as family members, enabling the entire “family” to cut ahead of everyone else. In response, Disney modified the policy favoring disabled patrons, causing numerous complaints about Disney’s insensitivity to their concerns.<sup>11</sup>

10. [\(Shiner 2007\)](#). The comparison of the organizing systems in casinos and airports comes from [\(Curran 2011\)](#). [\(Venturi 1972\)](#)
11. The number of queues, their locations and their layout (if spatial) is referred to as the “queue configuration.”

There are many other examples of how values and biases become part of built environments. In the mid-20th century the road systems of Long Island in New York were designed with low overpasses, which prevented public buses from passing under them, effectively segregating the beaches. The trend in college campus design after the student protests of the 1960s and 1970s was to create layouts that would prevent or frustrate large demonstrations.<sup>12</sup>

The “queue discipline” is the policy for selecting the next customer from the queue. Most common discipline is “First come, first served.” Frequent, higher-paying, or some other customer segment might have their own queue with FCFS applied within it.

See the New York Post article at <http://nypost.com/2013/05/14/rich-manhattan-moms-hire-handicapped-tour-guides-so-kids-can-cut-lines-at-disney-world/>

12. The designer of the road system, Robert Moses, heralded as the master builder of mid-20th century New York City, built roads to enforce his idea of who should frequent Long Island (affluent whites). The overpasses were intentionally designed with clearances (often around nine feet) that were too low for public buses. Consequently, low-income bus riders (largely people of color) had no way to get to beaches. See ([Winner 1980](#)).



## *Orientation and Wayfinding Mechanisms*

It is easy to move through an environment and stay oriented if the design is simple and consistent, but most built environments must include additional features or descriptions to assist people in these tasks. Distinctive architectural elements can create landmarks for orientation, and spaces can be differentiated with color, lighting, furnishings, or other means. More ubiquitous mechanisms include signs, room numbers, or directional arrows highlighting the way and distance to important destinations.

In airports, for example, there are many orientation signs and display terminals that help passengers find their departure gates, baggage, or ground transportation services. In contrast, casinos provide little orientation and navigation support because increased confusion leads to lengthier visits, and more gambling on the part of the casino's visitors.

A recent innovation in wayfinding and orientation mechanisms is to give them sensing and communication capabilities so they can identify people by their smartphones and then provide personalized directions or information. These so-called “beacon” systems have been deployed at numerous airports, including London's Gatwick, San Francisco, and Miami.<sup>13</sup>

13. ([Arthur and Passini 1992](#)) ([McCartney 2015](#)) McCartney, Scott. Technology will speed you through the airport of the future. Wall Street Journal, July 15 2015.

## Organizing Digital Resources

Organizing systems that arrange digital resources like digital documents or information services have some important differences from those that organize physical resources. Because digital resources can be easily copied or interlinked, they are free from the “one place at a time” limitation.<sup>14</sup> The actual storage locations for digital resources are no longer visible or very important. It hardly matters if a digital document or video resides on a computer in Berkeley or Bangalore if it can be located and accessed efficiently.<sup>15</sup>

14. In principle, it is easy to make perfect copies of digital resources. In practice, however, many industries employ a wide range of technologies including digital rights management, watermarking, and license servers to prevent copying of documents, music or video files, and other digital resources. The degree of copying allowed in digital organizing systems is a design choice that is shaped by law.
15. Web-based or “cloud” services are invoked through URIs, and good design practice makes them permanent even if the implementation or location of the resource they identify changes ([Berners-Lee 1998](#)). Digital resources are often replicated in content delivery networks to improve performance, reliability, scalability, and security ([Pathan et al. 2008](#)); the web pages served by a busy site

Moreover, because the functions and capabilities of digital resources are not directly manifested as physical properties, the constraints imposed on all [material](#) objects do not matter to digital content in many circumstances.<sup>16</sup>

might actually be delivered from different parts of the world, depending on where the accessing user is located.

16. Whether a digital resource seems intangible or tangible depends on the scale of the digital collection and whether we focus on individual resources or the entire collection. An email message is an identified digital resource in a standard format, RFC 2822 ([Resnick 2001](#)). We can compare different email systems according to the kinds of interactions they support and how easy it is to carry them out, but how email resources are represented does not matter to us and they surely seem intangible. Similarly, the organizing system we use to manage email might employ a complex hierarchy of folders or just a single searchable in-box, but whether that organization is implemented in the computer or smart phone we use for email or exists somewhere “in the cloud” for web-based email does not much matter to us either. An email message is tangible when we print it on paper, but all that matters then is that there is well-defined mapping between the different representations of the abstract email resource.

On the other hand, at the scale at which Google and

### Materiality

An emerging issue in the field of digital humanities is the requirement to recognize the *materiality* of the environment that enables people to create and interact with digital resources. Even if the resources themselves are intangible, it can be necessary to study and preserve the technological and social context in which they exist to fully understand them.<sup>17</sup>

An organizing system for digital resources can also use digital

Microsoft handle billions of email messages in their Gmail and Hotmail services the implementation of the email organizing system is extremely relevant and involves many tangible considerations. The location and design of data centers, the configuration of processors and storage devices, the network capacity for delivering messages, whether messages and folder structures are server or client based, and numerous other considerations contribute to the quality of service that we experience when we interact with the email organizing system.

17. ([Schreibman, Siemens, and Unsworth 2005](#)) and ([Leonardi 2010](#)). For example, a “Born-Digital Archives” program at Emory University is preserving a collection of the author Salman Rushdie’s work that includes his four personal computers and an external hard drive. ([Kirschenbaum 2008](#)), and ([Kirschenbaum et al. 2009](#)).

description resources that are associated with them. Since the incremental costs of adding processing and storage capacity to digital organizing systems are small, collections of both primary digital resources and description resources can be arbitrarily large. Digital organizing systems can support collections and interactions at a scale that is impossible in organizing systems that are entirely physical, and they can implement services and functions that exploit the exponentially growing processing, storage and communication capabilities available today. This all sounds good, unless you are the small local business with limited onsite inventory that cannot compete with global web retailers that offer many more choices from a network of warehouses.<sup>18</sup>

There are inherently more arrangements of digital resources than there are for physical ones, but this difference emerges because of multiple implementation platforms for the organizing system as much as in the nature of the resources. Nevertheless, the organizing systems for digital books, music and video collections often

18. For example, a car dealer might be able to keep track of a few dozen new and used cars on his lot even without a computerized inventory system, but web-based AutoTrader.com offered more than 2,000,000 cars in 2012. The cars are physical resources where they are located in the world, but they are represented in the AutoTrader.com organizing system as digital resources, and cars can be searched for using any combination of the many resource properties in the car listings: price, body style, make, model, year, mileage, color, location, and even specific car features like sunroofs or heated seats.

maintain the distinctions embodied in the organizing system for physical resources because it enables their co-existence or simply because of legacy inertia. As a result, the organizing systems for collections of digital resources tend to be coarsely distinguished by media type (e.g., document management, digital music collection, digital video collection, digital photo collection, etc.).

Information resources in either physical or digital form are typically organized using intrinsic properties like author names, creation dates, publisher, or the set of words that they contain. Information resources can also be organized using assigned properties like subject classifications, names, or identifiers. Information resources can also be organized using behavioral or transactional properties collected about individuals or about groups of people with similar interaction histories. For example, Amazon and Netflix use browsing and purchasing behavior to make book and movie recommendations.<sup>19</sup>

19. Even when organizing principles such as [\*alphabetical\*](#), [\*chronological\*](#), or numerical ordering do not explicitly consider physical properties, how the resources are arranged in the “storage tier” of the organizing system can still be constrained by their physical properties and by the physical characteristics of the environments in which they are arranged. Books can only be stacked so high whether they are arranged alphabetically or by frequency of use, and large picture books often end up on the taller bottom shelf of bookcases because that is the only shelf they fit. Nevertheless, it is important to treat these idiosyncratic outcomes in physical storage as

Complex organization and interactions are possible when organizing systems with digital resources are based on the data type or data model of the digital content (e.g., text, numeric, multimedia, statistical, geospatial, logical, scientific, personnel, and so on).

Interactions with numeric data can be further distinguished according to the *levels of measurement* embodied in the number, which determine how much quantitative processing makes sense:

- *Nominal* level data uses a number as an identifier for an instance or a category to distinguish it from other ones. Products in a catalog might have numbers associated with them, but the products have no intrinsic order, so no measurements using the numbers are meaningful other than the frequency with which they occur in the dataset. The most frequently occurring value is called the *mode*.
- *Ordinal* level data indicates a direction or ranking on some naturally ordered scale. We know that the first place finisher in a race came in ahead of the second place one, who finished ahead of the third place finisher, but this result conveys no information about the spacing among the racers at the finish line. The middle value in a sorted list is the *median*.
- *Interval* level data conveys order information, but in addition, the values that subdivide the scale are equally spaced. This makes it meaningful to calculate the distance between values, the *mean* or *average* value (the value for which the sum of its absolute distances to each other value is zero), the standard deviation, and other descriptive statistics about the data.

exceptions and not let them distort the choice of the organizing principles in the “logic tier.”

- *Ratio* level data is interval data with a fixed zero point, which makes assertions about proportions meaningful. \$10,000 is twice as much as \$5,000.

These distinctions are data type and levels of measurement are often strongly identifiable with business functions: operational, transactional, process control, and predictive analytics activities require the most fine-grained data and quantitative measurement scales, while strategic functions might rely on more qualitative analyses represented in narrative text formats.

Just as there are many laws and regulations that restrict the organization of physical resources, there are laws and regulations that constrain the arrangements of digital ones. Many information systems that generate or collect transactional data are prohibited from sharing any records that identify specific people. Banking, accounting, and legal organizing systems are made more homogeneous by compliance and reporting standards and rules.

#### Organizing Mental Resources

Memories can be viewed either as physical (because at some level they are represented in the brain) or as digital (because they are retrieved as electrical impulses), but memory techniques like the method of loci and memory palaces reify this duality in an interesting way.

While physical resources must be stored in physical locations, our powerful spatial memory provides an opportunity for us to, in a sense, store mental resources in physical locations. Our hippocampus, the brain component dedicated to memory, is highly developed for storing and recalling memories of physical locations.



The ancient Greeks relied on this capability and devised a mnemonic system—the method of loci—which involved attaching things to remember, the key ideas in a speech perhaps, to well-known physical locations. While giving the speech, then, all one must do is imagine walking through that physical location from idea to idea. Today, champion memorizers use this technique to associate items with places in vividly imagined “memory palaces.” While you may not be interested in memorizing the order of a deck of cards, recognizing the power of our spatial memory may be worth considering when designing your organizing system or when analyzing the successes or failures of a system.<sup>20</sup>

20. ([Spence 1985](#)) This memory technique has continued to be used since, and in addition to being found in tips for studying and public speaking, is applied in memorization competitions. For example, journalist and author Joshua Foer, in his book on memory and his journey from beginner to winning the 2006 U.S. Memory Championship ([Foer 2011](#)), wrote that Scott Hagwood, a four-time winner of the same competition, used locations in Architectural Digest to place his memories.

## Organizing Web-based Resources

The Domain Name System(DNS) is the most inherent scheme for organizing web resources. Top-level domains for countries (.us, .jp, .cn, etc.) and generic resource categories (.com, .edu, .org, gov, etc.) provide some clues about the resources organized by a website. These clues are most reliable for large established enterprises and publishers; we know what to expect at `ibm.com`, `Berkeley.edu`, and `sfgov.org`.<sup>21</sup>

The network of hyperlinks among web resources challenges the notion of a collection, because it makes it impractical to define a precise boundary around any collection smaller than the complete web.<sup>22</sup>

21. The Domain Name System(DNS) ([Mockapetris 1987](#)) is the hierarchical naming system that enables the assignment of meaningful domain names to groups of Internet resources. The responsibility for assigning names is delegated in a distributed way by the *Internet Corporation for Assigned Names and Numbers*(ICANN) (<http://www.icann.org>). DNS is an essential part of the Web's organizing system but predates it by almost twenty years.
22. [HTML5](#) defines a “manifest” mechanism for making the boundary around a collection of web resources explicit even if somewhat arbitrary to support an “offline” mode of interaction in which all needed resources are continually downloaded (<http://www.w3.org/TR/html5/>

Furthermore, authors are increasingly using “web-native” publication models, creating networks of articles that blur the notions of articles and journals. For example, scientific authors are interconnecting scientific findings with their underlying research data, to discipline-specific data repositories, or to software for analyzing, visualizing, simulation, or otherwise interacting with the information.<sup>23</sup>

The conventional library is both a collection of books and the physical space in which the collection is managed. On the web, rich hyper linking and the fact that the actual storage location of web resources is unimportant to the end users fundamentally undermine the idea that organizing systems must collect resources and then arrange them under local control to be effective. The spectacular rise during the 1990s of the AOL “walled garden,” created on the assumption that the open web was unreliable, insecure, and pernicious, was for a time a striking historical reminder and warning to designers of closed resource collections until its equally spectacular collapse in the following decade.<sup>24</sup> But Facebook so far is succeeding by following a walled garden strategy.

### *“Information Architecture” and Organizing Systems*

The discipline known as [information architecture](#) can be viewed as a specialized approach for designing the information models and

[browsers.html#offline](#)), but many people consider it unreliable and subject to strange side effects.

23. ([Aalbersberg and Kahler 2011](#)).

24. ([Munk 2004](#)).

their systematic manifestations in user experiences on websites and in other information-intensive organizing systems.<sup>25</sup> Abstract patterns of information content or organization are sometimes called architectures, so it is straightforward from the perspective of the discipline of organizing to define the activity of *information architecture* as designing an abstract and effective organization of information and then exposing that organization to facilitate navigation and information use. Note how the first part of this definition refers to intentional arrangement of resources, and the second to the interactions enabled by that arrangement.

Our definition of information architecture implies a methodology for the design of user interfaces and interactions that puts conceptual modeling at the foundation. Best practices in information architecture emphasize the use of systematic principles or design patterns for organizing the resources and interactions in user interfaces. The logical design is then translated into a graphical design that arranges windows, panes, menus, and other user interface components. The logical and graphical organization of a user interface together affect how people interact with it and the actions they take (or do not take).

25. This definition of [information architecture](#) combines those in a Wikipedia article ([http://en.wikipedia.org/wiki/Information\\_architecture](http://en.wikipedia.org/wiki/Information_architecture)) and in a popular book with the words in its title ([Morville and Rosenfield 2006](#)). Given the abstract elegance of “information” and “architecture” any definition of “information architecture” can seem a little feeble.

See ([Resmini and Rosati 2011](#)) for a history of information architecture.

### The Activities of Information Architecture

IA is a relatively new field, but the ubiquity of the web and information-intensive applications that must implement many types of user interactions has inspired many conceptual and methodological innovations. Here are some of them.

**Selecting Resources:** To make good choices about what content to include in an information system or service, methods and tools for creating and organizing the information that is potentially available are important. Glushko and McGrath's method for creating a "Document Inventory" and Halvorson and Rach's "Information Inventory" both use a matrix or grid format to list information sources and various associated properties. Once the inventory is completed, the information must be evaluated with respect to the user and information requirements. This usually requires a more fine-grained analysis to choose the most reliable or reusable source when there are alternatives. This process is usually called content auditing, and tools or templates for organizing the work are easy to find on the web.

**Organizing Resources:** Tidwell proposes a set of design patterns for input forms, text and graphic editors, information graphics, calendars, and other common types of web applications that organize resources. Morville and Rosenfield classify design patterns as "organization schemes" and "organization structures," reinforcing the idea that information architecture is a sub-specialty of the discipline of organizing.

**Designing Interactions:** Kalbach presents design patterns and implementations for navigation interactions. Resmini and Rosati discuss architectures and examples for information architectures that interconnect physical and digital channels. Marcotte introduces techniques for adapting user interfaces to the size and capabilities of different devices, collectively called responsive web design.

Information architects use a variety of tools for representing information and process models. Common ones include site maps, workflow and dataflow diagrams, and wireframe models. Brown's [Communicating Design](#) and Abel and Baillie's [The Language of Content Strategy](#) are concise sources.<sup>26</sup>

Some information design conventions have become design patterns. Documents use headings, boxes, white space, and horizontal rules to organize information by type and category. Large type signifies more important content than small type, red type indicates an advisory or warning, and italics or bold says “pay attention.”

Some patterns are general and apply to an entire website, page, or interface genre such as a government site, e-commerce site, blog, social network site, home page, “about us” page, and so on. Other patterns are more specific and affect a part of a site or a

26. See ([Halvorson and Rach 2012](#)), ([Tidwell 2008](#)), ([Morville and Rosenfield 2006](#)), ([Kalbach 2007](#)), ([Resmini and Rosati 2011](#)), ([Marcotte 2011](#)), ([Brown 2010](#)), ([Abel and Baillie 2014](#))

single component of a page (e.g., autocompletion of a text field, breadcrumb menu, slideshow).

In websites, different categories of content or interactions are typically arranged in different menus. The choices within each menu are then arranged to reflect typical workflows or ordered according to some commonly used property like size, percentage, or price.

All design patterns reflect and reinforce the user's past experiences with content and interface components, and this familiarity reduces the cognitive complexity of user interface interaction, requiring users to pay less attention.<sup>27</sup>

However, interface designers can take advantage of this familiarity and employ design patterns in a less beneficial way to manipulate users, control their behaviors, or trick them into taking actions they do not intend. Patterns used this way are sometimes called [Dark Patterns](#).

#### Dark Patterns

Some websites and applications employ Dark Patterns, which rely on user familiarity with good design patterns to induce users to take actions or fail to take actions in ways counter to their best interests. For example, a website may exploit familiar patterns to induce users to click on an ad disguised as a news item,

27. Some popular collections of design patterns are [\(Van Duyne et. al, 2006\)](#), [\(Tidwell 2010\)](#), and <http://ui-patterns.com/>

sign up for unwanted e-mails, disclose personal information, or ignore important terms and conditions because they are buried in tiny text or in unusual locations.

[Darkpatterns.org](http://Darkpatterns.org) collects and classifies dark patterns. The largest categories are “bait and switch” (suggest one action but cause another), “trick questions” (misleading phrasing of an option), and “misdirection” (focusing attention on one thing to distract from another). The website has numerous examples of interfaces that try to get users to install additional software or change their defaults to a company’s product during installation. Other examples are from commerce sites that conceal the cheapest options, add additional fees at the very end of the purchase process, or make it difficult to accurately compare costs.

These practices are enough of a concern that some governments have begun to regulate the information that must be provided to consumers when purchasing digital products. The Directive on Consumer Rights published by the European Commission in June 2014 contains instructions about design choices that should be avoided, such as allowing additional purchases and payments without the consumer’s consent. The Directive even includes a model set of patterns to help designers comply with it.<sup>28</sup>

28. The Directives can be found at <http://ec.europa.eu/>



Dark patterns can be used to manipulate interactions with physical resources too. Gas pumps with three or four grades of gasoline invariably arrange the pumps in order of price, with the cheapest gas at the left and the most expensive on the right. Some gas stations put the cheapest gas in the middle, which causes inattentive customers who are relying on the usual pattern to buy more expensive gas than they intended.

Many organizing systems need to support interactions to find, identify, and select resources. Some of these systems contain both physical and digital resources, as in a bookstore with both web and physical channels, and many interactions are implemented across more than one device. Both the cross-channel and multiple-device situations create user expectations that interactions will be consistent across these different contexts. Starting with a conceptual model and separating content and structure from presentation, as we discussed in [“The Concept of “Organizing Principle”](#)”, gives organizing systems more implementation alternatives and makes them more robust in the face of technology diversity and change.

A model-based foundation is also essential in information visualization applications, which depict the structure and relationships in large data collections using spatial and graphical conventions to enable user interactions for exploration and analysis. By transforming data and applying color, texture, density, and other

[consumers/consumer\\_rights/rights-contracts/  
directive/index\\_en.htm](https://consumers.consumer_rights/rights-contracts/directive/index_en.htm)

properties that are more directly perceptible, information visualization applications enable people to obtain more information than they can from text displays.<sup>29</sup>

Some designers of information systems put less emphasis on conceptual modeling as an “inside-out” foundation for interaction design and more emphasis on an “outside-in” approach that highlights layout and other presentation-tier considerations with the goal of making interactions easy and enjoyable. This focus is typically called user experience design, and information architecture methods remain an important part of it, but not beginning with explicit organizing principles implies more heuristic methods and yields less predictable results.

## Organizing With Descriptive Statistics

Descriptive statistics, about a collection or dataset, summarize it concisely and can identify the properties that might be most useful as organizing principles. The simplest statistical description of a collection is how big it is; how many resources or observations does it contain?

Descriptive statistics summarize a collection of resources or dataset with two types of measures:

- Measures of *central tendency*: Mean, median, and mode; which measure is appropriate depends on the level of measurement represented in the numbers being described (these measures

29. The classic text about information visualization is [\*The Visual Display of Quantitative Information\* \(Tufte 1983\)](#). More recent texts include [\*\(Few 2012\)\*](#) and [\*\(Yau 2011\)\*](#).

and the concept of levels of measurements are defined in [“Organizing Digital Resources”](#)).

- Measures of *variability*: Range (the difference between the maximum and minimum values), and standard deviation (a measure of the spread of values around the mean).

Statistical descriptions can be created for any resource property, with the simplest being the number of resources that have the property or some particular value of it, such as the number of times a particular word occurs in a document or the number of copies a book has sold. Comparing summary statistics about a collection with the values for individual resources helps you understand how typical or representative that resource is. If you can compare your height of 6 feet,  $\frac{1}{2}$  inch with that of the average adult male, which is 5 feet, 10 inches, the difference is two and a half inches, but what does this mean? It is more informative to make this comparison using the standard deviation, which is three inches, because this tells you that 68% of adult men have heights between 5 feet, 7 inches and 6 feet, 1 inch. When measurements are normally distributed in the familiar bell-shaped curve around the mean, the standard deviation makes it easy to identify statistical outliers.

No matter how measurements are distributed, it can be useful to employ descriptive statistics to organize resources or observations into categories or quantiles that have the same number of them. Quartiles (4 categories), deciles (10), and percentiles (100) are commonly used partitions.

Alternatively, resources or observations can be organized by visualizing them in a histogram, which divides the range of values into units with equal intervals. Because values tend to vary around some central tendency, the intervals are unlikely to contain the same number of observations. Descriptive statistics and associated visualizations can suggest which properties make good organizing principles because they exhibit enough variation to distinguish

resources in their most useful interactions. For example, it probably isn't useful to organize books according to their weight because almost all books weigh between  $\frac{1}{2}$  and 2 pounds, unless you are in the business of shipping books and paying according to how much they weigh.

### *Exploratory Analysis to Understand Data*

Many experts recommend that data analysts should undertake some exploratory analysis with descriptive statistics and simple information visualizations to understand their data before applying sophisticated computational techniques to the dataset. In particular, because the human visual system quickly perceives shapes and patterns, analyzing and graphing the values of data attributes and other resource descriptions can suggest which properties might be useful and comprehensible organizing principles. In addition, data visualization makes it easy to recognize values that are typical or that are outliers. Some of this analysis might form part of data quality assessment during resource selection, but if not done then, it should be done as part of the organizing process.

A dataset whose fields or attributes lack information about data types and units of measure has little use because the data lacks meaning. When some, but not all parts of the data are named or annotated, avoid over-interpreting these descriptions' meanings. (See [“Naming Resources”](#).)

We will do some exploratory analysis to understand what an example dataset contains and how we might use it. For our example, we consider a collection of a few hundred records from a healthcare study, whose first eight records and first five data fields in each record are shown in [Figure: Example Dataset](#).

Example Dataset

| ID  | Sex | Temp | Age | Weight |
|-----|-----|------|-----|--------|
| 1   | 1   | 97.6 | 32  | 135    |
| 2   | 0   | 97.6 | 19  | 118    |
| 3   | 0   | 97.6 | 23  | 128    |
| 4   | 1   | 98.7 | 34  | 140    |
| 5   | 1   | 98.5 | 52  | 162    |
| 6   | 1   | 98.7 | 60  | 160    |
| 7   | 0   | 98.3 | 36  | 148    |
| 8   | 0   | 98.3 | 38  | 155    |
| ... | ... |      |     |        |
| 260 | 1   | 99.0 | 23  | 123    |

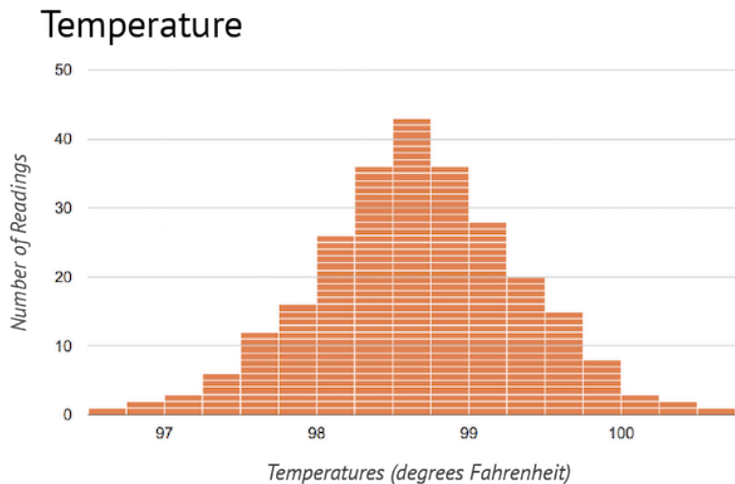
The “ID” column contains numeric data, but every value is a different integer, and the values are contiguous. The field label “ID” suggests that this is the resource identifier for the participants in the healthcare study. Further examination of other tables will reveal that this is a key value that points into a different dataset containing the resource names.

The “Sex” column is also numeric, but there are only two different values, 0 and 1, and in the complete dataset they are approximately equal in frequency. This attribute seems to be categorical or Boolean data. This makes sense for a “Sex” categorization, and it is likely to prove useful in understanding the dataset.

## Histogram

A histogram is the simplest visualization of one-dimensional data. It is a bar graph that takes the full range of values, organizes them into a set of intervals of equal size on one axis, and then counts the number of values in each interval on the other axis.

The “Temp” column contains several hundred different numeric values in the complete dataset, ranging from 96.8 to 100.6, with a mean of 98.6. These values are sensible if the label “Temp” means the under-the-tongue body temperature in degrees Fahrenheit of the study participant when the other measures were obtained. This type of data is usefully viewed as a histogram to get a sense of the spread and shape, shown in [Figure: Temperature](#).



The data values of the “Temp” column follow the familiar normal or bell-shaped distribution, for which simple and useful descriptive statistics are the mean and the standard deviation. The mean (or

average) is at the center of the distribution, and the standard deviation captures the width of the bell shape. In this dataset, the very narrow range of data values here suggests that this attribute is not useful as an organizing principle, since it does not distinguish the resources in any significant way. In a larger sample, however, there might be a few very low or very high temperatures, and it would be useful to investigate these “hypothermic” or “hyperthermic” outliers.

#### Median versus Average

If ten people are in a bar, all of whom make \$50,000 a year, when a movie star who made \$25,000,000 this year walks in, the average income is now \$2.3 million. The median income is still \$50,000.

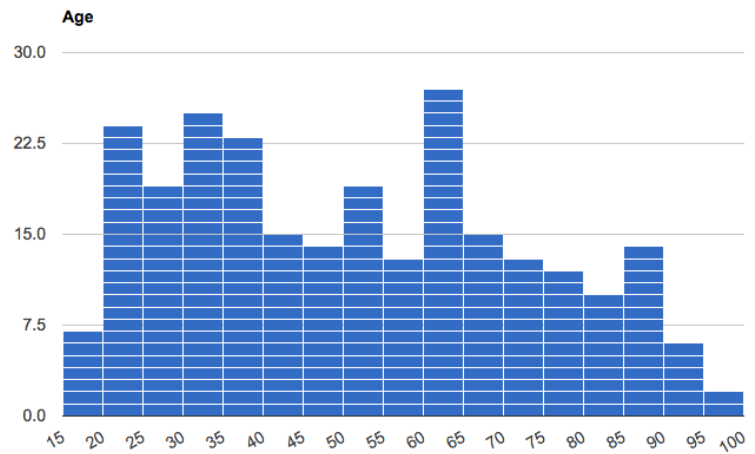
[\*The End of Average\*](#) tells the story of how the U.S. military designed aircraft cockpits beginning in 1926 on the basis of the average dimensions of a 1926 pilot. In 1950, researchers measured over four thousand pilots only to discover that no actual pilot had average values on all the measures, and recommended adjustable seats and controls in cockpit design.<sup>30</sup>

The data values of the “Age” column range from 18 to 97, and are spread broadly across the entire range; this is the age, in years, of the study participants. When a distribution is very broad and flat, or highly skewed with many values at one end or another, the mean value is less useful as a descriptive statistic. Instead of the mean, it is

30. [\(Rose 2016\)](#)

better to use the median or middle value as a summary of the data; the median value for “Age” in the complete dataset is 39.

Age



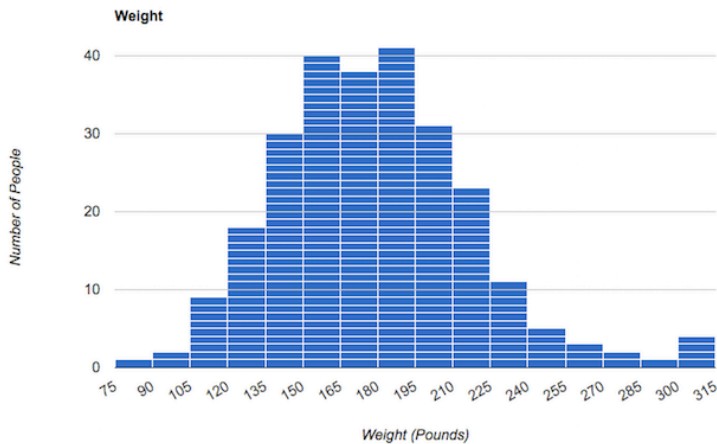
The “Weight” column has about 220 different numeric values, from 82 to 300, and judging from this range we can infer that the weights are measured in pounds. The data follows an uneven distribution with peaks around 160 and 200, and a small peak at 300. This odd shape appears in the histogram of [Figure: Weight](#). The two peaks in this so-called multi-modal histogram suggest that this measure is mixing two different kinds of resources, and indeed it is because weights of men and women follow different distributions. It would thus be useful to use the categorical “Sex” data to separate these populations, and [Figure: Sex and Weight: Female](#) shows how analyzing weight for women and men as different populations is much more informative as an organizing principle than combining them.

What about the odd peak in the distribution at 300? End of range anomalies like this generally reflect a limitation in the device or

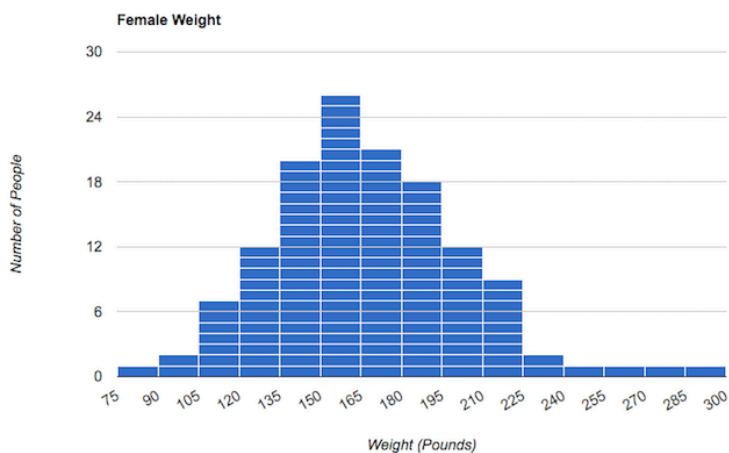


system that created the data. In this case, the weight scale must have an upper limit of 300 pounds, so the peak represents the people whose weight is 300 or greater.

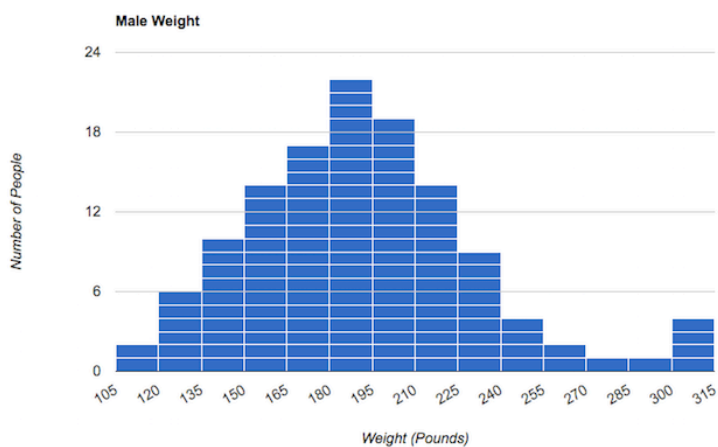
Weight



Sex and Weight: Female



## Sex and Weight: Male



## Detecting Errors and Fraud in Data

There are numerous techniques for evaluating individual data items

or datasets to ensure that they have not been changed or corrupted during transmission, storage, or copying. These include parity bits, check digits, check sums, and cryptographic hash functions. They share the idea that a calculation will yield some particular value or match a stored result when the original data has not been changed. Another basic technique for detecting errors is to look for data values that are different or anomalous because they do not fall into expected ranges or categories.

More interesting challenges arise when the data might have been changed by intentional actions to commit fraud, launder money, or carry out some other crime. In these situations, the person tampering with data or creating fake data will try to make the data look normal or expected.

Forensic accountants and statisticians use many techniques for detecting possibly fraudulent data in these adversarial contexts. Some are quite simple:

- If expenses are reimbursed up to some maximum allowed value, look for data items with that exact value.
- When any value exceeding some threshold triggers more careful analysis, look for other data items just below that threshold.
- When invoices or claims are paid on receipt, and only a sample are subsequently audited, look for duplicate submissions.
- Calculate the ratio of the maximum to the minimum value for purchases in some category (such as the unit price paid for items from suppliers); items with large ratios might indicate fraud where the supplier “kicks back” some of the money to the purchaser.

Benford's Law, the observation that the leading digits in data sets are distributed in a non-uniform manner, is an effective technique

for detecting fraudulent data because it is based on a counter-intuitive fact not known to most fraudsters, who often make up data to look random. You might think that the number 1 would occur 11% of the time as the first digit (since there are 9 possibilities), but for data sets whose values span several orders of magnitude, the number 1 is the first digit about 30% of the time, and 7, 8, and 9 occur around 5%.

Because of the very high transaction rate and the relatively small probability of fraud, credit card fraud is detected using machine learning algorithms. The classifier is trained with known good and bad transactions using properties like average amount, frequency, and location to develop a model of each cardholder's "data behavior" so that a transaction can quickly be assigned a probability that it is fraudulent. (More about this kind of computational classification in [\*Categorization: Describing Resource Classes and Types\*](#).)<sup>31</sup>

## Organizing with Multiple Resource Properties

Multiple properties of the resources, the person organizing or intending to use them, and the social and technological environment in which they are being organized can collectively shape their organization. For example, the way you organize your home kitchen is influenced by the physical layout of counters, cabinets, and drawers; the dishes you cook most often; your skills as a cook, which may influence the number of cookbooks, specialized

31. See [https://chapters.theiia.org/ottawa/Documents/Digital\\_Analysis.pdf](https://chapters.theiia.org/ottawa/Documents/Digital_Analysis.pdf) for a short introduction to data analysis for fraud detection. See [\(Durtschi et al 2004\)](#) for the use of Benford's Law in forensic accounting.

appliances and tools you own and how you use them; the sizes and shapes of the packages in the pantry and refrigerator; and even your height.

If multiple resource properties are considered in a fixed order, the resulting arrangement forms a *logical hierarchy*. The top level categories of resources are created based on the values of the property evaluated first, and then each category is further subdivided using other properties until each resource is classified in only a single category. Consider the hierarchical system of folders used by a professor to arrange the digital resources on his computer; the first level distinguishes personal documents from work-related documents; work is then subdivided into teaching and research, teaching is subdivided by year, and year divided by course.

For physical resources, mapping categories to physical locations is another required step; for example, resources in the “kitchen utensils” category might all be arranged in drawers near a workspace, with “silverware” arranged more precisely to separate knives, forks, and spoons.

An alternative to hierarchical organization that is often used in digital organizing systems is [\*faceted classification\*](#), in which the different properties for the resources can be evaluated in any order. For example, you can select wines from the wine.com store catalog by type of grape, cost, or region and consider these property facets in any order. Three people might each end up choosing the same moderately-priced Kendall Jackson California Chardonnay, but one of them might have started the search based on price, one based on the grape varietal, and the third with the region. This kind of interaction in effect generates a different logical hierarchy for every different combination of property values, and each user made his final selection from a different set of wines.

[\*Faceted classification\*](#) allows a collection of description resources to be dynamically re-organized into as many categories as there are combinations of values on the descriptive facets, depending on the

priority or point of view the user applies to the facets. Of course this only works because the physical resources are not themselves being rearranged, only their digital descriptions.

Applications that organize large collections of digital information, including those for search, natural language processing, image classification, personalized recommendation, and other computationally intensive domains, often use huge numbers of resource properties (which are often called “features” or “dimensions”). For example, in document collections each unique word might initially be treated as a feature by machine learning algorithms, so there might be tens of thousands of features.

[Classification: Assigning Resources to Categories](#) explains principles and methods for hierarchical and faceted classification in more detail.

## 22. Designing Resource-based Interactions

We need to focus on the interactions that are enabled because of the intentional acts of description or arrangement that transform a collection of resources into an organizing system. With physical resources, it is easy to distinguish the interactions that are designed into and directly supported by an organizing system because of intentional acts of description or arrangement from those that can take place with resources after they have been accessed. For example, when a book is checked out of a library it might be read, translated, summarized, criticized, or otherwise used—but none of these interactions would be considered a capability of the book that had been designed into the library. Some physical resources can initiate interactions, as surely “human resources” and “smart” objects with sensors and other capabilities can, but most physical resources are passive. We will discuss this idea of resource [agency](#) in “[Resource Agency](#)”.

In contrast, in organizing systems that contain digital resources the logical boundary between the resources and their interactions is less clear because what you can do with a digital resource is often not apparent. Furthermore, some of the interactions that are outside of the boundary with physical resources can be inside of it with digital ones. For example, when you check a printed book out of the library, it is no longer in the library when you translate it. But a digital book in the Google Books library is not removed when you start reading it, and a language translation service runs “inside” of it.

Additional issues in the design of interactions with resources are whether users have direct or mediated access to the resources, and whether they interact with the resources themselves or only

with copies or descriptions of them. For example, users have direct access to original resources in a collection when they browse through library stacks or wander in museum galleries.<sup>1</sup> Users have mediated or indirect access when they use catalogs or search engines. Because digital resources can be easily reproduced, it can be difficult to distinguish a copy from the original, which raises questions of authenticity we will discuss in [“Authenticity”](#).

## Affordance and Capability

The concept of *affordance*, introduced by J. J. Gibson, then extended and popularized by Donald Norman, captures the idea that physical resources and their environments have inherent actionable properties that determine, in conjunction with an actor’s capabilities and cognition, what can be done with the resource.<sup>2</sup>

1. Except when the resources on display are replicas of the originals, which is more common than you might suspect. Many nineteenth-century museums in the United States largely contained copies of pieces from European museums. Today, museums sometimes display replicas when the originals are too fragile or valuable to risk damage ([Wallach 1998](#)). Whether the “resource-based interaction” is identical for the replica and original is subjective and depends on how well the replica is implemented.
2. ([Gibson 1977](#)), ([Norman 1988](#)). See also ([Norman 1999](#)) for



Including capabilities and cognition brings accessibility considerations into the definition of affordance. A resource is only accessible when it supports interactions, and it is ineffective design to implement interactions with resources that some people are unable to perform. A person who cannot see text cannot read it, or if they are confined to a wheelchair they cannot select a book from a tall library shelf. Describing or transforming resources to ensure their accessibility is discussed in greater detail in [“Accessibility”](#).

When organizing resources involves arranging physical resources using boxes, bins, cabinets, or shelves, the affordances and the implications for access and use can be easily perceived. Resources of a certain size and weight can be picked up and carried away. Books on the lower shelves of bookcases are easy to reach, but those stored ten feet from the ground cannot be easily accessed.

We can analyze the organizing systems with physical resources to identify the affordances and the possible interactions they imply. We can compare the affordances or overall interaction [capability](#) enabled by different organizing systems for some type of physical resources, and we often do this without thinking about it. The tradeoffs between the amount of work that goes into organizing a collection of resources and the amount of work required to find and use them are inescapable when the resources are physical objects or information resources are in physical form. We can immediately see that storing information on scrolls does not enable the random access capability that is possible with books.

What and how to count to compare the capabilities of organizing systems becomes more challenging the further we get from collections of static physical resources, like books or shoes, where it

a short and simple explanation of Norman’s (re-)interpretation of Gibson.

is usually easy to perceive and understand the possible interactions. With computers, information systems, and digital resources in general, considerations about affordances and capabilities are not as straightforward.

First, the affordances we can perceive might not be tied to any useful interaction. Donald Norman joked that every computer screen within reaching distance affords touching, but unless the display is touch-sensitive, this affordance only benefits companies that sell screen-cleaning materials.<sup>3</sup>

Second, most of the interactions that are supported by digital resources are not apparent when you encounter them. You cannot tell from their names, but you probably know from past experience what interactions are possible with files of types “.doc” and “.pdf.” You probably do not know what interactions take place with “.xpi” and “.mobi” files.<sup>4</sup>

A similar difficulty exists when we look at resource descriptions

3. ([Norman 1999, p. 39](#)).

4. The “.xpi” file type is used for Mozilla/Firefox browser extensions, small computer programs that can be installed in the browser to provide some additional user interface functionality or interaction. The “.mobi” file type was originally developed to enable better document display and interactions on devices with small screens. Today its primary use is as the base ebook format for the Amazon Kindle, except that the Kindle version is more highly compressed and locked down with digital rights management.

and data collections, where we often cannot tell just by examining their values what kinds of interactions and operations with them are sensible. Think of all the different kinds of information that might be associated with a collection of people like the students in a university. A database might contain student names, student IDs, gender, birth dates, addresses, a numeric code for academic major, course units completed, grade point average, and other information. These pieces of information differ in their data type; some are integers, some are real numbers, some are Boolean, and some are just text strings. The numeric data also differs in the level of measurement it represents. Student IDs and the academic major codes are nominal data, the house or apartment number in the address is ordinal data, and the course units and grade point average are interval data. Data type and level of measurement influence the kind of interactions that are meaningful; we can create an alphabetical list of students using their last names, count up the number of students with the same academic major, and calculate the average GPA or units completed. But it makes no sense to use the numeric codes for academic major to compute an average major.

Once you have discovered it, the [capability](#) of digital resources and information systems can be assessed by counting the number of functions, services, or application program interfaces. However, this very coarse measure does not take into account differences in the capability or generality of a particular interaction. For example, two organizing systems might both have a search function, but differences in the operators they allow, the sophistication of pre-processing of the content to create index terms, or their usability can make them vastly differ in power, precision, and effectiveness.<sup>5</sup>

An analogous measure of *functional capability* for a system with

5. See ([Hearst 2009](#)), ([Buettcher et al. 2010](#)).

dynamic or living resources is the *behavioral repertoire*, the number of different activities, or range of actions, that can be initiated.

We should not assume that supporting more types of interactions necessarily makes a system better or more capable; what matters is how much value is created or invoked in each interaction. A smartphone cluttered with features and apps you never use enables a great many interactions, but most of them add little value. Doors that open automatically when their sensors detect an approaching person do not need handles or require explicit interactions. Organizing systems can use stored or computed information about user preferences or past interactions to anticipate user needs or personalize recommendations. This has the effect of substituting information for interaction to make interactions unnecessary or simpler.

For example, a “smart travel agent” service can use a user’s appointment calendar, past travel history, and information sources like airline and hotel reservation services to transform a minimal interaction like “book a business trip to New York for next week’s meeting” into numerous hidden queries that would have otherwise required separate interactions. These queries are interconnected by logical or causal dependencies that are represented by information that overlaps between them. For example, all travel-related services (airlines, hotels, ground transportation) need the traveler’s identity and the time and location of his travel. A New York trip might involve all of these services, and they need to fit together in time and location for the trip to make sense. The hotel reservation needs to begin the day the flight arrives in the destination city, the limousine service needs to meet the traveler shortly after the plane lands, and the restaurant reservation should be convenient in time and location to the hotel.<sup>6</sup>

6. ([Glushko and Nomorosa 2013](#)).

## Interaction and Value Creation

A useful way to distinguish types of interactions with resources is according to the way in which they create value, using a classification proposed by Apte and Mason. They noted that interactions differ not just in their overall intensity but in the absolute and relative amounts of physical manipulation, interpersonal or empathetic contact, and symbolic manipulation or information exchange involved in the interaction.

Furthermore, Apte and Mason recognized that the proportions of these three types of value creating activities can be treated as design parameters, especially where the value created by retrieving or computing information could be completely separated from the value created by physical actions and person-to-person encounters. This configuration of value creation enables automated self-service, in which the human service provider can be replaced by technology, and outsourcing, in which the human provider is separated in space or time from the customer.<sup>7</sup>

### *Value Creation with Physical Resources*

Physical manipulation is often the intrinsic type of interaction with collections of physical resources. The resource might have to be handled or directly perceived in order to interact with it, and often the experience of interacting with the resource is satisfying or entertaining, making it a goal in its own right. People often visit museums, galleries, zoos, animal theme parks or other institutions

7. ([Apte and Mason 1995](#)) introduced this framework to analyze services rather than interactions *per se*.

that contain physical resources because they value the direct, perceptual, or otherwise unmediated interaction that these organizing systems support.

Physical manipulation and interpersonal contact might be required to interact with information resources in physical form like the printed books in libraries.

A large university library contains millions of books and academic journals, and access to those resources can require a long walk deep into the library stacks after a consultation with a reference librarian or a search in a library catalog. For decades library users searched through description resources—first printed library cards, and then online catalogs and databases of bibliographic citations—to locate the primary resources they wanted to access. The surrogate descriptions of the resources needed to be detailed so that users could assess the relevance of the resource without expending the significant effort of obtaining and examining the primary resource.<sup>8</sup>

However, for most people the primary purpose of interacting with a library is to access the information contained in its resources. Many people prefer accessing digital documents or books to accessing the original physical resource because the incidental physical and interpersonal interactions are unnecessary. In addition, many library searches are for known items, which is easily supported by digital search.<sup>9</sup>

8. Furthermore, many of the resources might not be available in the user's own library and could only be obtained through inter-library loan, which could take days or weeks.

9. In contrast, far fewer interactions in museum collections

In some organizing systems robotic devices, computational processes, or other entities that can act autonomously with no need for a human agent carry out interactions with physical resources. Robots have profoundly increased efficiency in materials management, “picking and packing” in warehouse fulfillment, office mail delivery, and in many other domains where human agents once located, retrieved, and delivered physical resources. A “library robot” system that can locate books and grasp them from the shelves can manage seven times as many books in the same space used by conventional open stacks.<sup>10</sup>

are searches for known items, and serendipitous interactions with previously unknown resources are often the goal of museum visitors. As a result, few museum visitors would prefer an online visit to experiencing an original painting, sculpture, or other physical artifact. However, it is precisely because of the unique character of museum resources that museums allow access to them but do not allow visitors to borrow them, in clear contrast to libraries.

10. ([Viswanadham 2002](#)), ([Madrigal 2009](#)), ([Prats et al. 2008](#)). A video of a robot librarian in action at the University of Missouri, Kansas City is at <http://www.youtube.com/watch?v=8wJJLlTq7ts>.

See also the Popular Science article How It Works: Underground Robot Library available at <http://www.popsoci.com/content/underground-robot-library>.

Interactions with physical resources often have highly tangible results; in the preceding examples of fulfillment and delivery interactions, resources move from one location to another. However, an abstract or architectural perspective on interaction design and value creation can create more flexibility in carrying out the interactions while still producing the expected value for the user. In general, more abstract descriptions of interactions and services allow for transparent substitution of the implementation, potentially enabling a computational process to be a substitute for one carried out by a person, or vice versa.

For example, a user buying from an internet-based store need not know and probably does not care which service delivers the package from the warehouse. Presenting the interaction to the shopper as the “delivery service” rather than as a “FedEx” or “UPS” service allows the retailer to choose the best service provider for each delivery. Going even further, if you need printed documents at a conference, sales meeting, or anywhere other than your current location, the interaction you desire is “provide me with documents” and not “deliver my documents.” It does not matter that FedEx will print your documents at their destination rather than shipping them there.



Library Robot





An automated robot library system at San Francisco State University.

The automated robot library system installed by the Dematic Group stores books in bins stacked on three-story-tall metal racks in five long aisles. Instead of using a library classification scheme, books are stored according to their sizes in one-foot deep metal bins, which contain about one hundred books each. Given an online catalog request for a book, the system looks up the bin where it was last stored, and then directs a robot to bring that bin to the circulation desk. Human librarians then find the requested book in the bin and scan its barcode, which notifies the requester that the book can be picked up. To store a book, the librarian scans its barcode, and it is then stored in the closest bin with available space.

(Photo by Scott Abel. Used with permission.)

## *Value Creation with Digital Resources*

With digital resources, neither physical manipulation nor interpersonal contact is required, and the essence of the interaction is information exchange or symbolic manipulation of the information contained in the resource.<sup>11</sup> Put another way, by replacing interactions that involve people and physical resources with symbolic ones, organizing systems can lower costs without reducing user satisfaction. This is why so many businesses have automated their information-intensive processes with self-service technology.

Similarly, web search engines eliminate the physical effort required to visit a library and enable users to consult more readily accessible digital resources. A search engine returns a list of the page titles of resources that can be directly accessed with just another click, so it takes little effort to go from the query results to the primary resource. This reduces the need for the rich surrogate descriptions that libraries have always been known for because it enables rapid evaluation and iterative query refinement.<sup>12</sup>

11. Providing access to knowledge is a core mission of libraries, and it is worth pointing out that library users obtain knowledge both from the primary resources in the library collection and from the organizing system that manages the collection.
12. It also erodes the authority and privilege that apply to resources because they are inside the library when a web search engine can search the “holdings” of the web

### Stop and Think: Browsing for Books

How does the experience of browsing for books in a library or bookstore compare with browsing using a search engine? What aspects are the same or similar? What aspects are different?

The ease of use and speed of search engines in finding web resources creates the expectation that any resource worth looking at can be found on the web. This is certainly false, or Google would never have begun its ambitious and audacious project to digitize millions of books from research libraries. While research libraries strive to provide access to authoritative and specialized resources, the web is undeniably good enough for answering most of the questions ordinary users put to search engines, which largely deal with everyday life, popular culture, personalities, and news of the day.

Libraries recognize that they need to do a better job integrating their collections into the “web spaces” and web-based activities of their users if they hope to change the provably suboptimal strategies of “information foraging” most people have adopted that rely too much on the web and too little on the library.<sup>13</sup> Some libraries are experimenting with [Semantic Web](#) and “Linked Data”

faster and more comprehensively than you can search a library’s collection through its online catalog.

13. ([Pirolli 2007](#)).

technologies that would integrate their extensive bibliographic resources with resources on the open web.<sup>14</sup>

Museums have aggressively embraced the web to provide access to their collections. While few museum visitors would prefer viewing a digital image over experiencing an original painting, sculpture, or other physical artifact, the alternative is often no access at all. Most museum collections are far larger than the space available to display them, so the web makes it possible to provide access to otherwise hidden resources.<sup>15</sup>

The variety and functions of interactions with digital resources are determined by the amount of structure and semantics represented in their digital encoding, in the descriptions associated with the

14. [\(Byrne and Goddard 2010\)](#).
15. See [\(Simon 2010\)](#). An exemplary project to enhance museum access is Delphi [\(Schmitz and Black 2008\)](#), the collections browser for the Phoebe A. Hearst Museum of Anthropology at University of California, Berkeley. Delphi very cleverly uses natural language processing techniques to build an easy-to-use faceted browsing user interface that lets users view over 600,000 items stored in museum warehouses. Delphi is being integrated into Collection Space (<http://www.collectionspace.org/>), an open source web collections management system for museum collections, collaboratively being developed by University of California, Berkeley, Cambridge University, *Ontario Academy of Art and Design*(OCAD), and numerous museums.

resources, or by the intelligence of the computational processes applied to them. Digital resources can support enhanced interactions of searching, copying, zooming, and other transformations. Digital or “ebooks” demonstrate how access to content can be enhanced once it is no longer tied to the container of the printed book, but ebook readers vary substantially in their interaction repertoires; the baseline they all share is “page turning,” resizing, and full-text search.

To augment digital resources with text structures, multimedia, animation, interactive 3-D graphics, mathematical functions, and other richer content types requires much more sophisticated representation formats that tend to require a great deal of “hand-crafting.” An alternative to hand-crafted resource description is sophisticated computer processing guided by human inputs. For example, Facebook and many web-based photo organizing systems implement face recognition analysis that detects faces in photos, compares features of detected faces to features of previously identified faces, and encourages people to tag photos to make the recognition more accurate. Some online services use similar image classification techniques to bring together shoes, jewelry, or other items that look alike.

Richer interactions with digital text resources are possible when they are encoded in an application or presentation-independent format. Automated content reuse and “single-source publishing” is most efficiently accomplished when text is encoded in XML, but much of this XML is produced by transforming text originally created in word processing formats. Once it is in XML, digital information can be distributed, processed, reused, transformed, mixed, remixed, and recombined into different formats for different purposes, applications, devices, or users in ways that are almost impossible to imagine when it is represented in a tangible (and

therefore static) medium like a book on a shelf or a box full of paper files.<sup>16</sup>

Businesses that create or own their information resources can readily take advantage of the enhanced interactions that digital formats enable. For libraries, however, copyright is often a barrier to digitization, both as a matter of law and because digitization enables copyright enforcement to a degree not possible with physical resources.

As a result, digital books are somewhat controversial and problematic for libraries, whose access models were created based on the economics of print publication and the social contract of the copyright *first sale* doctrine that allowed libraries to lend printed books.<sup>17</sup>

16. Even sophisticated text representation formats such as XML have inherent limitations: one important problem that arises in complex management scenarios, humanities scholarship, and bioinformatics is that XML markup cannot easily represent overlapping substructures in the same resource ([Schmidt 2009](#)).
17. Digital books change the economics and *first sale* is not as well-established for digital works, which are licensed rather than sold ([Aufderheide and Jaszi 2011](#)). To protect their business models, many publishers are limiting the number of times ebooks can be lent before they “self-destruct.” Some librarians have called for boycotts of publishers in response (<http://boycottharpercollins.com>).

Software-based agents do analogous work to robots in “moving information around” after accessing digital resources such as web services or physical resources with sensors attached that produce digital information. Agents can control or choreograph a set of interactions with digital resources to carry out complex business processes.

## *Accessibility*

The United Nations Convention on the Rights of Persons with Disabilities recognizes accessibility to information and communications technologies as a basic human right. There is also a strong business case for accessibility: studies show that accessible websites are used more often, are easier to maintain, and produce better search results.<sup>18</sup>

In contrast to these new access restrictions imposed by publishers on digital works, many governments as well as some progressive information providers and scientific researchers have begun to encourage the reuse and reorganization of their content by making geospatial, demographic, environmental, economic, and other datasets available in open formats, as web services, or as data feeds rather than as “fixed” publications ([Bizer 2009a](#)), ([Robinson et al. 2008](#)). And we have made this book available as an open content repository so that it can be collaboratively maintained and customized.

18. We cannot explain this any better than the UN does: “The Convention follows decades of work by the United

Many of the techniques for making a resource accessible involve transforming the resource or its description into a different form so someone who could not perceive it or interact with it in its original form can now do so. The most common operating systems all come with general-purpose accessibility features such as reading text aloud, recognizing speech, magnifying text, increasing cursor size, signaling with flashing lights instead of with sounds, lights to signal keyboard shortcuts for selecting and navigating, and connecting to devices for displaying Braille. Google Translate converts text in one language to another, and many people use it to create a rough draft that is finished by a human translator.<sup>19</sup>

Other techniques are not generic and automatic, and instead require investment by authors or designers to make information accessible. Websites are more accessible when images or other non-text content types have straightforward titles, captions, and “alt text” that describes what they are about. Consistent placement and appearance of navigation controls and interaction widgets is

Nations to change attitudes and approaches to persons with disabilities. It takes to a new height the movement from viewing persons with disabilities as ‘objects’ of charity, medical treatment and social protection towards viewing persons with disabilities as ‘subjects’ with rights, who are capable of claiming those rights and making decisions for their lives based on their free and informed consent as well as being active members of society.” See

<https://www.un.org/disabilities/default.asp?navid=12&pid=150>

19. See [Microsoft Windows Accessibility](#), [Apple Accessibility](#), and [Android Accessibility Features](#).



essential; for example, in a shopping site “My Cart” might always be found at the top right corner of the page.<sup>20</sup>

If authors apply semantic and structural markup to the text and use formats that distinguish it from presentation instructions, page outlines and summaries can be generated to enhance navigation, and search can be made more precise by limiting it to particular sections or content types. As the “Information IQ” of the source format increases, more can be done to make it more accessible (see [“Resource Format”](#) and [Figure: Information IQ](#)).<sup>21</sup>

The Smithsonian Museum in Washington, DC invites visitors to record audio descriptions on mobile devices of the nearly 137 million objects in its collection, and then makes these available to everyone. This is just a small part of its efforts to make its exhibits more accessible. A company called D-Scriptive enables blind people to enjoy Broadway shows more by recording hundreds of audio descriptions that are synchronized with dialog spoken by the actors.<sup>22</sup>

20. The [Web Accessibility Initiative](#) works to make the Web accessible to people with visual, auditory, speech, cognitive, neurological, and physical disabilities.

21. [Accessible Rich Internet Applications\(ARIA\)](#)

22. Smithsonian [Guidelines for Accessible Exhibition Design](#)

Because not every performance of a Broadway is exactly the same, the D-Scriptive audio snippets are tied to particular bits of dialog. The theater's stage manager triggers the D-Scriptive system to broadcast the corresponding visual explanations to audience members

Transforming recorded spoken language to text to make it accessible and searchable is called transcription. At times transcription is necessary to comply with accessibility requirements, but is often done simply to add organization to content, as when a script is created to separate the multiple voices in a radio or television interview or story.

Transcriptions created by skilled people are highly accurate but labor-intensive to produce, so speech-to-text software is increasingly being used to transcribe speech using pre-trained acoustic and language models. Training these models is computationally intensive, and there are many clever techniques to acquire the “labeled” inputs. However, most of them are conceptually simple; they take the huge amount of data collected by voice search applications and analyze what the searcher does with the results to assess the accuracy of the transcription. Transcription accuracy can be improved when models can be specialized by industry or application. For example, speech-to-text software for doctors is trained to recognize medical terminology, while software for use by generic voice recognition services like Apple’s is trained to understand dictation and commands or questions one would ask of a smartphone.

Since text transcripts are machine-readable, unlike audio or video

listening on earpieces. For example, in the Lion King a snippet might explain that “on the left are two giraffes and a cheetah.” ([Giridharadas 2014](#))

In 2015 Netflix began a similar audio description service to accompany some of its original series. See <http://blog.netflix.com/2015/04/netflix-begins-audio-description-for.html>

files, adding text transcripts makes it possible for search engines to index audio and video in ways that were previously impossible. Pop Up Archive, an audio search company in Oakland, California, works with speech-to-text software specially trained for news media and spoken word content to make radio, podcasts, and archival audio searchable. A challenge for audio search is that even though a transcription with a few mistakes works just fine for search engines, people often expect transcriptions to be perfect.<sup>23</sup>

When the speech is in a language that is not understood, it needs to be translated as well. Perhaps you have watched a movie on an international flight and were able to choose from subtitles in many different languages. Creating subtitles for a foreign film is an asynchronous task that is substantially easier task than doing a real-time translation, and the demand for skilled translators for speeches and other synchronous situations (and interpreters, who translate speech to sign language for people with hearing disabilities) remains high.

## Access Policies

Different levels of interactions or access can apply to different

23. For a recent historical and highly technical review of speech recognition written by some of the most prominent researchers in the field, see ([Huang, Baker, and Reddy 2014](#)) An easier to read story about Apple's Siri voice recognition program is ([Geller 2012](#)). Popup archive is <https://www.popuparchive.org/> and its audio search service is <https://www.audiosear.ch/>

resources in a collection or to different categories of users. For example, library collections can range from completely open and public, to allowing limited access, to wholly private and restricted.

The library stacks might be open to anyone, but rare documents in a special collection are only accessible to authorized researchers. The same is true of museums, which typically have only a fraction of their collections on public display.

Because of their commercial and competitive purposes, organizing systems in business domains are more likely to enforce a granular level of access control that distinguishes people according to their roles and the nature of their interactions with resources. For example, administrative assistants in a company's Human Resources department are not allowed to see salaries; HR employees in a benefits administration role can see salaries but not change them; management-level employees in HR can change the salaries. Some firms limit access to specific times from authorized computers or IP addresses.<sup>24</sup>

24. These access controls to the organizing system or its host computer are enforced using passwords and more sophisticated software and hardware techniques. Some access control policies are mandated by regulations to ensure privacy of personal data, and policies differ from industry to industry and from country to country. Access controls can improve the credibility of information by identifying who created or changed it, especially important when traceability is required (e.g., financial accounting).

An important difference between interactions with

A noteworthy situation arises when the person accessing the organizing system is the one who designed and implemented it. In this case, the person will have qualitatively better knowledge of the resources and the supported interactions. This situation most often arises in the organizing systems in kitchens, home closets, and other highly personal domains but can also occur in knowledge-intensive business and professional domains like consulting, customer relationship management, and scientific research.

Many of the organizing systems used by individuals are embedded in physical contexts where access controls are applied in a coarse

physical resources and those with digital resources is how they use resource descriptions for access control. Resources sometimes have associated security classifications like “Top Secret” that restrict who can learn about their existence or obtain them. Nonetheless, if you get your hands on a top secret printed document, nothing can prevent you from reading it. Similarly, printed resources often have “All rights reserved” copyright notices that say that you cannot copy them, but nothing can prevent you from making copies with a copy machine. On the other hand, learning of the existence of a digital resource might be of little value if copyright or licensing restrictions prevent you from obtaining it. Moreover, obtaining a digital resource might be of no value if its content is only available using a password, decryption key, or other resource description that enforces access control directly rather than indirectly like the security classifications.

manner. We need a key to get into the house, but we do not need additional permissions or passwords to enter our closets or to take a book from a bookshelf. In our online lives, however, we readily accept and impose more granular access controls. For example, we might allow or block individual “friend” requests on Facebook or mark photos on Flickr as public, private, or viewable only by named groups or individuals.

We can further contrast access policies based on their origins or motivations.

Designed resource access policies are established by the designer or operator of an organizing system to satisfy internally generated requirements. Examples of designed access policies are:

- giving more access to “inside” users (e.g., residents of a community, students or faculty members at a university, or employees of a company) than to anonymous or “outside” users;
- giving more access to paying users than to users who do not pay;
- giving more access to users with capabilities or competencies that can add value to the organizing system (e.g., material culture researchers like archaeologists or anthropologists, who often work with resources in museum collections that are not on display).

*Imposed Policies* are mandated by an external entity and the organizing system must comply with them. For example, an organizing system might have to follow information privacy, security, or other regulations that restrict access to resources or the interactions that can be made with them.

University libraries typically complement or replace parts of their print collections with networked access to digital content licensed from publishers. Typical licensing terms then require them to

restrict access to users that are associated with the university, either by being on campus or by using virtual private network (VPN) software that controls remote access to the library network.<sup>25</sup>

Copyright law limits the uses of a substantial majority of the books in the collections of major libraries, prohibiting them from being made fully available in digital formats. Museums often prohibit photography because they do not own the rights to modern works they display.

Whether an access policy is designed or imposed is not always clear. Policies that were originally designed for a particular organizing system may over time become best practices or industry standards, which regulators or industry groups not satisfied with “self-regulation” later impose. Museums might aggressively enforce a ban on photography not just to comply with copyright law, but also to enhance the revenue they get from selling posters and reproductions.

25. In response to this trend, however, many libraries are supporting “open access” initiatives that strive to make scholarly publications available without restriction ([Bailey 2007](#)). Libraries and ebook vendors are engaged in a tussle about the extent to which the “first sale” rule that allows libraries to lend physical books without restrictions also applies to ebooks ([Howard 2011](#)).

## 23. Maintaining Resources

*Maintaining resources* is an important activity in every organizing system because resources must be available at the time they are needed. Beyond these basic shared motivations are substantial differences in [maintenance](#) goals and methods depending on the domain of the organizing system.

However, different domains sometimes use the same terms to describe different [maintenance](#) activities and different terms for similar activities. Common maintenance activities are [storage](#), [preservation](#), [curation](#), and [governance](#). Storage is most often used when referring to physical or technological aspects of maintaining resources; backup (for short-term storage), archiving (for long-term storage), and migration (moving stored resources from one storage device to another) are similar in this respect. The other three terms generally refer to activities or methods that more closely overlap in meaning; we will distinguish them in [“Preservation”](#) through [“Governance”](#).

Selection and maintenance are interdependent. Selection is based on an initial set of rules that determine which resources enter the organizing system. Maintenance includes the work to preserve the resources, the processes for evaluating and revising the original selection criteria, and the removal of resources from the system when they no longer need to be preserved. More stringent rules for selecting resources generally imply a maintenance plan that carefully enforces the same constraints that limit selection. This is just common sense whether the resource is a piece of art, an automobile, a software package, or a star basketball player; if you worked hard to find or paid a lot to acquire a resource, you are going to take care of it and will not soon be buying another one.

Ideally, [maintenance](#) requirements for resources should be



anticipated when organizing principles are defined and implemented. Resource descriptions to support preservation of digital resources are especially important.<sup>1</sup>

## Motivations for Maintaining Resources

The concept of *memory institution* broadly applies to a great many organizing systems that share the goal of preserving knowledge and cultural heritage.<sup>2</sup> The primary resources in libraries, museums, data archives or other [memory institutions](#) are fixed cultural, historic, or scientific artifacts that are maintained because they are unique and original items with future value. This is why the Musée du Louvre preserves the portrait of the [Mona Lisa](#) and the United States National Archives preserves the Declaration of Independence.<sup>3</sup>

1. [\(Guenther and Wolfe 2009\)](#).
2. This is the historical and dominant conception of the research library, but libraries are now fighting to prove that they are much more than just repositories because many of their users place greater value “on-the-fly access” of current materials. See [\(Teper 2005\)](#) for a sobering analysis of this dilemma.
3. Today the United States National Archives displays the Declaration of Independence, Bill of Rights, and the U.S. Constitution in sealed titanium cases filled with inert argon gas. Unfortunately, for over a century these documents were barely preserved at all; the Declaration

In contrast, in businesses organizing systems, many of the resources that are collected and managed have limited intrinsic value. The motivation for preservation and maintenance is economic; resources are maintained because they are essential in running the business. For example, businesses collect and preserve information about employees, inventory, orders, invoices, etc., because it ensures internal goals of efficiency, revenue generation, and competitive advantage. The same resources (e.g., customer information) are often used by more than one part of the business.<sup>4</sup> Maintaining the accuracy and consistency of changing resources is a major challenge in business organizing systems.<sup>5</sup>

Many business organizing systems preserve information needed to satisfy externally imposed regulatory or compliance policies and serve largely to avoid possible catastrophic costs from penalties and lawsuits. In all these cases, resources are maintained as one of the

hung on the wall at the United States Patent Office in direct sunlight for about 40 years.

4. Customer information drives day-to-day operations, but is also used in decision support and strategic planning.
5. For businesses “in the world,” a “customer” is usually an actual person whose identity was learned in a transaction, but for many web-based businesses and search engines a customer is a computational model extracted from browser access and click logs that is a kind of “theoretical customer” whose actual identity is often unknown. These computational customers are the targets of the computational advertising in search engines.

means employed to preserve the business as an ongoing enterprise, not as an end in itself.

Unlike libraries, archives, and museums, indefinite preservation is not the central goal of most business organizing systems. These organizing systems mostly manage information needed to carry out day-to-day operations or relatively recent historical information used in decision support and strategic planning. In addition to these internal mandates, businesses have to conform to securities, taxation, and compliance regulations that impose requirements for long-term information preservation.<sup>6</sup>

Of course, libraries, museums, and archives also confront economic issues as they seek to preserve and maintain their collections and

6. The Sarbanes-Oxley Act in the United States and similar legislation in other countries require firms to preserve transactional and accounting records and any document that relates to “internal controls,” which arguably includes any information in any format created by any employee ([Langevoort 2006](#)). Civil procedure rules that permit discovery of evidence in lawsuits have long required firms to retain documents, and the proliferation of digital document types like email, voice mail, shared calendars and instant messages imposes new storage requirements and challenges ([Levy and Casey 2006](#)). However, if a company has a data retention policy that includes the systematic deletion of documents when they are no longer needed, courts have noted that this is not willful destruction of evidence.

themselves as [memory institutions](#).<sup>7</sup> They view their collections as intrinsically valuable in ways that firms generally do not. Because of this, extensive energy goes into preservation, protection, and storage of resources in memory institutions, and it is more rare that resources may be discarded or de-accessioned. Art galleries are an interesting hybrid because they organize and preserve collections that are valuable, but if they do not manage to sell some things, they will not stay in business.

In between these contrasting purposes of preservation and maintenance are the motives in personal collections, which occasionally are created because of the inherent value of the items but more typically because of their value in supporting personal activities. Some people treasure old photos or collectibles that belonged to their parents or grandparents and imagine their own children or grandchildren enjoying them, but many old collections seem to end up as offerings on eBay. In addition, many personal organizing systems are task-oriented, so their contents need not be preserved after the task is completed.<sup>8</sup>

7. Libraries are increasingly faced with the choice of providing access to digital resources through renewable licensing agreements, “pay per view” arrangements, or not at all. To some librarians, however, the failure to obtain permanent access rights “offends the traditional ideal of libraries” as [memory institution](#) (Carr 2010).

8. For example, students writing a term paper usually organize the printed and digital resources they rely on; the former are probably kept in folders or in piles on the desk, and the latter in a computer file system. This organizing system is not likely to be preserved after the

## Preservation

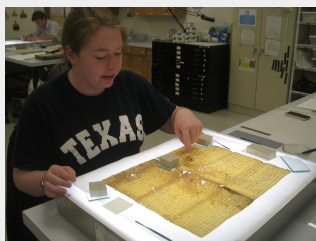
At the most basic level, *preservation* of resources means maintaining them in conditions that protect them from physical damage or deterioration. Libraries, museums, and archives aim for stable temperatures and low humidity. Permanently or temporarily out-of-service aircraft are parked in deserts where dry conditions reduce corrosion. Risk-aware businesses create continuity plans that involve off-site storage of the data and documents needed to stay in business in the event of a natural disaster or other disruption.

When the goal is indefinite preservation, other [maintenance](#) issues arise if resources deteriorate or are damaged. How much of an artifact's worth is locked in with the medium used to express it? How much restoration should be attempted? How much of an artifact's essence is retained when digitized?

Archivists at Work

term paper is finished. An exception that proves the rule is the task of paying income taxes for which (in the USA) taxpayers are legally required to keep evidence for up to seven years after filing a tax return (<http://www.irs.gov/Businesses/Small-Businesses-&-Self-Employed/How-long-should-I-keep-records%3F>).

The University of Texas School of Information has great expertise in document archiving and preservation and operates a conservation laboratory.



Catherine Bell works on a light table to see the tears and losses in a 19th-century document more clearly.



Heather Bollinger has repaired a 19th-century document with conservation quality tissue and wheat starch paste.

(Photos by R. Glushko.)

## *Digitization and Preserving Resources*

*Preservation* is often a key motive for *digitization*, but *digitization*

alone is not preservation. Digitization creates preservation challenges because technological obsolescence of computer software and hardware require ongoing efforts to ensure the digitized resources can be accessed.

Technological obsolescence is the major challenge in maintaining digital resources. The most visible one is a result of the relentless evolution of the physical media and environments used to store digital information in both institutional or business and personal organizing systems. Computer data began to be stored on magnetic tape and hard disk drives six decades ago, on floppy disks four decades ago, on CDs three decades ago, on DVDs two decades ago, on solid-state drives half a decade ago, and in “cloud-based” or “virtual” storage environments in the last decade. As the capacity of storage technologies grows, economic and efficiency considerations often make the case to adopt new technology to store newly acquired digital resources and raise questions about what to do with the existing ones.<sup>9</sup>

The second challenge might seem paradoxical. Even though digital storage capacity increases at a staggering pace, the expected useful lifetimes of the physical storage media are measured in years or at best in decades. Colloquial terms for this problem are [\*data rot\*](#) or “bit rot.” In contrast, books printed on acid-free paper can last for centuries. The contrast is striking; books on library shelves do not disappear if no one uses them, but digital data can be lost if no one wants access to it within a year or two after its creation.<sup>10</sup>

However, limits to the physical lifetime of digital storage media are much less significant than the third challenge, the fact that

9. [\(Rothenberg 1999\)](#).

10. [\(Pogue 2009\)](#).

the software and its associated computing environment used to parse and interpret the resource at the time of preservation might no longer be available when the resource needs to be accessed. Twenty-five years ago most digital documents were created using the Word Perfect word processor, but today the vast majority is created using Microsoft Word and few people use Word Perfect today. Software and services that convert documents from old formats to new ones are widely available, but they are only useful if the old file can be read from its legacy storage medium.<sup>11</sup>

11. Many of those Word Perfect documents were stored on floppy disks because floppy disk drives were built into almost every personal computer for decades, but it would be hard to find such disk drives today. And even if someone with a collection of word processor documents stored on floppy disks in 1995 had copied those files to newer storage technologies, it is unlikely that the current version of the word processor would be able to read them. Software application vendors usually preserve “backwards compatibility” for a few years with earlier versions to give users time to update their software, but few would support older versions indefinitely because to do so can make it difficult to implement new features.

Digital resources can be encoded using non-proprietary and standardized data formats to ensure “forward compatibility” in any software application that implements the version of the standard. However, if the ebook reader, web browser, or other software used to access the resource has capabilities that rely on later



Because almost every digital device has storage associated with it, problems posed by multiple storage environments can arise at all scales of organizing systems. Only a few years ago people often struggled with migrating files from their old computer, music player or phone when they got new ones. Web-based email and applications and web-based storage services like Dropbox, Amazon Cloud Drive, and Apple iCloud eliminate some data storage and migration problems by making them someone else's responsibility, but in doing so introduce privacy and reliability concerns.

It is easy to say that the solutions to the problems of digital preservation are regular recopying of the digital resources onto new storage media and then migrating them to new formats when significantly better ones come along. In practice, however, how libraries, businesses, government agencies or other enterprises deal with these problems depends on their budgets and on their technical sophistication. In addition, not every resource should or can always be migrated, and the co-existence of multiple storage technologies makes an organizing system more complex because different storage formats and devices can be collectively incompatible.

The Hathi Trust Digital Library

The [Hathi Trust](#) is a worldwide partnership of several dozen major research institutions and libraries dedicated to “collecting, organizing, preserving, communicating, and sharing the record of human

versions of the standards the “old data” will not have taken advantage of them.

knowledge.” The Hathi Trust was established in 2008 to coordinate the efforts of libraries in managing the digital copies of the books they received in return for providing books to Google for its book digitization projects. Since then the Hathi Trust has broadened its scope to include the public domain books collected by the Internet Archive and numerous other digital collections, and today its digital library has over ten million volumes. The costs of running the Hathi Trust and its digital library are shared in a transparent manner by the institutions that contributed digital collections or that want access to them, which reduces the costs for everyone compared to a “go it alone” strategy. The Hathi Trust Digital Library has separate modes for catalog search and full-text search of the library contents, unlike commercial search engines that do not distinguish them. A second important difference between the Hathi Trust Digital Library and commercial search engines is the absence of display advertising and “sponsored search” results.

(Interoperability and integration are discussed in [Interactions with Resources](#).)

### *Preserving the Web*

Preservation of web resources is inherently problematic. Unlike libraries, museums, archives, and many other kinds of organizing systems that contain collections of unchanging resources, organizing systems on the web often contain resources that are

highly dynamic. Some websites change by adding content, and others change by editing or removing it.<sup>12</sup>

Longitudinal studies have shown that hundreds of millions of web pages change at least once a week, even though most web pages never change or change infrequently.<sup>13</sup> Nevertheless, the continued

12. This is tautologically true for sites that publish news, weather, product catalogs with inventory information, stock prices, and similar continually updated content because many of their pages are automatically revised when events happen or as information arrives from other sources. It is also true for blogs, wikis, Facebook, Flickr, YouTube, Yelp and the great many other “Web 2.0” sites whose content changes as they incorporate a steady stream of user-generated content.

In some cases changes to web pages are attempts to rewrite history and prevent preservation by removing all traces of information that later turned out to be embarrassing, contradictory, or politically incorrect. When pages cannot be changed, like the archives of newspapers published on the web, only the search engine can remove them from search results, and in 2014 the European Court ruled that people could ask Google to do that.

13. [\(Fetterly et al. 2003\)](#).

Most people understand that web pages can change, but

existence of a particular web page is hardly sufficient to preserve it if it is not popular and relevant enough to show up in the first few pages of search results. Persistent access requires preservation, but preservation is not meaningful if there is no realistic probability of future access.

Comprehensive web search engines like Google and Bing use crawlers to continually update their indexed collections of web pages and their search results link to the current version, so preservation of older versions is explicitly not a goal. Furthermore, search engines do not reveal any details about how frequently they update their collections of indexed pages.<sup>14</sup>

The Internet Archive and the  
“Wayback Machine”

The Internet Archive (Archive.org), founded by Brewster Kahle, makes preservation of the web its first and foremost activity, and when you enter a URI into its “Wayback Machine” you can see what a site looked like at different moments in time. For example,

most changed web pages do not highlight the changes. A “diff” tool from Microsoft reveals them.  
<http://research.microsoft.com/en-us/projects/DiffIE/default.aspx>

14. However, when a website disappears its first page can often be found in the search engine’s index “cache” rather than by following what would be a broken link.

www.berkeley.edu was archived about 2500 times between October 1996 and January 2013, including about twice a week on average during all of 2012. Even so, since a large site like berkeley.edu often changes many times a day, the Wayback Machine's preservation of berkeley.edu is incomplete, and it only preserves a fraction of the web's sites. Since 2006 the Internet Archive has hosted the "Archive-It" service to enable hundreds of schools, libraries, historical societies, and other institutions to archive collections of digital resources.<sup>15</sup>

15. Brewster Kahle has been described as a computer engineer, Internet entrepreneur, Internet activist, advocate of universal access to knowledge, and digital librarian ([http://en.wikipedia.org/wiki/Brewster\\_Kahle](http://en.wikipedia.org/wiki/Brewster_Kahle)). In addition to websites, the Internet Archive preserves several million books, over a million pieces of video, 400,000 news programs from broadcast TV, over a million audio recordings, and over 100,000 live music concerts.

The Memento project has proposed a specification for using HTTP headers to perform "datetime negotiation" with the Wayback Machine and other archives of web pages, making it unnecessary for Memento to save anything on its own. Memento is implemented as a browser plug-in to "browse backwards in time"

## *Preserving Resource Instances*

A focus on preserving particular resource instances is most clear in museums and archives, where collections typically consist of unique and original items. There are many copies and derivative works of the [Mona Lisa](#), but if the original [Mona Lisa](#) were destroyed none of them would be acceptable as a replacement.<sup>16</sup>

Archivists and historians argue that it is essential to preserve original documents because they convey more information than just their textual content. Paul Duguid recounts how a medical historian used faint smells of vinegar in 18th-century letters to investigate a cholera epidemic because disinfecting letters with vinegar was thought to prevent the spread of the disease. Obviously, the vinegar smell would not have been part of a digitized letter.<sup>17</sup>

Zoos often give a distinctive or attractive animal a name and then market it as a special or unique instance. For example, the Berlin Zoo successfully marketed a polar bear named Knut to become a world famous celebrity, and the zoo made millions of dollars a year through increased visits and sales of branded merchandise. Merchandise sales have continued even though Knut died

whenever older versions of pages are available from archives that use its specification. ([VandeSompel 2010](#)).

16. People might still enjoy the many [Mona Lisa](#) parodies and recreations. See <http://www.megamonalisa.com>, [http://www.oddee.com/item\\_96790.aspx](http://www.oddee.com/item_96790.aspx), [http://www.chilloutpoint.com/art\\_and\\_design/the-best-mona-lisa-parodies.html](http://www.chilloutpoint.com/art_and_design/the-best-mona-lisa-parodies.html).
17. ([Brown and Duguid 2002](#)).

unexpectedly in March 2011, which suggests that the zoo was less interested in preserving that particular polar bear than in preserving the revenue stream based on that resource.<sup>18</sup>

Most business organizing systems, especially those that “run the business” by supporting day-to-day operations, are designed to preserve instances. These include systems for order management, customer relationship management, inventory management, digital asset management, record management, email archiving, and more general-purpose document management. In all of these domains, it is often necessary to retrieve specific information resources to serve customers or to meet compliance or traceability goals.

Recent developments in sensor technology enable very extensive data collection about the state and performance of machines, engines, equipment, and other types of physical resources, including human ones. (Are you wearing an activity tracker right now?) When combined with historical information about maintenance activity, predictive analytics techniques can use this data to determine normal operating ranges and indicators of coming performance degradation or failures. Predictive maintenance can maximize resource lifetimes while minimizing maintenance and inventory costs. These techniques have recently been used to predict when professional basketball players are at risk of an injury, potentially enabling NBA teams to identify the best time to rest their star players without impairing their competitive strategy.<sup>19</sup>

18. ([Savodnik 2011](#)).

19. ([Talukder 2016](#))

## *Preserving Resource Types*

### “Shamu” the Killer Whale



This photo of “Shamu” was taken at one of the three Sea World marine parks in the US, but it does not matter which one because each of them has a killer whale (orca) performing there called Shamu. Similarly, it does not matter when this

photo was taken because if a particular orca dies, it is replaced by another that also performs using Shamu as a stage name.

([Photo by Mike Saechang](#). Creative Commons CC BY-SA 2.0 license.)

Some business organizing systems are designed to preserve types or classes of resources rather than resource instances. In particular, systems for content management typically organize a repository of reusable or “source” information resources from which specific “product” resources are then generated. For example, content management systems might contain modular information about a company’s products that are assembled and delivered in sales or



product catalogs, installation guides, operating guides, or repair manuals.<sup>20</sup>

Businesses strive to preserve the collective knowledge embodied in the company's people, systems, management techniques, past decisions, customer relationships, and intellectual property. Much of this knowledge is “know how”—knowing how to get things done or knowing how things work—that is tacit or informal. *Knowledge management systems*(KMS) are a type of business organizing system whose goal is to capture and systematize these information resources.<sup>21</sup> As with content management, the focus of knowledge management is the reuse of “knowledge as type,” putting the focus on the knowledge rather than the specifics of how it found its way into the organizing system.

Libraries have a similar emphasis on preserving resource types rather than instances. The bulk of most library collections, especially public libraries, is made up of books that have many equivalent copies in other collections. When a library has a copy of *Moby Dick* it is preserving the abstract [work](#) rather than the

20. The set of content modules and their assembly structure for each kind of generated document conforms to a template or pattern that is called the document type model when it is expressed in XML.
21. Company intranets, wikis, and blogs are often used as knowledge management technologies; Lotus Notes and Microsoft SharePoint are popular commercial systems. (See the case study in [“Knowledge Management for a Small Consulting Firm”](#).)

particular physical [instance](#)—unless the copy of *Moby Dick* is a rare first edition signed by Melville.

Even when zoos give their popular animals individual names, it seems logical that the zoo's goal is to preserve animal species rather than instances because any particular animal has a finite lifespan and cannot be preserved forever.<sup>22</sup>

### *Preserving Resource Collections*

In some organizing systems any specific resource might be of little interest or importance in its own right but is valuable because of its membership in a collection of essentially identical items. This is the situation in the data warehouses used by businesses to identify trends in customer or transaction data or in the huge data collections created by scientists. These collections are typically analyzed as complete sets. A scientist does not borrow a single data point when she accesses a data collection; she borrows the complete dataset consisting of millions or billions of data points. This requirement raises difficult questions about what additional

22. In addition, the line between “preserving species” and “preserving marketing brands” is a fine one for zoos with celebrity animals, and in animal theme parks like Sea World, it seems to have been crossed. “Shamu” was the first killer whale (orca) to survive long in captivity and performed for several years at SeaWorld San Diego. Shamu died in 1971 but over forty years later all three US-based SeaWorld parks have Shamu shows and Shamu webcams.

software or equipment need to be preserved in an organizing system along with the data to ensure that it can be reanalyzed.<sup>23</sup>

Sometimes, specific items in a collection might have some value or interest on their own, but they acquire even greater significance and enhanced meaning because of the context created by other items in the collection that are related in some essential way. The odd collection of “things people swallow that they should not” at the Mütter Museum is a perfect example.<sup>24</sup>

23. ([Manyika et al. 2011](#)).

24. The College of Physicians of Philadelphia’s Mütter Museum houses a novel collection of artifacts meant to “educate future doctors about anatomy and human medical anomalies.” No museum in the world is like it; it contains display cases full of human skulls, abnormal fetuses in jars, preserved human bodies, a garden of medicinal herbs, and many other unique collections of resources.

However, one sub-collection best reflects the distinctive and idiosyncratic [selection](#) and arrangement of resources in the museum. Chevalier Jackson, a distinguished laryngologist, collected over 2,000 objects extracted from the throats of patients. Because of the peculiar focus and educational focus of this collection, and because there are few shared characteristics of “things people swallow that they should not,” the characteristics and principles used to organize and describe the collection would be of little use in another organizing

## Curation

For almost a century *curation* has referred to the processes by which a resource in a collection is maintained over time, which may include actions to improve access or to restore or transform its representation or presentation.<sup>25</sup>

Furthermore, especially in cultural heritage collections, curation also includes research to identify, describe, and authenticate resources in a collection. Resource descriptions are often updated to reflect new knowledge or interpretations about the primary resources.<sup>26</sup>

system. What other collection would include toys, bones, sewing needles, coins, shells, and dental material? It is hard to imagine that any other collection that would include all of these items plus fully annotated record of sex and approximate age of patient, the amount of time the extraction procedure took, the tool used, and whether or not the patient survived.

25. Curation is a very old concept whose Medieval meaning focused on the “preservation and cure of souls” by a pastor, priest, or “curate” ([Simpson and Weiner 1989](#)). A set of related and systematized curation practices for some class of resources is often called a curation system, especially when they are embodied in technology.

26. Information about which resources are most often

Curation takes place in all organizing systems—at a personal scale when we rearrange a bookshelf to accommodate new books or create new file folders for this year’s health insurance claims, at an institutional scale when a museum designs a new exhibit or a zoo creates a new habitat, and at web scale when people select photos to upload to Flickr or Facebook and then tag or “Like” those uploaded by others.

An individual, company, or any other creator of a website can make decisions and employ technology that maintains the contents, quality and character of the site over time. In that respect website curation and governance practices are little different than those for the organizing systems in memory institutions or business enterprises. The key to curation is having clear policies for collecting resources and maintaining them over time that enable people and automated processes to ensure that resource descriptions or data are authoritative, accurate, complete, consistent, and non-redundant.

### *Institutional Curation*

Curation is most necessary and explicit in institutional organizing systems where the large number of resources or their heterogeneity requires choices to be made about which ones should be most accessible, how they should be organized to ensure this access, and which ones need most to be preserved to ensure continued accessibility over time. Curation might be thought of as an ongoing or deferred selection activity because curation decisions must often be made on an item-by-item basis.

interacted with in scientific or archival collections is essential in understanding resource value and quality.

Curation in these institutional contexts requires extensive professional training. The institutional authority empowers individuals or groups to make curation decisions. No one questions whether a museum curator or a compliance manager should be doing what they do.<sup>27</sup>

Institutional curation may be supported by automated methods. An “approval plan” is often implemented for the acquisition of new books by libraries that involves an initial selection of certain criteria (such as “published by an American university press; costs less than \$100; not a reissue of an earlier edition; classed within a particular Library of Congress range”) that enable libraries to automatically purchase all books meeting the criteria. While the approval plan can certainly be considered a selection activity, we cite it in maintenance as an example of a strategy to maintain the currency and relevancy of a given collection.

### *Individual Curation*

The Life-Changing Magic of Tidying Up

27. In [memory institutions](#), the most common job titles include “curator” or “conservator.” In for-profit contexts where “[governance](#)” is more common than “curation” job titles reflect that difference. In addition to “governance,” job titles often include “recordkeeping,” “compliance,” or “regulatory” prefixes to “officer,” “accountant,” or “analyst” job classifications.

Many of the ever growing number of self-help books about organizing seem to approach it as an intellectual contest to devise more elaborate and optimized storage strategies. Marie Kondo's wildly popular 2014 book [\*The Life-changing Magic of Tidying Up\*](#), an international best-seller, has upended the conversation with an unapologetic dogma of removal that promises to yield a happier—and much more minimalist—life for individuals with their at-home organizing systems..

Kondo's method mandates that only what brings one joy may be kept. Everything else must be tossed — unused gifts, books kept only for reference but never referenced, unworn clothing, and anything else that does not bring its owner joy. Kondo's approach is designed for personal organizing systems, and would be difficult to implement in systems in systems used by multiple individuals, much less institutions. However, Kondo's rejection of the concept that things should be saved for a rainy day might benefit organizations by making them more attentive to the costs of maintaining resources with no current use.

While people must make up their own minds about how they manage their possessions, there is compelling evidence from cognitive science and behavioral economics that decision-making throughout the day can be mentally exhausting. Kondo's approach implicitly recognizes this limitation by requiring cognitive energy up front to reduce the total number of resources to the bare minimum necessary (by one's own "joy standards"). This philosophy has people spend decision-making

energy where it counts the most and makes it easier to make maintenance decisions over time.

Curation by individuals has been studied a great deal in the research discipline of Personal Information Management (PIM).<sup>28</sup> Much of this work has been influenced for decades by a seminal article written by Vannevar Bush titled [As We May Think](#). Bush envisioned the Memex, “a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility.” Bush’s most influential idea was his proposal for organizing sets of related resources as “trails” connected by associative links, the ancestor of the [hypertext links](#) that define today’s web.<sup>29</sup>

28. Because personal collections are strongly biased by the experiences and goals of the organizer, they are highly idiosyncratic, but still often embody well-thought-out and carefully executed curation activities ([Kirsh 2000](#)), ([Jones 2007](#)), ([Marshall 2007](#)), ([Marshall 2008](#)).
29. ([Bush 1945](#)). Bush imagined that Memex users could share these packages of trails and that a profession of trailbuilders would emerge. However, he did not envision that the Memexes themselves could be interconnected, nor did he imagine that their contents could be searched computationally.



## *Social and Web Curation*

Many individuals spend a great amount of time curating their own websites, but when a site can attract large numbers of users, it often allows users to annotate, “tag,” “like,” “+1,” and otherwise evaluate its resources. The concept of curation has recently been adapted to refer to these volunteer efforts of individuals to create, maintain, and evaluate web resources.<sup>30</sup> The massive scale of these bottom-up and distributed activities is curation by “crowdsourcing,” the continuously aggregated actions and contributions of users.<sup>31</sup>

The informal and organic “folksonomies” that result from their aggregated effort create organization and authority through network effects.<sup>32</sup> This undermines traditional centralized mechanisms of organization and governance and threatens any business model in publishing, education, and entertainment that

30. ([Howe 2008](#)).

31. The most salient example of this so called “community curation” activity is the work to maintain the Wikipedia open-source encyclopedia according to a curation system of roles and functions that governs how and under what conditions contributors can add, revise, or delete articles; receive notifications of changes to articles; and resolve editing disputes ([Lovink and Tkacz 2011](#)). Some museums and scientific data repositories also encourage voluntary curation to analyze and classify specimens or photographs ([Wright 2010](#)).

32. ([Trant 2009](#)).

has relied on top-down control and professional curation.<sup>33</sup> Professional curators are not pleased to have the *ad hoc* work of untrained people working on websites described as curation.

Most websites are not curated in a systematic way, and the decentralized nature of the web and its easy extensibility means that the web as a whole defies curation. It is easy to find many copies of the same document, image, music file, or video and not easy to determine which is the original, authoritative or authorized version. Broken links return “Error 404 Not Found” messages.<sup>34</sup>

Problems that result from lazy or careless webmastering are minor compared to those that result from deliberate misclassification, falsification, or malice. An entirely new vocabulary has emerged to describe these web resources with bad intent: “spam,” “phishing,” “malware,” “fakeware,” “spyware,” “keyword stuffing,” “spamdexing,” “META tag abuse,” “link farms,” “cybersquatters,” “phantom sites,” and many more.<sup>35</sup> Internet service providers, security software

33. Some popular “community content” sites like Yelp where people rate local businesses have been criticized for allowing positive rating manipulation. Yelp has also been criticized for allowing negative manipulation of ratings when competitors slam their rivals.
34. The resource might have been put someplace else when the site was reorganized or a new web server was installed. It is no longer the same resource because it will have another URI, even if its content did not change.
35. All of these terms refer to types of web resources or techniques whose purpose is to mislead people into

firms, email services, and search engines are engaged in a constant war against these kinds of malicious resources and techniques.<sup>36</sup>

Since we cannot prevent these deceptions by controlling what web resources are created in the first place, we have to defend ourselves from them after the fact. “Defensive curation” techniques include filters and firewalls that block access to particular sites or resource types, but whether this is curation or censorship is often debated,

doing things or letting things be done to their computers that will cost them their money, time, privacy, reputation, or worse. We know too well what spam is. Phishing is a type of spam that directs recipients to a fake website designed to look like a legitimate one to trick them into entering account numbers, passwords, or other sensitive personal information. Malware, fakeware, or spyware sites offer tempting downloadable content that installs software designed to steal information from or take control of the visiting computer. Keyword stuffing, spamdexing, and META tag abuse are techniques that try to mislead search engines about the content of a resource by annotating it with false descriptions. Link farms or scraper sites contain little useful or original content and exist solely for the purpose of manipulating search engine rankings to increase advertising revenue. Similarly, cybersquatters register domain names with the hope of profiting from the goodwill of a trademark they do not own.

36. [\(Brown 2009\)](#).

and from the perspective of the government or organization doing the censorship it is certainly curation. Nevertheless, the decentralized nature of the web and its open protocols can sometimes enable these controls to be bypassed.

### *Computational Curation*

Search engines continuously curate the web because the algorithms they use for determining relevance and ranking determine what resources people are likely to access. At a smaller scale, there are many kinds of tools for managing the quality of a website, such as ensuring that HTML content is valid, that links work, and that the site is being crawled completely. Another familiar example is the spam and content filtering that takes place in our email systems that automatically classifies incoming messages and sorts them into appropriate folders.

One might think that computational curation is always more reliable than any curation carried out by people. Certainly, it seems that we should always be able to trust any assertion created by context-aware resources like temperature or location sensors. But can we trust the accuracy of web content? Search engines use the popularity of web pages and the structure of links between them to compute relevance. But popularity and relevance do not always ensure accuracy. We can easily find popular pages that prove the existence of UFOs or claim to validate wacky conspiracy theories.

Computational curation is more predictable than curation done by people, but search engines have long been accused of bias built into their algorithms. For example, Google's search engine has been criticized for giving too much credibility to websites with .edu domain names, to sites that have been around for a long time, or

that are owned by or that partner with the company, like Google Maps or YouTube.<sup>37</sup>

In organizing systems that contain data, there are numerous tools for *name matching*, the task of determining when two different text strings denote the same person, object, or other named entity. This problem of eliminating duplicates and establishing a controlled or authoritative version of the data item arises in numerous application areas but familiar ones include law-enforcement and counter-terrorism. Done incorrectly, it might mean that you end up on a “watch list” and experience difficulties every time you want to fly commercially.

An extremely promising new approach to computational curation involves using scientific measuring equipment to analyze damaged physical resources and then building software models of the resources that can be manipulated to restore the resources or otherwise improve access to their content. For example, the first sound recordings were made using rotating wax cylinders; sounds caused a diaphragm to vibrate, the pattern of vibration was transferred into a connected stylus, which then cut a groove into the wax. When the cylinder was rotated past a passive stylus, it would vibrate according to the groove pattern, and the amplified vibrations could be heard as the replayed sound. Unfortunately, wax cylinders from the 19th century are now so fragile that they would fall apart if they were played. This dilemma was resolved by Carl Haber, an experimental physicist at the Lawrence Berkeley Laboratory. Haber used image processing techniques to convert microscope-detailed scans of the grooves in the wax cylinders.

37. ([Diaz 2005](#)), ([Grimmelmann 2009](#)).

Measurements of the grooves could then be transformed to reproduce the sounds captured in the grooves.<sup>38</sup>

A second example of computational curation applied to digital preservation is work done by a research team led by Melissa Terras and Tim Weyrich at University College London to build a 3-dimensional model of a 17th-century “Great Parchment Book” damaged in an 18th-century fire. The parchment was singed, shriveled, creased, folded, and nearly impossible to read (see [website](#)). After traditional document restoration techniques (e.g., illustrated in photos in [“Preservation”](#)) went as far as they could, the researchers used digital image capture and modeling techniques to create a software model of the parchment that could stretch and flatten the digital document to discover text hidden by the damage.

### *Discarding, Removing, and Not Keeping*

So far, we have discussed maintenance as activities involved in preserving and protecting resources in an organizing system over time. An essential part of maintenance is the phasing out of resources that are damaged or unusable, expired or past their effectivity dates, or no longer relevant to any interaction.

Many organizations admit to a distinct lack of strategy in the removal aspect of maintenance. A firm with outdated storage technology might have to discard older data simply to make room

38. See [video of Haber](#) explaining how this works, Haber has recently been able to build a version of his scanning and image processing technology for use outside the laboratory that he calls Irene (Image, Reconstruct, Erase Noise, Etc.). ([Cowen 2015](#)) and ([Wilkinson 2014](#))

for new data, and might do so without considering that keeping some summary statistics would be valuable for historical analysis. Other firms might be biased towards keeping information just because they went to the trouble of collecting or acquiring it. Some amount of “intelligent” removal is an essential ingredient in any maintenance regime, and a popular book argues forcefully for continually discarding resources from personal organizing systems as a method of focusing on the resources that really matter. (See the sidebar, [The Life-Changing Magic of Tidying Up](#).)

In memory institutions, common terms for getting rid of resources include discarding, de-accession, de-selection, and weeding.

Efforts by libraries to automate the discarding of books that have not circulated for several years might seem like the obvious counterpart to their automated acquisition, but such efforts often produce passionate complaints from library patrons.<sup>39</sup>

Other domains have other mechanisms and terms for removing

39. For an explanation of automated acquisition see Eva Guggemos, Professional archivist and academic librarian <https://www.quora.com/How-do-libraries-decide-which-books-to-purchase-and-which-books-to-remove-from-circulation>.

For a cogent discussion of when and for what reasons weeding must take place in university libraries, see <https://mrlibrarydude.wordpress.com/2014/03/12/why-we-weed-book-deselection-in-academic-libraries/>.

A typical reaction when libraries discard books is described in [\(Jackman 2015\)](#)

resources. Employees are removed by firing, layoff, or retirement. Athletes are cut or waived or sent down from a sports team if their performance deteriorates.

Keeping an organizing system current often involves some amount of elimination of older resources in order to make space for the new: in fashion retail, the floor is constantly restocked with the latest styles. Software development teams will halt active support and documentation efforts of legacy versions.

Information resources are often discarded to comply with laws about retaining sensitive data. Governments and office holders sometimes destroy documents that might prove damaging or embarrassing if they are discovered through Freedom of Information requests or by opposing political parties.

More positively, the “right to be forgotten” movement and intentional destruction of information records about prior bankruptcy, credit problems, or juvenile arrests after a certain period of time has passed can be seen as a policy of “social forgetfulness” that gives people a chance to get on with their lives.<sup>40</sup>

It is worth noting that the ability to discard without having to reuse is relatively recent. Historically, the urge and need to discard has clashed with the availability of resources. In the Middle Ages, liturgical texts or music would be phased out, perhaps when the music had gone out of style or when entire sections of the liturgy were phased out by decree. When this happened, they would reuse the parchment or vellum, either by scraping it down or by flipping it over, pasting it in a book, and using the other side. The former of these solutions often created a *palimpsest*, a document or other

40. ([Blanchette and Johnson 2002](#))



resource in which the remnants of older content remain visible under the new.

Some people have difficulty in discarding things, regardless of their actual value. This behavior is called hoarding, and is now regarded as a kind of obsessive-compulsive disorder that requires treatment because it can cause emotional, physical, social, and even legal problems for the hoarder and family members. It seems unsympathetic that many TV shows and stories have been produced about especially compulsive hoarding. A famous example is that of the Collyer brothers in New York, who shut themselves off from the world for years, and when they were found dead inside their home in 1947 it contained 140 tons of collected items, including 25,000 books, fourteen pianos, thousands of bottles and tin cans, hundreds of yards of fabrics, and even a Model T car chassis.<sup>41</sup>

## Governance

[Governance](#) overlaps with [curation](#) in meaning, but typically has more of policy focus (what should be done), rather than a process focus (how to do it). Governance is also more frequently used to describe curation in business and scientific organizing systems rather than in libraries, archives, and museums. Governance has a broader scope than curation because it extends beyond the resources in a collection and also applies to the software, computing, and networking environments needed to use them. This broader scope also means that governance must specify the rights and responsibilities for the people who might interact with the

41. ([Neziroglu 2014](#)) and ([Lidz 2003](#)).

resources, the circumstances under which that might take place, and the methods they would be allowed to use.

*Corporate governance* is a common term applied to the ongoing maintenance and management of the relationship between operating practices and long-term strategic goals.<sup>42</sup>

*Data governance* policies are often shaped by laws, regulations or policies that prohibit the collection of certain kinds of objects or types of information. Privacy laws prohibit the collection or misuse of personally identifiable information about healthcare, education, telecommunications, video rental, and in some countries restrict the information collected during web browsing.<sup>43</sup>

### *Governance in Business Organizing Systems*

Governance is essential to deal with the frequent changes in

42. Libraries and museums must also deal with long-term strategy, but the lesser visibility of *library governance* and *museum governance* might simply reflect the greater concerns about fraud and malfeasance in for-profit business contexts than in non-profit contexts and the greater number of standards or “best practices” for corporate governance. ([Kim, Nofsinger, and Mohr 2009](#)).
43. Data governance decisions are also often shaped by the need to conform to information or process model standards, or to standards for IT service management like the *Information Technology Infrastructure Library*(ITIL). See <http://www.itil-officialsite.com/>.

business organizing systems and the associated activities of data quality management, access control to ensure security and privacy, compliance, deletion, and archiving. For many of these activities, effective governance involves the design and implementation of standard services to ensure that the activities are performed in an effective and consistent manner.<sup>44</sup>

#### Stop and Think: Business Data Governance

Ebay, Target, and other large companies have had tens of millions of passwords, credit card numbers, and other sensitive personal information breached by hackers or security lapses. Consider a data breach you have heard of or experienced. What secure information was leaked? How might the business's governance policies and practices have affected the severity of the breach? What changes could the businesses make to protect people's data better?

Today's information-intensive businesses capture and create large amounts of digital data. The concept of "business intelligence"

44. In this context, these management and maintenance activities are often described as "IT governance" ([Weill and Ross 2004](#)). Data classification is an essential IT governance activity because the confidentiality, competitive value, or currency of information are factors that determine who has access to it, how long it should be preserved, and where it should be stored at different points in its lifecycle.

emphasizes the value of data in identifying strategic directions and the tactics to implement them in marketing, customer relationship management, supply chain management and other information-intensive parts of the business.<sup>45</sup> A management aspect of governance in this domain is determining which resources and information will potentially provide economic or competitive advantages and determining which will not. A conceptual and technological aspect of governance is determining how best to organize the useful resources and information in business operations and information systems to secure the potential advantages.

Business intelligence is only as good as the data it is based on, which makes business data governance a critical concern that has rapidly developed its own specialized techniques and vocabulary. The most fundamental governance activity in information-driven businesses is identifying the “master data” about customers, employees, materials, products, suppliers, etc., that is reused by different business functions and is thus central to business operations.<sup>46</sup>

45. [\(Turban et al. 2010\)](#).

46. This master data must be continually “cleansed” to remove errors or inconsistencies, and “de-duplication” techniques are applied to ensure an authoritative source of data and to prevent the redundant storage of many copies of the same resource. Redundant storage can result in wasted time searching for the most recent or authoritative version, cause problems if an outdated version is used, and increase the risk of important data being lost or stolen. [\(Loshin 2008\)](#).

Because digital data can be easily copied, data governance policies might require that all sensitive data be anonymized or encrypted to reduce the risk of privacy breaches. To identify the source of a data breach or to facilitate the assertion of a copyright infringement claim a digital watermark can be embedded in digital resources.<sup>47</sup>

### *Governance in Scientific Organizing Systems*

Scientific data poses special [governance](#) problems because of its enormous scale, which dwarfs the datasets managed in most business organizing systems. A scientific data collection might contain tens of millions of files and many petabytes of data. Furthermore, because scientific data is often created using specialized equipment or computers and undergoes complex workflows, it can be necessary to curate the technology and processing context along with data in order to preserve it. An additional barrier to effective scientific data curation is the lack of incentives in scientific culture and publication norms to invest in data retention for reuse by others.<sup>48</sup>

47. [\(Cox et al. 2007\)](#).

48. Recently imposed requirements by the National Science Foundation(NSF), National Institute of Health(NIH) and other research granting agencies for researchers to submit “data management plans” as part of their proposals should make digital data curation a much more important concern [\(Borgman 2011\)](#). (NSF Data Management Plan Requirements: <http://www.nsf.gov/eng/general/dmp.jsp>).

### The Long Tail of Dark Data

Almost all scientists admit that they are holding “dark data,” data that has never been made available to the rest of the scientific community. There may only be a few scientists worldwide that would want to see a particular dataset, but there are many thousands of these datasets. Other dark data comes from research that fails to find effects; because these negative findings are less likely to be published, literature reviews can be skewed by their omission. Just as Netflix makes the long tail of movies available, perhaps dark data would become more accessible if it could be easily uploaded to a Netflix for Science. ([Heidorn 2008](#))

## 24. Key Points in Chapter Three

- Which activities are common to all organizing systems?

Selection, organizing, interaction design, and maintenance activities occur in every organizing system.

(See [“Introduction”](#))

- Are selection, organizing, interaction design, and maintenance the same activities in every organizing systems?

These activities are not identical in every domain, but the general terms enable communication and learning about domain-specific methods and vocabularies.

(See [“Introduction”](#))

- What is the first decision to be made when creating an organizing system?

The most fundamental decision for an organizing system is determining its resource domain, the group or type of resources that are being organized.

(See [“Selecting Resources”](#))

- Why does selection by memory institutions differ from sampling in scientific research?

Memory institutions select rare and distinctive resources, but in scientific research, a sample must contain representative instances.

(See [“Selecting Resources”](#))

- Does making selection principles clear and consistent ensure that they are good ones?

Even when the selection principles behind a collection are clear and consistent, they can be unconventional, idiosyncratic, or otherwise biased.

(See [“Selection Criteria”](#))

- How does “looking upstream” support better resource selection?

If you can determine where the resources come from, you can make better selection decisions by evaluating the people, processes, and organizing systems that create them.

(See [“Looking “Upstream” and “Downstream” to Select Resources”](#))

- What is a resource property?

In this book we use *property* in a generic and ordinary sense as a synonym for [feature](#) or “characteristic.” Many cognitive and computer scientists are more precise in defining these terms and reserve [property](#) for binary predicates (e.g., something is red or not, round or not). If multiple values are possible, the [property](#) is called an [attribute](#), “dimension,” or “variable.”

(See [“Organizing Resources”](#))

- What is the relationship between resource properties and organizing principles?

Most organizing systems use principles that are based on specific resource properties or properties derived from the collection as a whole.

(See [“Organizing Resources”](#))



- How do businesses differ in the ways they organize people?

There are a huge number of ways to organize people that differ in the extent of hierarchical structure, the flow of information up and down the hierarchy, the span of control for managers, and the discretion people have to deviate or innovate with respect to the work they have been assigned to do.

(See the sidebar, [Organizing People into Businesses](#))

- What problems can arise when arranging physical resources?

Some arrangements of physical resources are constrained or precluded by resource properties that might cause problems for other resources or for their users.

(See [“Organizing with Properties of Physical Resources”](#))

- What are some of the ways in which the mind follows Gestalt principles and imposes simpler interpretations on visual sensations?

There are always multiple interpretations of the sensory stimuli gathered by our visual system, but the mind imposes the simplest ones: things near each other are grouped, complex shapes are viewed as simple shapes that are overlapping, missing information needed to see separate visual patterns as continuous or whole is filled in, and ambiguous figure-ground illusions are given one interpretation at a time.

(See the sidebar, [Gestalt Principles](#))

- How can built environments influence the expectations, behaviors, and experiences of everyone who enters the space?

Built environments can be designed to encourage or discourage interactions between people, to create a sense of freedom or confinement, to reward exploration or enforce efficiency.

(See [“Organizing Built Environments”](#))

- How can we define the activity of “Information Architecture” using the language of the discipline of organizing?

It is straightforward from the perspective of the discipline of organizing to define the activity of information architecture as designing an abstract and effective organization of information and then exposing that organization to facilitate navigation and information use.

(See [“Information Architecture” and Organizing Systems](#))

- What is materiality?

An emerging issue in the field of digital humanities is the requirement to recognize the *materiality* of the environment that enables people to create and interact with digital resources

(See [“Organizing Digital Resources”](#))

- Why is the level of measurement important when organizing numeric data?

The level of measurement (nominal, ordinal, interval, or ratio) of data determines how much quantitative organization of your data will be sensible.

(See [“Organizing With Descriptive Statistics”](#))

- How can statistics help organize a set of resources?

Statistical descriptions summarize a set of resources, and reveal other details that enable comparison of instances with the collection as a whole (such as identifying outliers).

(See [“Organizing With Descriptive Statistics”](#))

- What factors affect the organization of resources?

Multiple properties of the resources, the person organizing or intending to use them, and the social and technological environment in which they are being organized can collectively shape their organization.

(See [“Organizing with Multiple Resource Properties”](#))

- What is the fundamental tradeoff faced when organizing physical resources?

The tradeoff between the amount of work that goes into organizing a collection of resources and the amount of work required to find and use them is inescapable when the resources are physical objects or information resources are in physical form.

(See [“Affordance and Capability”](#))

- What are affordance and capability?

The concept of *affordance*, introduced by J. J. Gibson, then extended and popularized by Donald Norman, captures the idea that physical resources and their environments have inherent actionable properties that determine, in conjunction with an actor’s capabilities and cognition, what can be done with the resource.

(See [“Affordance and Capability”](#))

- Does supporting more interactions mean that an organizing system has more capability?

We should not assume that supporting more types of interactions necessarily makes a system better or more capable; what matters is how much value is created or invoked in each interaction.

(See [“Affordance and Capability”](#))

- What does it mean for a resource to be accessible?

A resource is only accessible when it supports interactions, and it is ineffective design to implement interactions with resources that some people are unable to perform.

(See [“Affordance and Capability”](#))

- Why are techniques for transforming the format of a resource or its description important in achieving accessibility?

Many of the techniques for making a resource accessible involve transforming the resource or its description into a different form so someone who could not perceive it or interact with it in its original form can now do so.

(See [“Affordance and Capability”](#))

- What is the basis of value creation when interacting with a digital resource?

With digital resources, the essence of the interaction is information exchange or symbolic manipulation of the information contained in the resource.

(See [“Value Creation with Digital Resources”](#))

- What factors improve the usability of digital resources?

The variety and functions of interactions with digital resources are determined by the amount of structure and semantics represented in their digital encoding, in the descriptions associated with the resources, or by the intelligence of the computational processes applied to them.

(See [“Value Creation with Digital Resources”](#))

- What is preservation?

Preservation of resources means maintaining them in

conditions that protect them from physical damage or deterioration.

(See [“Preservation”](#))

- What is the relationship between digitization and preservation?

Preservation is often a key motive for digitization, but digitization alone is not preservation.

(See [“Digitization and Preserving Resources”](#))

- What are curation and governance?

The essence of curation and governance is having clear policies for collecting resources and maintaining them over time that enable people and automated processes to ensure that resource descriptions or data are authoritative, accurate, complete, consistent, and non-redundant.

(See [“Curation”](#) and [“Governance”](#))

- In what ways can computation improve the maintenance of resources?

Data cleaning algorithms can eliminate duplicate data, search engines can improve the relevance of results using selection and navigation behavior, and sensor data can predict when machines need servicing.

(See [“Computational Curation”](#))

- For what reasons is discarding resources an essential maintenance activity?

An essential part of maintenance is the phasing out of resources that are damaged or unusable, expired or past their effectivity dates, or no longer relevant to any interaction.

(See [“Discarding, Removing, and Not Keeping”](#))

- What is the role of governance in business organizing systems?

Governance is essential to deal with frequent changes in business organizing systems, data quality management, access control to ensure security and privacy, compliance, deletion, and archiving.

(See [“Governance in Business Organizing Systems”](#))

- How is governance different in scientific organizing systems?

Scientific data poses special governance problems because of its scale.

(See [“Governance in Scientific Organizing Systems”](#))



PART IV

# RESOURCES IN ORGANIZING SYSTEMS

**Robert J. Glushko**

**Daniel D. Turner**

**Kimra McPherson**

**Jess Hemerly**





## 25. Introduction (IV)

This chapter builds upon the foundational concepts introduced in [Foundations for Organizing Systems](#) to explain more carefully what we mean by [resource](#). In particular, we focus on the issue of identity—what will be treated as a separate resource—and discuss the issues and principles we need to consider when we give each resource a name or identifier.

### Navigating This Chapter

In [“Four Distinctions about Resources”](#) we introduce four distinctions we can make when we discuss resources: [domain](#), [format](#), [agency](#), and [focus](#). In [“Resource Identity”](#) we apply these distinctions as we discuss how resource identity is determined for physical resources, bibliographic resources, resources in information systems, as well as for [active resources](#) and [smart things](#). [“Naming Resources”](#) then tackles the problems and principles for naming: once we have

identified resources, how do we name and distinguish them? Finally, [“Resources over Time”](#) considers issues that emerge with respect to resources over time.

## What Is a Resource?

[Resources](#) are what we [organize](#).

We introduced the concept of [resource](#) in [“The Concept of “Resource””](#) with its ordinary sense of “anything of value that can support goal-oriented activity” and emphasized that a group of resources can be treated as a [collection](#) in an organizing system. And what do we mean by “anything of value,” exactly? It might seem that the question of [identity](#), of what a single resource is, should not be hard to answer. After all, we live in a world of resources, and finding, selecting, describing, arranging, and referring to them are everyday activities. And while human resources are not a primary focus of this book, it would be remiss not to explain why it makes sense to think of people that way. See the sidebar, [People as Resources](#).

Nevertheless, even when the resources we are dealing with are tangible things, how we go about organizing them is not always obvious, or at least not obvious to each of us in the same way at all times. Not everyone thinks of them in the same way. Recognizing something in the sense of perceiving it as a tangible thing is only the first step toward being able to organize it and other resources like it. Which properties garner our attention, and which we use in organizing depends on our experiences, purposes, and context.

We add information to a resource when we name it or describe it; it then becomes more than “it.” We can describe the same resource

in many different ways. At various times we can consider any given resource to be one of many members of a broad category, as one of the few members of a narrow category, or as a unique instance of a category with only one member. For example, we might recognize something as a piece of clothing, as a sock, or as the specific dirty sock with the hole worn in the heel from yesterday's long hike.

However, even after we categorize something, we might not be careful how we talk about it; we often refer to two objects as “the same thing” when what we mean is that they are “the same type of thing.” Indeed, we could debate whether a category with only one possible member is really a category, because it blurs an important distinction between particular items or instances and the class or type to which they belong.

The issues that matter and the decisions we need to make about resource instances and resource classes and types are not completely separable. Nevertheless, we will strive to focus on the former ones in this chapter and the latter ones in [\*Categorization: Describing Resource Classes and Types\*](#).

### *Resources with Parts*

As tricky as it can be to decide what a resource is when you are dealing with single objects, it is even more challenging when the resources are objects or systems composed of other parts. In these cases, we must focus on the entirety of the object or system and treat it as a resource, treat its constituent parts as resources, and deal with the relationships between the parts and the whole, as we do with engineering drawings and assembly procedures.

How Many Things is a Chess Set?

A chess set exemplifies the many different ways to decide what to count as a separate resource. Is this a chess set, two sets of chess pieces, six types of chess pieces (1 king, 1 queen, 2 rooks, 2 bishops, 2 knights, 8 pawns for each side), or 33 separate things (the 32 pieces and a board on which to play the game)?



([Photo by Emma Jane Hogbin Westby](#). Creative Commons CC-BY-2.0 license.)

How many things is a car? If you are imagining the car being assembled you might think of several dozen large parts like the frame, suspension, drive train, gas tank, brakes, engine, exhaust system, passenger compartment, doors, and other pre-assembled components. Of course, each of those components is itself made up of many parts—think of the engine, or even just the radio. Some sources have counted ten or fifteen thousand parts in the average car, but even at that precise granularity a lot of parts are still complex things. There are screws and wires and fasteners and on and on; really too many to count.

Ambiguity about the number of parts in the whole holds for information resources too; a newspaper can be considered a single resource but it might also consist of multiple sections, each of which contains separate stories, each of which has many paragraphs, and so on. Similarly, while a web page can be treated as a single resource, it can also be considered as a collection of more granular parts, each of which can be separately identified

as the source or anchor of a link. Likewise, a bank's credit card application might ask about outstanding loans, payment history, current income, and other information, or the bank might just look up your credit score, which is a statistical index that combines this financial information into a single number.

### *Bibliographic Resources, Information Components, and “Smart Things” as Resources*

Information resources generally pose additional challenges in their identification and description because their most important property is usually their content, which is not easily and consistently recognizable. Organizing systems for information resources in physical form, like those for libraries, have to juggle the duality of their tangible embodiment with what is inherently an abstract information resource; that is, the printed book versus the knowledge the book contains. Here, the organizing system emphasizes description resources or surrogates, like bibliographic records that describe the information content, rather than their physical properties.

Another question about resource that is especially critical in libraries is: What set of resources should be treated as the same [work](#) because they contain essentially similar intellectual or artistic content? We may talk about Shakespeare's play [Macbeth](#), but what is this thing we call “Macbeth”? Is it a particular string of words, saved in a computer file or handwritten upon a folio? Is it the collection of words printed with some predetermined font and pagination? Are all the editions and printings of these words the same [Macbeth](#)? How should we organize the numerous live and recorded performances of plays and movies that share the Macbeth name? What about creations based on or inspired by [Macbeth](#) that do not share the title “Macbeth,” like the Kurosawa film “[Kumonosu-jo](#)” (Throne of Blood)

that transposes the plot to feudal Japan? Patrick Wilson proposed a genealogical analogy, characterizing a [work](#) as “a group or family of texts,” with the idea that a creation like Shakespeare’s [Macbeth](#) is the “ancestor of later members of the family.”<sup>1</sup>

Information system designers and architects face analogous design challenges when they describe the [information components](#) in business or scientific organizing systems. Information content is intrinsically merged or confounded with structure and presentation whenever it is used in a specific instance and context. From a logical perspective, an order form contains [information components](#) for ITEM, CUSTOMER NAME, ADDRESS, and PAYMENT INFORMATION, but the arrangement of these components, their type font and size, and other non-semantic properties can vary a great deal in different order forms and even across a single information system that re-purposes these components for letters, delivery notices, mailing labels, and database entries.<sup>2</sup>

1. ([Shakespeare 1623](#)), ([Kurosawa 1957](#)), and ([Wilson 1968, p. 9](#)).
2. Separating information content from its structure and presentation is essential to re-purposing it for different scenarios, applications, devices, or users. The global information economy is increasingly driven by automated information exchange between business processes. When information flows efficiently from one type of document to another in this chain of related documents, the overlapping content components act as the “glue” that connects the information systems or web

Similar questions about resource identity are posed by the emergence of ubiquitous or pervasive computing, in which information processing capability and connectivity are embedded into physical objects, in devices like smart phones, and in the surrounding environment. Equipped with sensors, radio-frequency identification (RFID) tags, [GPS](#) data, and user-contributed metadata, these [smart things](#) create a jumbled torrent of information about location and other properties that must be sorted into identified streams and then matched or associated with the original resource.

[“Resource Identity”](#) discusses the issues and methods for determining “what is a [resource](#)?” for physical resources, as well as for the [bibliographic resources](#), [information components](#) and [smart things](#) discussed here, in [“Resources with Parts”](#).

## Identity, Identifiers, and Names

The answer to the question posed in [“What Is a Resource?”](#) has two parts.

- The first part is [identity](#): what thing are we treating as the resource?
- The second part is [identification](#): differentiating between this single resource and other resources like it.

These problems are closely related. Once you have decided what to treat as a resource, you create a name or an identifier so that you can refer to it reliably. A *name* is a label for a resource that

services that produce and consume the documents. [\(Glushko and McGrath 2005\)](#).



is used to distinguish one from another. An *identifier* is a special kind of name assigned in a controlled manner and governed by rules that define possible values and naming conventions. For a digital resource, its identifier serves as the input to the system or function that determines its location so it can be retrieved, a process called *resolving* the identifier or [resolution](#).

Choosing names and identifiers—be it for a person, a service, a place, a trend, a work, a document, a concept, etc.—is often challenging and highly contentious. Naming is made difficult by countless factors, including the audience that will need to access, share, and use the names, the limitations of language, institutional politics, and personal and cultural biases.

A common complication arises when a resource has more than one name or identifier. When something has more than one name each of the multiple names is a [synonym](#) or [alias](#). A particular physical instance of a book might be called a hardcover or paperback or simply a text. George Furnas and his research collaborators called this issue of multiple names for the same resource or concept the [vocabulary problem](#).<sup>3</sup>

Whether we call it a book or a text, the resource will usually have a Library of Congress catalog number as well as an ISBN as an [identifier](#). When the book is in a carton of books being shipped from the publisher to a bookstore or library, that carton will have a bar-coded tracking number assigned by the delivery service, and a manifest or receipt document created by the publisher whose identifier associates the shipment with the customer. Each of these identifiers is unique with respect to some established scope or context.

### 3. [\(Furnas, Landauer, Gomez, and Dumais 1987\)](#).

A partial solution to the [vocabulary problem](#) is to use a [controlled vocabulary](#). We can impose rules that standardize the way in which names and labels for resources are assigned in the first place. Alternatively, we can define mappings from terms used in our natural language to the authoritative or controlled terms. However, vocabulary control cannot remove all ambiguity. Even if a passport or national identity system requires authoritative full names rather than nicknames, there could easily be more than one Robert John Smith in the system.

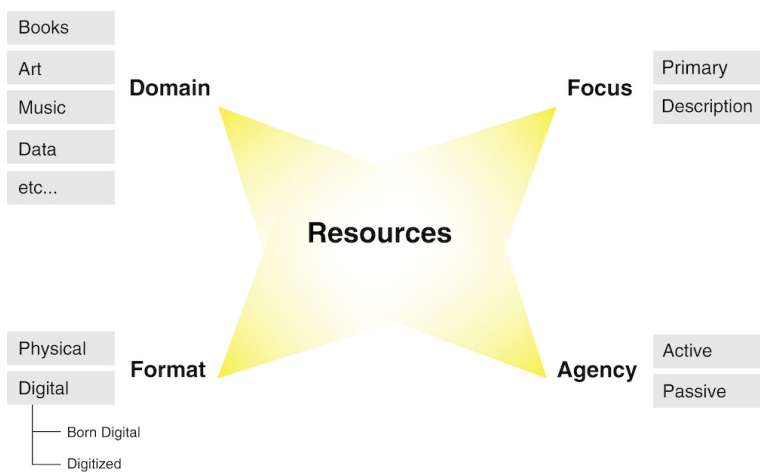
Controlling the language used for a particular purpose raises other questions: Who writes and enforces these rules? What happens when organizing systems that follow different rules get compared, combined, or otherwise brought together in contexts different from those for which they were originally intended?



# 26. Four Distinctions about Resources

The nature of the resource is critical for the creation and maintenance of quality organizing systems. There are four distinctions we make in discussing resources: [domain](#), [format](#), [agency](#), and [focus](#). [Figure: Resource Domain, Format, Focus and Agency](#). depicts these four distinctions, perspectives or points of view on resources; because they are not independent, we cannot portray these distinctions as categories of resources.

Resource Domain, Format, Focus and Agency



Four distinctions we can make when discussing resources concern their domain (their type of matter or content), format (physical or digital), agency (active or passive), and focus (primary or description).

## Resource Domain

*Resource domain* is an intuitive notion that groups resources according to the set of natural or intuitive characteristics that distinguishes them from other resources. It contrasts with the idea of ad hoc or arbitrary groupings of resources that happen to be in the same place at some time.

For physical resources, domains can be coarsely distinguished according to the type of matter they are made of using easily perceived properties. The top-level classification of all things into the animal, vegetable, and mineral kingdoms by Carl Linnaeus in 1735 is deeply embedded in most languages and cultures to create a hierarchical system of domain categories.<sup>1</sup> Many aspects of this system of domain categories are determined by natural constraints on category membership that exist as patterns of shared and correlated properties; a resource identified as a member of one category must also be a member of another with which it shares some but not all properties. For example, a marble statue in a museum must also be a kind of material, and a fish in an aquarium must also be a kind of animal.

1. ([Linnaeus 1735](#)). Linnaeus is sometimes called the father of modern taxonomy (which is unfair to Aristotle) but he certainly deserves enormous credit for the systematic approach to biological classification that he proposed in [Systema Naturae](#), published in 1735. This seminal work contains the familiar kingdom, class, order, family, genus, species hierarchy.

### The Document Type Spectrum

Different domains or types of documents can be distinguished according to the extent to which their content is semantically prescribed, by the amount of internal structure, and by the correlations of their presentation and formatting to their content and structure. These three characteristics of content, structure, and presentation vary systematically from narrative document types like novels to transactional document types like invoices.

Narrative types are authored by people and are heterogeneous in structure and content, and their content is usually just prose and graphic elements. Their presentational characteristics carefully reinforce their structure and semantics; for example, the text of titles or major headings is large because the content is important, in contrast to the small text of footnotes. Transactional document types are usually created mechanically and, as a result, are homogeneous in structure and content; their content is largely “data”—strongly typed content with precise semantics that can be processed by computers.

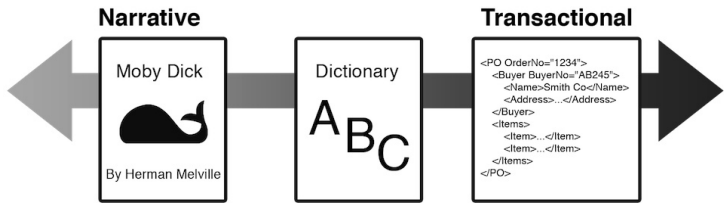
In the middle of the spectrum are hybrid document types like textbooks, encyclopedias, and technical manuals that contain a mixture of narrative text and structured content in figures, data tables, code examples, and so on.

For information resources, easily perceived properties like a book's color or size are less reliably correlated with resource domain, so we more often distinguish domains based on semantic properties;

the definitions of the “encyclopedia,” “novel,” and “invoice” resource types distinguish them according to their typical subject matter, or the type of content, rather than according to the great variety of physical forms in which we might encounter them. Arranging books by color or size might be sensible for very small collections, or in a photo studio, but organizing according to physical properties would make it extremely impractical to find books in a large library.

We can arrange types of information resources in a hierarchy. However, because the category boundaries are not sharp it is more useful to view domains of information resources on a continuum from weakly-structured narrative content to highly structured transactional content. This [framework](#), called the [Document Type Spectrum](#) by Glushko and McGrath, captures the idea that the boundaries between resource domains, like those between colors in the rainbow, are easy to see for colors far apart in the spectrum but hard to see for adjacent ones.<sup>2</sup> (See the sidebar, [The Document Type Spectrum](#), and its corresponding depiction as [Figure: Document Type Spectrum](#).)

Document Type Spectrum



The Document Type Spectrum is a continuum of document types from narrative ones that are mostly text, like novels, to transactional ones with highly-structured information, like invoices. In between

2. ([Glushko and McGrath 2005](#)).

are hybrid types that contain both narrative and transactional content, like dictionaries and encyclopedias.

## Resource Format

Information resources can exist in numerous formats with the most basic distinction between physical and digital ones. This distinction is most important in the implementation of a resource storage or preservation system because that is where physical properties are usually considerations, and very possibly constraints. This distinction is less important at the logical level when we design interactions with resources because digital surrogates for the physical resources can overcome the constraints posed by physical properties. When we search for cars or appliances in an online store it does not matter where the actual cars or appliances are located or how they are organized. (See the sidebar, [The Three Tiers of Organizing Systems](#)).

Many digital representations can be associated with either physical or digital resources, but it is important to know which one is the original or primary resource, especially for unique or valuable ones.

Today many resources in organizing systems are *born digital*, created in word processors, digital cameras, audio and video recorders. Other digital resources are by sensors in “smart things” and by the systems that create digital resources when they interact with barcodes, QR codes, RFID tags, or other mechanisms for tracking identity and location.<sup>3</sup>

Other digital resources are created by *digitization*, the process for

3. ([Kuniavsky 2010](#)).



transforming an artifact whose original format is physical so it can be stored and manipulated by a computer. We digitize the printed word, photographs, blueprints, and record albums. Printed text, for example, can be digitized by scanning pages and using character recognition software or simply re-typing it.<sup>4</sup>

There are a vast number of digital formats that differ in many ways, but we can coarsely compare them on two dimensions: the degree to which they distinguish information content from presentation or rendering, and the explicitness with which content distinctions are represented. Taken together, these two dimensions allow us

4. Project Gutenberg, begun in 1971, was the first large-scale effort to digitize books; its thousands of volunteers have created about 40,000 digital versions of classic printed works. Systematic research in digital libraries began in the 1990s when the [US National Science Foundation\(NSF\)](#), the [Advanced Research Projects Agency\(ARPA\)](#), and NASA launched a Digital Library Initiative that emphasized the enabling technologies and infrastructure. At about the same time numerous pragmatic efforts to digitize library collections began, characterized by some as a race against time as old books in libraries were literally disintegrating and turning into dust. The Internet Archive, started in 1996, now has a collection of over 3 million texts and has estimated the cost of digitizing to be about \$30 for the average book. Multiply this by the scores of millions of books held in the world's research libraries and it is easy to see why many libraries endorsed Google's offer to digitize their collections.

to compare formats on their overall “Information IQ” —with the overarching principle being that “smarter” formats contain more computer-processable information, as illustrated in [Figure: Information IQ](#).

Simple digital formats for “plain text” documents contain only the characters that you see on your computer keyboard. [ASCII](#) is the most commonly used simple format, but [ASCII](#) is inadequate for most languages, which have larger character sets, and it also cannot handle mathematical characters.<sup>5</sup> The Unicode standard was

5. The [ASCII](#) scheme was standardized in the 1960s when computer memory was expensive and most computing was in English-speaking countries, so it is minimal and distinguishes only 128 characters. ([Cerf1969](#)) [American Standard Code for Information Interchange\(ASCII\)](#) is an ANSI specification. (See <http://en.wikipedia.org/wiki/ASCII>.)

designed to overcome these limitations.<sup>6</sup> (ASCII and Unicode are discussed in great detail in [“Notations”](#).)

Most document formats also explicitly encode a hierarchy of structural components, such as chapters, sections or semantic components like descriptions or procedural steps, and sometimes the appearance of the rendered or printed form.<sup>7</sup> Another

6. Unicode 6.0 (<http://www.unicode.org/>) has room to encode 109,449 characters for all the [writing systems](#) in the world, so a single standard can represent the characters of every existing language, even “dead” ones like Sumerian and Hittite. Unicode encodes the scripts used in languages, rather than languages *per se*, so there only needs to be one representation of the Latin, Cyrillic, Arabic, etc scripts that are used for writing multiple language. Unicode also distinguishes [characters](#) from [glyphs](#), the different forms for the same character—enabling different [fonts](#) to be identified as the same character.

7. Encoding of structure in documents is valuable because titles, sections, links and other structural elements can be leveraged to enhance the user interface and navigational interactions with the digital document and enable more precise information retrieval. Some uses of documents require formats that preserve their printed appearance. “Presentational fidelity” is essential if we imagine a banker or customs inspector carefully

important distinction to note is whether the information is encoded as a sequence of text characters so that it is human readable as well as computer readable. Encoding character content with XML, for example, allows for layering of intentional coding or markup interwoven with the “plain text” content. Because XML processors are required to support Unicode, any character can appear in an XML document. The most complex digital formats are those for multimedia resources and multidimensional data, where the data format is highly optimized for specialized analysis or applications.<sup>8</sup>

Digitization of non-text resources such as film photography, drawings, and analog audio and visual recordings raises a complicated set of choices about pixel density, color depth, sampling rate, frequency filtering, compression, and numerous other technical issues that determine the digital representation.<sup>9</sup>

There may be multiple intended uses and devices for a digitized resource that could require different digitization approaches and formats. Downstream users of digitized resources need to know the format in which a digital artifact has been created, so they can reuse it as is, or process it in other ways.

Some digital formats support interactions that are qualitatively different and more powerful than those possible with physical resources. Museums are using virtual world technology to create interactive exhibits in which visitors can fly through the solar

comparing a printed document with a computer-generated one to ensure they are identical.

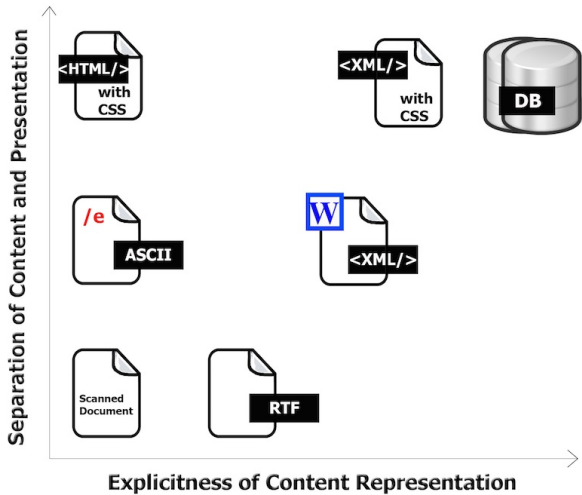
8. Text encoding specs are well-documented; see (<http://www.wotsit.org/list.asp?fc=10>).
9. ([Chapman and Chapman 2009](#)).

system, scan their own bodies, and change gravity so they can bounce off walls. Sophisticated digital document formats can enable interactions with annotated digital images or video, 3-D graphics or embedded datasets. The Google Art Project contains extremely high resolution photographs of famous paintings that make it possible to see details that are undetectable under the normal viewing conditions in museums.<sup>10</sup>

Nevertheless, digital representations of physical resources can also lose important information and capabilities. The distinctive sounds of hip hop music produced by “scratching” vinyl records on turntables cannot be produced from digital MP3 music files.<sup>11</sup>

10. The ambitious use of virtual world technology to create novel forms of interaction described by [\(Rothfarb and Doherty 2007\)](#) reflects the highly interactive character of its host museum, the Exploratorium in San Francisco (<http://www.exploratorium.edu/>). Similarly, the Google Art Project (<http://googleartproject.com>) is notable for its goal of complementing and extending, rather than merely imitating, the museum visitor’s encounter with artwork [\(Proctor 2011\)](#). A feature that let people create a “personal art collection” is very popular, enabling a fan of Vincent Van Gogh to bring together paintings that hang in different museums.
11. However, scratching can be simulated using a smart phone or tablet app called djay. See <http://www.algoriddim.com/djay>.

## INFORMATION IQ



The notion of Information IQ captures the idea that document formats differ on two dimensions: the explicitness of content representation, and the separation of content and presentation. A scanned document is just a picture of a document with neither of these distinctions, so it is low on both dimensions. A database or XML document distinguishes explicitly between types of content and presentation is separately assigned, so they are high on both dimensions and have the highest Information IQ. An HTML document's content distinctions are usually presentational and, thus, it has lower IQ. Formats with high Information IQ facilitate computer processing.

Copyright often presents a barrier to digitization, both as a matter of law and because digitization itself enables copyright enforcement to a degree not possible prior to the advent of digitization, by eliminating common forms of access and interactions that are

inherently possible with physical printed books like the ability to give or sell them to someone else.<sup>12</sup>

## Resource Agency

### People as Resources

Earlier editions of this book sidestepped the question of people as resources to avoid complicating the Discipline of Organizing. People organize themselves in innumerable ways to coexist, share knowledge, and accomplish more than they could as individuals, and

12. As a result, digital books are somewhat controversial and problematic for libraries, whose access models were created based on the economics of print publication and the social contract of the copyright first sale doctrine that allowed libraries to lend printed books. Digital books change the economics and first sale is not as well-established for digital works, which are licensed rather than sold ([Aufderheide and Jaszi 2011](#)). To protect their business models, many publishers are limiting the number of times ebooks can be lent before they “self-destruct.” Some librarians have called for boycotts of publishers in response (<http://boycottharpercollins.com>).

behaviors such as trust and reciprocity might be considered “organizing principles” for human society. But these organic relationships and interactions usually lack the intentional arrangement to be considered true Organizing Systems, except when the people are living in “intentional communities.”<sup>13</sup>

However, people do qualify as resources in Organizing Systems under our definition: just like digital and physical resources, human resources can be identified, categorized, described in terms of their attributes and relationships, and take part in interactions to create value. In businesses, people are organized to amplify their skills, knowledge, and agency. A company’s organizational chart is often a formal hierarchy in which each worker’s role is defined through his or her responsibilities and relationships to others in the company. Treating an employee abstractly as a resource with specific and predictable functions, inputs, and outputs enables employees or processes to depend upon each other without being distracted by the details of one another’s work. This so-called “black boxing” can encourage specialization and allow an organization to function more efficiently.

Nevertheless, while organizational charts are typically presented as neat hierarchies, human relationships cut across the hierarchy to create a network, and may be complicated by differing values and motivations. Conflicting incentives and lack of communication

13. [link to footnote]



between people may cause breakdowns in the system. People are more than the sum of properties used to organize them: understanding and defining employees' roles too narrowly could exclude the aspects of the job that they find rewarding or consider part of their professional identity, while black boxing people's labor and treating them as "remote person calls" also risks dehumanizing them and ignoring their working conditions.

Agency is the extent to which a resource can initiate actions on its own. We can define a continuum between completely passive resources that cannot initiate any actions and active resources that can initiate actions based on information they sense from their environments or obtain through interactions with other resources. A book being read at the beach will grow warm from absorbing the sun's energy but it has no way of measuring its temperature and is a completely passive resource. An ordinary mercury thermometer senses and displays the temperature but is not capable of communicating its own reading, whereas a digital wireless thermometer or weather station can.

Passive resources serve as nouns or operands that are acted upon, while active resources serve as verbs or operants that cause and carry out actions. We need a concept of agency to bring resources that are active information sources, or computational in character, into the organizing system framework. This concept also lets us include living resources, or more specifically, humans, into

discussions about organizing systems in a more general way that emphasizes their agency.<sup>14</sup>

### *Passive or Operand Resources*

Organizing systems that contain passive or operand resources are ubiquitous for the simple reason that we live in a world of physical resources that we identify and name in order to interact with them. *Passive resources* are usually tangible and static and thus they become valuable only as a result of some action or interaction with them.

Most organizing systems with physical resources or those that contain resources that are digitized equivalents treat those resources as passive. A printed book on a library shelf, a digital book in an ebook reader, a statue in a museum gallery, or a case of beer in a supermarket refrigerator only create value when they are checked out, viewed, or consumed. None of these resources exhibits any agency and cannot initiate any actions to create value on their own.

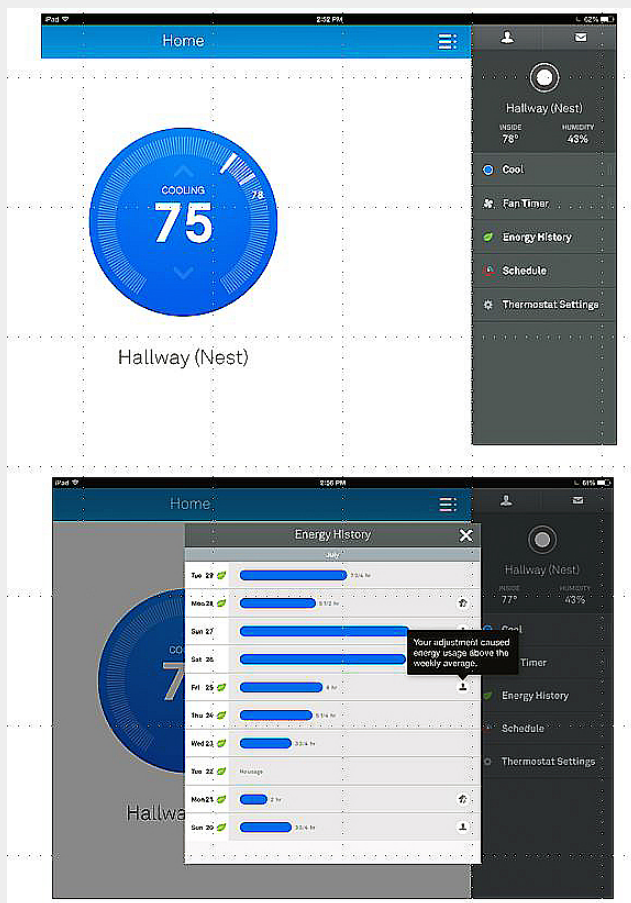
### *Active or Operant Resources*

[Active resources](#) create effects or value on their own, sometimes

14. The opposing categories of operands and operants have their roots in debates in political economics about the nature of work and the creation of value ([Vargo and Lusch 2004](#)) and have more recently played a central role in the development of modern thinking about service design ([Constantin and Lusch 1994](#)), ([Maglio et al. 2009](#)).

when they initiate interactions with passive resources. Active resources can be people, other living resources, computational agents, active information sources, web-based services, self-driving cars, robots, appliances, machines or otherwise ordinary objects like light bulbs, umbrellas, and shoes that have been made “smarter.” We can exploit computing capability, storage capacity, and communication bandwidth to create active resources that can do things and support interactions that are impossible for ordinary physical passive resources.

Active Resources:  
The Nest Thermostat “Ecosystem”



These two screenshots of the Nest iPad app show the thermostat control panel and an energy history report with a pop-up note explaining that resetting the temperature resulted in higher than average energy use on that day. The Nest thermostat serves as a hub device, communicating with lights, appliances, smoke alarms,

your car, your wearable fitness sensor, and other active resources (<https://nest.com/works-with-nest/>).

(Screenshots by Andrea Angquist. Used with permission.)

We can analyze active resources according to five capabilities that progressively increase their agency. These capabilities build on each other to give resources and the organizing systems in which they participate more ways to create value through interactions and information exchanges.

### **Sensing or awareness**

The minimal capability for a resource to have some agency is for it to be able to sense or be aware of some aspect of its environment or its interactions with other resources. A thermometer measures temperature, a photodetector measures light, a gauge measures the fuel left in a car's gas tank, a GPS device computes its location after detecting and analyzing signals from satellites, a wearable fitness sensor tracks your heartbeat and how far you walk. But sensing something in itself does not create any value in an organizing system. Something needs to be done.

### **Actuation**

A resource has the capability to actuate when it can create effects or value by initiating some action as a result of the information it senses; “actuator” is often used to describe a resource that can move or control a physical mechanism or system, while “effector” is used when the resource is a biological one. Resources can actuate by turning on lights,

speakers, cameras, motors, switches, by sending a message about the state or value of a sensor, or by moving themselves around (as with robots).

A potential or latent actuation is created when a resource can display or broadcast some aspect of its state, but value is only created if another resource (possibly human) happens to see the display or hear the broadcast and then acts upon it.

For example, RFID chips, which are essentially bar codes with built-in radio transponders, can be attached to otherwise passive resources to make them active. RFID chips begin transmitting when they detect the presence of a RFID reading device. This enables automated location tracking and context sensing. RFID receivers are built into assembly lines, loading docks, parking lots, toll booths, or store shelves to detect when some RFID-tagged resource is at some meaningful location. RFID tags can be made more useful by having them record and transmit information from sensors that detect temperature, humidity, acceleration, and even biological contamination.<sup>15</sup>

## Connectivity

For an active resource to do useful work it must be connected in some way to the actuation mechanism that manipulates or controls some other resource. This connection might be a direct and permanent one between the resource and the thing it actuates, like that of a thermostat whose temperature sensing capability has a fixed connection to a heating or cooling system that it turns off or on depending on the temperature.

An important innovation in the design of active resources is “wrapping” physical resources with software so they can be

15. See ([Allmendinger and Lombreglia 2005](#)), ([Want 2006](#)).  
([Crawford and Johnson, 2012](#))

given IP addresses and make connections with Internet protocols, which allows them to send information to an application with more capability to act on it. Such resources are said to be part of the “Internet of Things.”

Smart phones are [active resources](#) that can identify and share their own location, orientation, acceleration and a growing number of other [contextual parameters](#) to enable personalization of information services. Smart phones can also run the applications that receive messages from and send messages to other smart resources to monitor and optimize how they work.

## **Computation or programmability**

Simple active resources operate in a deterministic manner: given this sensor reading, do this. Other active resources have computational capabilities that enable them to analyze the current and historical information from their sensors, identify significant data values or patterns in these interaction resources, and then adapt their behavior accordingly.

Many thermostats are programmable, but most people do not bother to program them so they miss out on potential energy savings. Nest Labs makes a *learning thermostat* that programs itself. The Nest thermostat uses sensors for temperature, humidity, motion, and light to figure out whether people are at home, and a Wi-fi connection to get local weather data.

The Roomba vacuum cleaning robot navigates around furniture, power cords, stairs, and optimizes its cleaning paths to go over particularly dirty places. But vacuuming is all it does. More sophisticated robots are designed to be versatile and adaptable so they can repetitively perform whatever task is needed for some manufacturing process, and their capabilities can be continually upgraded by software updates, just like the apps on your smart phone. A new generation of robots typified by one called [Baxter](#) can be trained by example; a person moves

Baxter's arms and hands to show him what to do, and when Baxter has programmed himself to repeat it, he nods.

## **Composability and cooperating**

The “smartest” active resources can do more than analyze the information they collect and adapt what they do. In addition, they expose what they know and can do to other resources using standard or non-proprietary formats and protocols. This means that active resources that were independently designed and implemented can work together to create value.

Many organizing systems on the web consist of collections or configurations of active digital resources. Interactions among these active resources often implement information-intensive business models where value is created by exchanging, manipulating, transforming, or otherwise processing information, rather than by manipulating, transforming, or otherwise processing physical resources.

We are beginning to see the same principles of modularity and composability applied to physical resources, with open source software libraries for using sensors and micro-controllers and easy to use APIs. In essence, we are using software and physical resources in much the same way as functional building blocks, and standards will be critically important.

*Service Oriented Architecture*(SOA) is an emerging design discipline for organizing active software resources as functional business components that can be combined in different ways. SOA is generally implemented using web services that exchange XML documents in real-time information flows to interconnect the business service components.

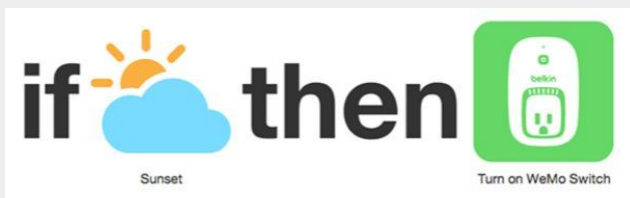
A familiar design pattern for an organizing system composed from active digital resources is the “online store.” The store can be analyzed as a composition or choreography in which some web



pages display catalog items, others serve as “shopping carts” to assemble the order, and then a “checkout” page collects the buyer’s payment and delivery information that gets passed on to other service providers who process payments and deliver the goods.

### If This, Then That

[IFTTT](#) is a visual programming system that lets non-programmers connect and control active resources in the physical and digital worlds. [IFTTT](#) programs, called recipes, can take information from a growing library of Internet services (date/time, calendar, weather, news, email, social media, and many others) and use this information with simple control logic to trigger actions in other services or resources. Example recipes can copy an Instagram photo to Google Drive, add daily Fitbit data to a spreadsheet, or control lights based on time, weather, or sunset.



“If sunset then electrical outlet on.” The icon on the left is the trigger; the icon on the right is the action.

(Photo by R. Glushko)

Design patterns for composing organizing systems from “smart” physical resources are emerging in work on the “smart home,” “smart office building,” or “smart city.” Many experiments are underway and new products emerging that are trying out different

combinations of hardware and software to understand the design tradeoffs between them to best determine where the “smarts” should go. For example, we can compare a “smart home” built around a super-intelligent hub device that communicates and coordinates with many other “not so smart” devices from the same manufacturer to one in which all of the devices are equally smart and come from different makers.

At more complex scales, a truly smart building will not just have programmable thermostats to control heating and cooling systems. It will take in weather forecasts, travel calendars, information about the cost of electricity from different sources, and other relevant information as inputs to a model of how the building heats and cools to optimize energy use and cost while keeping the rooms at appropriate temperatures.

Standard application interfaces enable active resources to interact with people to get information that might otherwise come from sensors or that enhances the value of the sensor information. A programmable thermostat that can record time-based preferences of the people who use the space controlled by the thermostat is more capable than one with just a single temperature threshold. A standard Internet protocol for communicating with the thermostat would enable it to be controlled remotely.

Open and standard data formats and communication protocols enable the aggregation and analysis of information from many instances of the same type of active resource. For example, smart phones running the Google Maps application transmit information about their speed and location. Machine learning and sophisticated optimization techniques of this dataset can yield collective intelligence that can then be given to the resources from which it was derived. In this case, Google can identify traffic jams and generate alternative routes for the drivers stuck in traffic.

### Stop and Think: The Internet of Things

There is a great deal of hype about the Internet of Things, but there is also a great deal of innovation. If you search for the phrase “Internet of Things” along with almost any physical resource, chances are you will find something. Try “baby,” “dog,” “fork,” “lettuce,” “pajamas,” “streetlamp,” and then a few of your own.

But not everything can be done best by computers. The web has enabled the use of people as [active resources](#) to carry out tasks of short duration that can be precisely described but which cannot be done reliably by computers. These tasks often require aesthetic or subjective judgment. The people doing these web-based tasks are often called “Mechanical Turks” by analogy to a fake chess playing machine from the 18th century that had a human hidden inside who secretly moved the pieces.<sup>16</sup>

16. Luis Von Ahn ([von Ahn 2004](#)) was the first to use the web to get people to perform “microwork” or “human computation” tasks when he released what he called “the ESP game” that randomly paired people trying to agree on labeling an image. Not long afterward Amazon created the MTurk platform (<http://www.mturk.com>) that lets people propose microwork and others sign up to do it, and today there are both hundreds of thousands of tasks offered and hundreds of thousands of people offering to be paid to do them.

## Resource Focus

A fourth contrast between types of resources distinguishes original or [primary resources](#) from resources that describe them. Any primary resource can have one or more [description resources](#) associated with it to facilitate finding, interacting with, or interpreting the primary one. *Description resources* are essential in organizing systems where the primary resources are not under its control and can only be accessed or interacted with through the description. [Description resources](#) are often called [metadata](#).

The distinction between [primary resources](#) and [description resources](#), or [metadata](#), is deeply embedded in library science and traditional organizing systems whose collections are predominantly text resources like books, articles, or other documents. In these contexts description resources are commonly called [bibliographic resources](#) or catalogs, and each primary resource is typically associated with one or more description resources.

In business enterprises, the organizing systems for digital information resources, such as business documents, or data records created by transactions or automated processes, almost always employ resources that describe, or are associated with, large sets or classes of primary resources.<sup>17</sup>

17. For semi-structured or more narrative documents these descriptions might be authoring templates used in word processors or other office applications, document schemas in [XML](#) applications, style sheets, or other kinds of transformations that change one resource representation into another one. Primary resources that are highly and regularly structured are invariably

The contrast between [primary resources](#) and [description resources](#) is very useful in many contexts, but when we look more broadly at organizing systems, it is often difficult to distinguish them, and determining which resources are primary and which are [metadata](#) is often just a decision about which resource is currently the focus of our attention.<sup>18</sup>

For example, many Twitter users treat the 140-character message body as the primary resource, while the associated metadata about the message and sender (is it a forward, reply, link, etc.) is less important. However, for firms that use Twitter metadata to measure

organized in databases or enterprise information management systems in which a [data schema](#) specifies the arrangement and type of data contained in each field or component of the resource.

18. Describing information as “metadata” suggests that it is of secondary importance, not as essential or informative as the resource being described. This is surely the reason why the US National Security Agency and those of other governments, whose unauthorized surveillance of global communications were revealed in 2013 by Edward Snowden, often stressed that they were only collecting message metadata, not its content. Of course, information about who you communicate with and when you do so defines your social network, information that is potentially very valuable, and the NSA knows this just as Facebook and Twitter do.

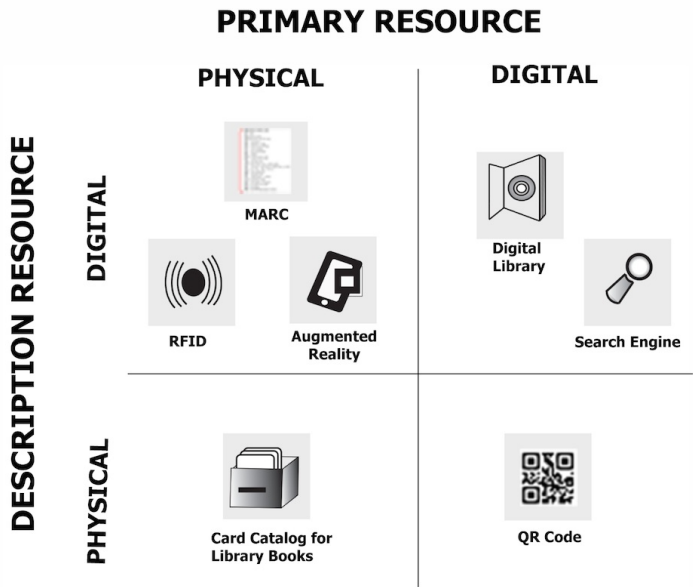
sender and brand impact, or identify social networks and trends, the focus is the metadata, not the content.<sup>19</sup>

As another example, players on professional sports teams are human resources, but millions of people participate in fantasy sports leagues where teams consist of resources based on the statistics generated by the actual human players. Put another way, the associated resources in the actual sports are treated as the primary ones in the fantasy leagues.<sup>20</sup>

## Resource Format x Focus

Applying the format contrast between physical and digital resources to the focus distinction between primary and descriptive resources yields a useful [framework](#) with four categories of resources ([Figure: Resource Format x Focus](#)).

19. There are a large number of third-party Twitter apps. See <http://twitter.pbworks.com/w/page/1779726/Apps>. For a scholarly analysis see [\(Efron 2011\)](#).
20. The basic idea behind fantasy sports is quite simple. You select a team of existing players in any sport, and then compare their statistical performance against other teams similarly selected by other people. Fantasy sports appeal mostly to die-hard fans who study player statistics carefully before “drafting” their players. The global fantasy sports business for companies who organize and operate fantasy leagues is estimated as between 1 and 2 billion US dollars annually [\(Montague 2010\)](#).



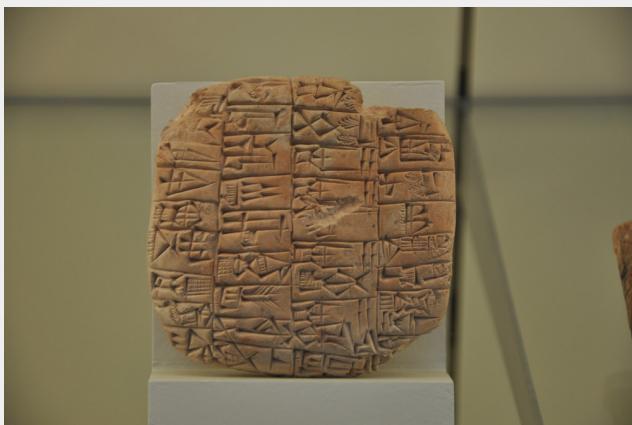
The distinctions of resource format and resource focus combine to distinguish four categories of resources: physical resources, digital resources, physical descriptions , and digital descriptions.

*Physical Description of a Primary Physical Resource*

The oldest relationship between descriptive resources and physical resources is when descriptions or other information about physical resources are themselves encoded in a physical form. Nearly ten thousand years ago in Mesopotamia small clay tokens kept in clay containers served as inventory information to count units of goods or livestock. It took 5000 years for the idea of stored tokens to

evolve into Cuneiform writing in which marks in clay stood for the tokens and made both the tokens and containers unnecessary.<sup>21</sup>

#### A Cuneiform Document at the Pergamon



The Pergamon Museum in Berlin contains a very large collection of Babylonian, Persian, and Assyrian artifacts that are nearly three thousand years old. including numerous cuneiform clay tablets like this one.

(Photo by R. Glushko.)

Printed cards served as physical description resources for books in libraries for nearly two centuries.<sup>22</sup>

21. [\(Schmandt-Besserat 1997\)](#) and [\(McGrath 2015\)](#).

22. The oldest known lists of books were created about 4000



## *Digital Description of a Primary Physical Resource*

Here, the digital resource describes a physical resource. The most familiar example of this relationship is the online library catalog used to find the shelf location of physical library resources, which beginning in the 1960s replaced the physical cards with database records. The online catalogs for museums usually contain a digital photograph of the painting, item of sculpture, or other museum object that each catalog entry describes.

Bar Code Shopping in A Virtual Supermarket

years ago in Sumeria. The first use of cards in library catalogs was literal; when the revolutionary government of France seized private book collections, an inventory was created stating in 1791 using the blank backs of playing cards. 110 years later the US Library of Congress began selling pre-printed catalog cards to libraries, but in the mid-1960s the creation of the *Machine-Readable Cataloging*(MARC) format marked the beginning of the end of printed cards. See ([Strout 1956](#)). The MARC standards are at <http://www.loc.gov/marc/>.



Woolworth's Australia created a "virtual supermarket" with product photos and bar codes. Scanning places an order, which is delivered from the customer's local supermarket.

(Photo by R. Glushko.)

Digital description resources for primary physical resources are

essential in supply chain management, logistics retailing, transportation, and every business model that depends on having timely and accurate information about where things are or about their current states. This digital description resource is created as a result of an interaction with a primary physical resource like a temperature sensor or with some secondary physical resource that is already associated with the primary physical resource like an RFID tag or barcode.

Augmented reality systems combine a layer of real-time digital information about some physical object to a digital view or representation of it. The yellow “first down” lines superimposed in broadcasts of football games are a familiar example. Augmented reality techniques that superimpose identifying or descriptive metadata are used in displays to support the operation or maintenance of complex equipment, in smart phone navigation and tourist guides, in advertising, and in other domains where users might otherwise need to consult a separate information source. Advanced airplane cockpit technology includes heads-up displays that present critical data based on available instrumentation, including augmented reality runway lights when visibility is poor because of clouds or fog.

Augmented reality displays have recently been incorporated into wearable technology like Google Glass, which mounts on eyeglass frames to display information obtained from the Internet after being requested by voice commands. Some luxury car brands have incorporated similar technology to project dashboard data, traffic conditions, and directions on the driver’s windshield.

### *Digital Description of a Primary Digital Resource*

A digital resource describes a digital resource. This is the relationship in a digital library or any web-based organizing system,

making it possible to access a primary digital resource directly from the digital secondary resource.

### *Physical Description of a Primary Digital Resource*

This is the relationship implemented when we encounter an embedded QR barcode in newspaper or magazine advertisements, on billboards, sidewalks, t-shirts, or on store shelves. Scanning the QR code with a mobile phone camera can launch a website that contains information about a product or service, place an order for one unit of the pointed-to- item in a web catalog, dial a phone number, or initiate another application or service identified by the QR code.<sup>23</sup>

23. We treat resource format and resource focus as distinct dimensions, so there are four categories here. This contrasts with David Weinberger's three "orders of order" that he proposes in the first chapter of a book called [\*Everything is Miscellaneous\* \(Weinberger 2007\)](#). Weinberger starts with the assumption that physical resources are inherently the primary ones, so the first "order of order" emerges when physical resources are arranged. The second "order of order" emerges when physical description resources are arranged, and the third "order of order" emerges when digital description resources for physical resources are arranged. Later in the book Weinberger mentions the use of bar codes associated with websites, a physical description of a digital resource, but because he started with the

assumption that physical resources define the “first order” this example does not fit into his orders of order.

## 27. Resource Identity

Determining the identity of resources that belong in a domain, deciding which properties are important or relevant to the people or systems operating in that domain, and then specifying the principles by which those properties encapsulate or define the relationships among the resources are the essential tasks when building any organizing system. In organizing systems used by individuals or with small scope, the methods for doing these tasks are often *ad hoc* and unsystematic, and the organizing systems are therefore idiosyncratic and do not scale well. At the other extreme, organizing systems designed for institutional or industry-wide use, especially in information-intensive domains, require systematic design methods to determine which resources will have separate identities and how they are related to each other. These resources and their relationships are then described in conceptual models which guide the implementation of the systems that manage the resources and support interactions with them.<sup>1</sup>

1. These methods go by different names in different disciplines, including “data modeling,” “systems analysis,” and “document engineering” (e.g., [\(Kent 2012\)](#), [\(Silverston 2000\)](#), [\(Glushko and McGrath 2005\)](#)). What they have in common is that they produce conceptual models of a domain that specify their components or parts and the relationships among these components or parts. These conceptual models are called “schemas” or “domain ontologies” in some modeling approaches, and are typically implemented in models that are optimized for particular technologies or applications.

## Identity and Physical Resources

Our human visual and cognitive systems do a remarkable job at picking out objects from their backgrounds and distinguishing them from each other. In fact, we have little difficulty recognizing an object or a person even if we are seeing them from a novel distance and viewing angle or with different lighting, shading, and so on. When we watch a football game, we do not have any trouble perceiving the players moving around the field, and their contrasting uniform colors allow us to see that there are two different teams.

The perceptual mechanisms that make us see things as permanent objects with contrasting visible properties are just the prerequisite for the organizing tasks of identifying the specific object, determining the categories of objects to which it belongs, and deciding which of those categories is appropriate to emphasize. Most of the time we carry out these tasks in an automatic, unconscious way; at other times we make conscious decisions about them. For some purposes we consider a sports team as a single resource, as a collection of separate players for others, as offense and defense, as starters and reserves, and so on.<sup>2</sup>

Although we have many choices about how we can organize football

2. Specifically, an NFL football team needs to be considered a single resource for games through the season and in playoffs, and 53 individual players for other situations, like the NFL draft or play-calling. The team and the team's roster can be thought of as resources, and the team's individual players are also resources that make up the whole team.

players, all of them will include the concept of a single player as the smallest identifiable resource. We are never going to think of a football player as an intentional collection of separately identified leg, arm, head, and body resources because there are no other ways to “assemble” a human from body parts. Put more generally, there are some natural constraints on the organization of matter into parts or collections based on sizes, shapes, materials, and other properties that make us identify some things as indivisible resources in some domain.

## Identity and Bibliographic Resources

Pondering the question of [identity](#) is something relatively recent in the world of librarians and catalogers. Libraries have been around for about 4000 years, but until the last few hundred years librarians created “bins” of headings and topics to organize resources without bothering to give each individual item a separate identifier or name. This meant searchers first had to make an educated guess as to which bin might house their desired information—“Histories”? “Medical and Chemical Philosophy”?—then scour everything in the category in a quest for their desired item. The choices were *ad hoc* and always local—that is, each cataloger decided the bins and groupings for each catalog.<sup>3</sup>

3. ([Denton 2007](#)) is a highly readable retelling of the history of cataloging that follows four themes—the use of axioms, user requirements, the [work](#), and standardization and internationalization—culminating with their synthesis in the [Functional Requirements for Bibliographic Records\(FRBR\)](#).



The first systematic approach to dealing with the concept of [identity](#) for bibliographic resources was developed by Antonio Panizzi at the British Museum in the mid-19th century. Panizzi wondered: How do we differentiate similar objects in a library catalog? His solution was a catalog organized by author name with an index of subjects, along with his newly concocted [Rules for the Compilation of the Catalogue](#). This contained 91 rules about how to identify and arrange author names and titles and what to do with anonymous works. The rules were meant to codify how to differentiate and describe each singular resource in his library. Taken together, the rules serve to group all the different editions and versions of a work together under a single identity.<sup>4</sup>

The concept of identity for bibliographic resources was refined in the 1950s by Lubetzky, who enlarged the concept of [the work](#) to make it a more abstract idea of an author's intellectual or artistic creation. According to Lubetzky's principle, an audio book, a video recording of a play, and an electronic book should be listed each as distinct items, yet still linked to the original because of their overlapping intellectual origin.<sup>5</sup>

4. This was a surprisingly controversial activity. Many opposed Panizzi's efforts as a waste of time and effort because they assumed that "building a catalog was a simple matter of writing down a list of titles"([Denton 2007, p. 38](#)).
5. Seymour Lubetzky worked for the US Library of Congress from 1943-1960 where he tirelessly sought to simplify the proliferating mass of special case cataloging rules proposed by the American Library Association, because at the time the [LOC](#) had the task of applying those rules

The distinctions put forth by Panizzi, Lubetzky, Svenonius and other library science theorists have evolved today into a four-step abstraction hierarchy (see [Figure: The FRBR Abstraction Hierarchy](#).) between the abstract *work*, an *expression* in multiple formats or genres, a particular *manifestation* in one of those formats or genres, and a specific physical *item*. The broad scope from the abstract work to the specific item is essential because organizing systems in libraries must organize tangible artifacts while expressing the conceptual structure of the domains of knowledge represented in their collections.

This hierarchy is defined in the [Functional Requirements for Bibliographical Records\(FRBR\)](#), published as a standard by the International Federation of Library Associations and Institutions(IFLA).<sup>6</sup>

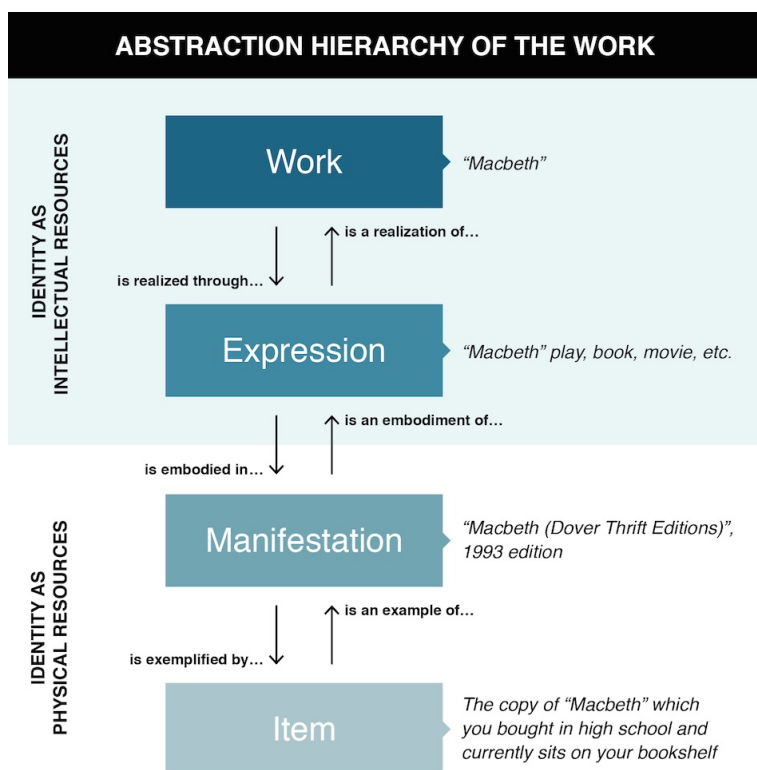
and making the catalog cards other libraries used. Lubetzky's book on [Cataloguing Rules and Principles \(Lubetzky 1953\)](#) bluntly asks "Is this rule necessary?" and was a turning point in cataloging.

6. In between the abstraction of the [work](#) and the specific single [item](#) are two additional levels in the FRBR abstraction hierarchy. An [expression](#) denotes the multiple the multiple realizations of a work in some particular medium or notation, where it can actually be perceived. There are many editions and translations of [Macbeth](#), but they are all the same expression, and they are a different expression from all of the film adaptations of [Macbeth](#). A [manifestation](#) is the set of physical artifacts with the

If we revisit the question “What is this thing we call [Macbeth](#)?” we can see how different ways of answering fit into this abstraction hierarchy. The most specific answer is that “ [Macbeth](#)” is a specific [item](#), a very particular and individual resource, like that dog-eared paperback with yellow marked pages that you owned when you read [Macbeth](#) in high school. A more abstract answer is that [Macbeth](#) is an idealization called a *work*, a category that includes all the plays, movies, ballets, or other intellectual creations that share a recognizable amount of the plot and meaning from the original Shakespeare play.

### The FRBR Abstraction Hierarchy

same expression. All of the copies of the Folger Library print edition of [Macbeth](#) are the same manifestation.



The abstraction hierarchy for identifying resources yields four different answers about the identity of an information resource.

## Identity and Information Components

In information-intensive domains, documents, databases, software applications, or other explicit repositories or sources of information are ubiquitous and essential to the creation of value for the user, reader, consumer, or customer. Value is created through the comparison, compilation, coordination or transformation of information in some chain or choreography of processes operating

on information flowing from one information source or process to another. These processes are employed in accounting, financial services, procurement, logistics, supply chain management, insurance underwriting and claims processing, legal and professional services, customer support, computer programming, and energy management.

The processes that create value in information-intensive domains are “glued together” by shared [information components](#) that are exchanged in documents, records, messages, or resource descriptions of some kind. Information components are the primitive and abstract resources in information-intensive domains. They are the units of meaning that serve as building blocks of composite descriptions and other information artifacts.

The value creation processes in information-intensive domains work best when their component parts come from a common [controlled vocabulary](#) for components, or when each uses a vocabulary with a granularity and semantic precision compatible with the others. For example, the value created by a personal health record emerges when information from doctors, clinics, hospitals, and insurance companies can be combined because they all share the same “patient” component as a logical piece of information.

This abstract definition of information components does not help identify them, so we will introduce some heuristic criteria: An [information component](#) can be: (1) Any piece of information that has a unique label or identifier or (2) Any piece of information that is self-contained and comprehensible on its own.<sup>7</sup>

7. This kind of advice can be found in many data or conceptual modeling texts, but this particular statement comes from [\(Glushko, Weaver, Coonan, and Lincoln 1988\)](#). Similar advice can also be found in the information

These two criteria for determining the identity of information components are often easy to satisfy through observations, interviews, and task analysis because people naturally use many different types of information and talk easily about specific components and the documents that contain them. Some common components (e.g., person, location, date, item) and familiar document types (e.g., report, catalog, calendar, receipt) can be identified in almost any domain. Other components need to be more precisely defined to meet the more specific semantic requirements of narrower domains. These smaller or more fine-grained components might be viewed as refined or qualified versions of the generic components and document types, like course grade and semester components in academic transcripts, airport codes and flight numbers in travel itineraries and tickets, and drug names and dosages in prescriptions.

Decades of practical and theoretical effort in conceptual modeling, relational theory, and database design have resulted in rigorous methods for identifying [information components](#) when requirements and business rules for information can be precisely specified. For example, in the domain of business transactions, required information like item numbers, quantities, prices, payment information, and so on must be encoded as a particular type of data—integer, decimal, Unicode string, etc.—with clearly defined possible values and that follows clear occurrence rules.<sup>8</sup>

science literature: “A unit of information...would have to be...correctly interpretable outside any context” ([Wilson 1968, p. 18](#)).

8. A group of techniques collectively called “normalization” produces a set of tightly defined [information components](#) that have minimal redundancy and ambiguity. Imagine

that a business keeps information about customer orders using a “spreadsheet” style of organization in which a row contains cells that record the date, order number, customer name, customer address, item ID, item description, quantity, unit price, and total price. If an order contains multiple products, these would be recorded on additional rows, as would subsequent orders from the same customer. All of this information is important to the business, but this way of organizing it has a great deal of redundancy and inefficiency. For example, the customer address recurs in every order, and the customer address field merges street, city, state and zip code into a large unstructured field rather than separating them as atomic components of different types of information with potentially varying uses. Similar redundancy exists for the products and prices. Canceling an order might result in the business deleting all the information it has about a particular customer or product.

Normalization divides this large body of information into four separate tables, one for customers, one for customer orders, one for the items contained in each order, and one for item information. This normalized information model encodes all of the information in the “spreadsheet style” model, but eliminates the redundancy and avoids the data integrity problems that are inherent in it.

Identifying components can seem superficially easy at the transactional end of the Document Type Spectrum (see the sidebar in [“Resource Domain”](#)), with orders or invoices, forms requiring data entry, or other highly-structured document types like product catalogs, where pieces of information are typically labeled and delimited by boxes, lines, white space or other presentation features that encode the distinctions between types of content. For example, the presence of ITEM, CUSTOMER NAME, ADDRESS, and PAYMENT INFORMATION labels on the fields of an online order form suggests these pieces of information are semantically distinct components in a retail application. In addition, these labels might have analogues in variable names in the source code that implements the order form, or as tags in a [XML](#) document created by the ordering application; `<CustName>John Smith</CustName>` and `<Item>A-19</Item>` in the order document can be easily identified when it is sent to the other services by the order management application.

But the theoretically grounded methods for identifying components like those of relational theory and normalization that work for structured data do not strictly apply when information requirements are more qualitative and less precise at the narrative end of the Document Type Spectrum. These information requirements are typical of narrative, unstructured and semi-structured types of documents, and information sources like those often found in law, education, and professional services. Narrative documents include technical publications, reports, policies, procedures and other less structured information, where semantic

Normalization is taught in every database design course. The concept and methods were proposed by [\(Codd 1970\)](#), who invented the relational data model, and has been taught to students in numerous database design textbooks like [\(Date 2003\)](#).



components are rarely labeled explicitly and are often surrounded by text that is more generic. Unlike transactional documents that depend on precise semantics because they are used by computers, narrative documents are used by people, who can ask if they are not sure what something means, so there is less need to explicitly define the meaning of the [information components](#). Occasional exceptions, such as where components in narrative documents are identified with explicit labels like NOTE and WARNING, only prove the rule.

## Identity and Active Resources

[Active resources](#) (“[Use Controlled Vocabularies](#)”) initiate effects or create value on their own. In many cases an inherently passive physical resource like a product package or shipping pallet is transformed into an active one when associated with an RFID tag or bar code. Mobile phones contain device or subscriber IDs so that any information they communicate can be associated both with the phone and often, through indirect reference, with a particular person. If the resource has an IP address, it is said to be part of the “Internet of Things.”<sup>9</sup>

9. The “Internet of Things” concept spread very quickly after it was proposed in 1999 by Kevin Ashton, who co-founded the Auto-ID center at MIT that year to standardize [RFID](#) and sensor information. For a popular introduction, see [\(Gershenfeld, Krikorian, and Cohen 2004\)](#). For a recent technical survey and a taxonomy of application domains and scenarios see [\(Atzori, Iera, and Morabito 2010\)](#).

Organizing systems that create value from active resources often co-exist with or complement organizing systems that treat its resources as passive. In a traditional library, books sat passively on shelves and required users to read their spines to identify them. Today, some library books contain active RFID tags that make them dynamic information sources that self-identify by publishing their own locations. Similarly, a supermarket or department store might organize its goods as physical resources on shelves, treating them as passive resources; superimposed on that traditional organizing system is one that uses point-of-sale transaction information created when items are scanned at checkout counters to automatically re-order goods and replenish the inventory at the store where they were sold. In some stores the shelves contain sensors that continually “talk to the goods” and the information they gather can maintain inventory levels and even help prevent theft of valuable merchandise by tracking goods through a store or warehouse. The inventory becomes a collection of active resources; each item eager to announce its own location and ready to conduct its own sale. Another category of inanimate objects that are active resources are those that use Twitter to communicate their status or sensor measurements. These include bridges, rivers, and the Curiosity Rover on Mars.

#### Big Data Makes “Smart” Soccer Players

The German World Cup soccer team, which won the 2014 World Cup, took advantage of sophisticated data collection and analysis to optimize player skill and strategy training. German software firm SAP analyzed video data from on-field cameras that captured thousands of data points per second about player position and movement to identify improvements in passing and ball handling for German players and detect

weaknesses in opponents. German sports equipment firm Adidas designed cleats with sensors that track mileage, field position, and movements. ([Norton 2014](#)) and ([Reynolds 2014](#)).

The extent to which an active resource is “smart” depends on how much computing capability it has available to refine the data it collects and communicates. A large collection of sensors can transmit a torrent of captured data that requires substantial processing to distinguish significant events from those that reflect normal operation, and also from those that are statistical outliers with strange values caused by random noise. This challenge gets qualitatively more difficult as the amount of data grows to [big data](#) size, because a one in million event might be a statistical outlier that can be ignored, but if there are a thousand similar outliers in a billion sensor readings, this cluster of data probably reveals something important. On the other hand, giving every sensor the computing capability to refine its data so that it only communicates significant information might make the sensors too expensive to deploy.<sup>10</sup>

10. Pattern analysis can help escape this dilemma by enabling predictive modeling to make optimal use of the data. In designing [smart things](#) and devices for people, it is helpful to create a smart model in order to predict the kinds of patterns and locations relevant to the data collected or monitored. These allow designers to develop a set of dimensions and principles that will act as smart

guides for the development of smart things. Modeling helps to enable automation, security, or energy efficiency, and baseline models can be used to detect anomalies. As for location, exact locations are unnecessary; use of a “symbolic space” to represent each “sensing zone”—e.g., rooms in a house—and an individual’s movement history as a string of symbols—e.g., abcdegia—works sufficiently as a model of prediction. See [\(Das et al. 2002\)](#).

## 28. Naming Resources

Determining the identity of the thing, document, information component, or data item we need is not always enough. We often need to give that resource a name, a label that will help us understand and talk about what it is. But naming is not just the simple task of assigning a sequence of characters. In this section, we will discuss why we name, some of the problems with naming, and the principles that help us name things in useful ways.

### What's in a Name?

When a child is born, its parents give it a name, often a very stressful and contentious decision. Names serve to distinguish one person from another, although names might not be unique—there are thousands of people named James Smith and Maria Garcia. Names also, intentionally or unintentionally, suggest characteristics or aspirations. The name given to us at birth is just one of the names we will be identified with during our lifetimes. We have nicknames, names we use professionally, names we use with friends, and names we use online. Our banks, our schools, and our governments will know who we are because of numbers they associate with our names. As long as it serves its purpose to identify you, your name could be anything.<sup>1</sup>

1. Well, maybe not anything. Books list traditional meanings of various names, charts rank names by popularity in different eras, and dozens of websites tout themselves as the place to find a special and unique name. See

Resources other than people need names so we can find them, describe them, reuse them, refer or link to them, record who owns them, and otherwise interact with them. In many domains the names assigned to resources are also influenced or constrained by rules, industry practice, or technology considerations.

<http://www.ssa.gov/oact/babynames/> for historical trends about baby names in the US with an interactive visualization at <http://www.babynamewizard.com/voyager#>.

Different countries have rules about characters or words that may be used in names. In Germany, for example, the government regulates the names parents can give to their children; there's even a book, the [\*International Handbook of Forenames\*](#), to guide them ([Kulish 2009](#)). In Portugal, the Ministry of Justice publishes lists of prohibited names (BBC News, 2007a). Meanwhile, in 2007, Swedish tax officials rejected a family's attempt to name their daughter Metallica (<http://news.bbc.co.uk/2/hi/6525475.stm>).

We can also change our names. Whether a woman takes on her husband's surname after marriage or, like the California man who changed his name to "Trout Fishing," we just find something that better suits us than our given name.

# The Problems of Naming

Giving names to anything, from a business to a concept to an action, can be a difficult process and it is possible to do it well or do it poorly. The following section details some of the major challenges in assigning a name to a resource.

## *The Vocabulary Problem*

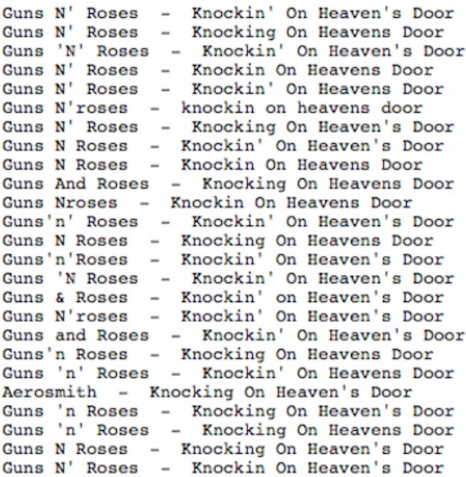
Every natural language offers more than one way to express any thought, and in particular there are usually many words that can be used to refer to the same thing or concept. The words people choose to name or describe things are embodied in their experiences and context, so people will often disagree in the words they use. Moreover, people are often a bit surprised when it happens, because what seems like the natural or obvious name to one person is not natural or obvious to another. One way to avoid surprises is to have people cooperate in choosing names for resources, and information architects often use participatory design techniques of card sorting or free listing for this purpose.<sup>23</sup>

2. While you may think that certain terms are more obviously “good” than others, studies show that “there is no one good access term for most objects. The idea of an ‘obvious,’ ‘self-evident,’ or ‘natural’ term is a myth!” ([Furnas et al. 1987, p. 967](#)).
3. ([Spencer 2009](#)). Free listing (see <http://boxesandarrows.com/beyond-cardsorting-free-listing-methods-to-explore-user-categorizations/>)

Back in the 1980s in the early days of computer user interface design, George Furnas and his colleagues at Bell Labs conducted a set of experiments to measure how much people would agree when they named some resource or function. The short answer: very little. Left to our own devices, we come up with a shockingly large number of names for a single common thing.

Unreliable Names: Knockin' On Heaven's Door

Figure 1: Top 25 Representations of "Knockin' On Heaven's Door" [35]



|                |   |                           |
|----------------|---|---------------------------|
| Guns N' Roses  | - | Knockin' On Heaven's Door |
| Guns N' Roses  | - | Knocking On Heavens Door  |
| Guns 'N' Roses | - | Knockin' On Heaven's Door |
| Guns N' Roses  | - | Knockin On Heavens Door   |
| Guns N' Roses  | - | Knockin' On Heavens Door  |
| Guns N'roses   | - | knockin on heavens door   |
| Guns N' Roses  | - | Knocking On Heaven's Door |
| Guns N Roses   | - | Knockin' On Heaven's Door |
| Guns N Roses   | - | Knockin On Heavens Door   |
| Guns And Roses | - | Knocking On Heavens Door  |
| Guns Nroses    | - | Knockin On Heavens Door   |
| Guns'n' Roses  | - | Knockin' On Heaven's Door |
| Guns N Roses   | - | Knocking On Heavens Door  |
| Guns'n'Roses   | - | Knockin' On Heaven's Door |
| Guns 'N Roses  | - | Knockin' On Heaven's Door |
| Guns & Roses   | - | Knockin' on Heaven's Door |
| Guns N'roses   | - | Knockin' On Heaven's Door |
| Guns and Roses | - | Knockin' On Heaven's Door |
| Guns'n Roses   | - | Knocking On Heavens Door  |
| Guns 'n' Roses | - | Knockin' On Heavens Door  |
| Aerosmith      | - | Knocking On Heaven's Door |
| Guns 'n Roses  | - | Knocking On Heaven's Door |
| Guns 'n' Roses | - | Knocking On Heavens Door  |
| Guns N Roses   | - | Knocking On Heaven's Door |
| Guns N' Roses  | - | Knockin On Heaven's Door  |

In 2008, Music recommendation service Last.fm employee Richard Jones compiled a list of the 100 most descriptions of the Guns N' Roses recording of Bob Dylan's song "Knockin' on Heaven's Door." The 21st most common description of the song incorrectly attributes the recording to Aerosmith.



Reprinted in Figure 1 of ([Hemerly 2011](#)). Used by permission here.

In one experiment, a thousand pairs of people were asked to “write the name you would give to a program that tells about interesting activities occurring in some major metropolitan area.” Less than 12 pairs of people agreed on a name. Furnas called this phenomenon [the vocabulary problem](#), concluding that no single word could ever be considered the “best” name.<sup>4</sup>

### *Homonymy, Polysemy, and False Cognates*

Sometimes the same word can refer to different resources—a “bank” can be a financial institution or the side of a river. When two words are spelled the same but have different meanings they are *homographs*; if they are also pronounced the same they are

4. The most common names for this service were activities, calendar and events, but in all over a hundred different names were suggested, including cityevents, whatup, sparetime, funtime, weekender, and nightout, “People use a surprisingly great variety of words to refer to the same thing,” Furnas wrote, “If everyone always agreed on what to call things, the user’s word would be the designer’s word would be the system’s word. ... Unfortunately, people often disagree on the words they use for things” ([Furnas et al. 1987, p. 964](#)).

*homonyms*. If the different meanings of the homographs are related, they are [polysemes](#).

Resources with homonymous and polysemous names are sometimes incorrectly identified, especially by an automated process that cannot use common sense or context to determine the correct referent. Polysemy can cause more trouble than simple homography because the overlapping meaning might obscure the misinterpretation. If one person thinks of a “shipping container” as being a cardboard box and orders some of them, while another person thinks of a “shipping container” as the large box carried by semi-trailers and stacked on cargo ships, their disagreement might not be discovered until the wrong kinds of containers arrive.<sup>5</sup>

Many words in different languages have common roots, and as a result are often spelled the same or nearly the same. This is especially true for technology words; for example, “computer” has been borrowed by many languages. The existence of these *cognates* and borrowed words makes us vulnerable to false cognates. When a word in one language has a different meaning and refers to different resources in another, the results can be embarrassing or disastrous. “Gift” is poison in German; “*pain*” is bread in French.

5. This example comes from [\(Farish 2002\)](#), who analyzes “What’s in a Name?” and suggests that multiple names for the same thing might be a good idea because non-technical business users, data analysts, and system implementers need to see things differently and no one standard for assigning names will work for all three audiences.

## *Names with Undesirable Associations*

False cognates are a special category of words that make poor names, and there are many stories relating product marketing mistakes, where a product name or description translates poorly, into other languages or cultures, with undesirable associations.<sup>6</sup> Furthermore, these undesirable associations differ across cultures. For example, even though floor numbers have the straightforward purpose to identify floors from lowest to highest levels, most buildings in Western cultures skip the 13<sup>th</sup> floor because many people think 13 is an unlucky number. In many East and Southeast Asian buildings, the 4<sup>th</sup> floor is skipped. In China the number 4 is dreaded because it sounds like the word for “death,” while 8 is prized because it sounds like the word for “wealth.”

While it can be tempting to dismiss unfamiliar biases and beliefs about names and identifiers as harmless superstitions and practices, their implications are ubiquitous and far from benign. [Alphabetical ordering](#) might seem like a fair and non-discriminatory arrangement of resources, but because it is easy to choose the name at the top of an alphabetical list, many firms in service businesses select names that begin with “A,” “AA,” or even “AAA” (look in any printed service directory). A consequence of this bias is that people or resources with names that begin with letters late in the alphabet are systematically discriminated against because they are often not considered, or because they are evaluated in the context created by resources earlier in the alphabet rather than on their own merit.<sup>7</sup>

6. See, for example, *Handbook of Cross-Cultural Marketing*, ([Kaynak 1997](#)).

7. See As easy as YZX, <http://www.economist.com/node/760345>. In addition, the convention to list the co-authors

## *Names that Assume Impermanent Attributes*

Many resources are given names based on attributes that can be problematic later if the attribute changes in value or interpretation.

From 'Kentucky Fried Chicken' to 'KFC'



of scientific publications in alphabetic order has been shown to affect reputation and employment by giving undeserved advantages to people whose names start with letters that come early in the alphabet. This bias might also affect admission to selective schools. ([Efthyvoulou 2008](#)).

“Kentucky Fried Chicken” was founded in 1930 by Harland Sanders as a tiny restaurant in a gas station storeroom in Corbin, Kentucky. It was one of the first fast-food chains to go international, and in 1987 was the first Western restaurant chain to open in China. It changed its name to “KFC” a few years later, no doubt in part because in Beijing, Moscow, London and other locations not anywhere near Kentucky many people have probably never heard of the place.

([Photo by Kyle Taylor. CC-BY-2.0 license.](#))

Web resources are often referred to using URLs that contain the domain name of the server on which the resource is located, followed by the directory path and file name on the computer running the server. This treats the current location of the resource as its name, so the name will change if the resource is moved. It also means that resources that are identical in content, like those at an archive or mirror website, will have different names than the original even though they are exact copies. An analogous problem is faced by restaurants or businesses with street names or numbers in their names if they lose their leases or want to expand.<sup>8</sup>

8. The Kentucky Fried Chicken franchise solved this problem by changing its name to KFC, which you can now find in Beijing, Moscow, London and other locations not anywhere near Kentucky and where many people have probably never heard of the place.

Some dynamic web resources that are generated by programs have URIs that contain information about the server technology used to create them. When the technology changes, the URIs will no longer work.<sup>9</sup>

Some resources have names that include page numbers, which disappear or change when the resource is accessed in a digital form. For example, the standard citation format for legal opinions uses the page number from the printed volume issued by West Publishing, which has a virtual monopoly on the publishing of court opinions and other types of legal documents.<sup>10</sup>

Why is the professional basketball team in Los Angeles called the “Lakers” when there are few natural lakes there? The team was originally located in Minneapolis, Minnesota, a state nicknamed “The Land of 10,000 Lakes.”

9. Tim Berners-Lee, the founder of the web, famously argued that [Cool URIs Don't Change \(Berners-Lee 1998\)](#).
10. Any online citation to one of the West printed court reports will use the West format. However, when Mead Data wanted to use the West page numbers in its LEXIS online service to link to specific pages, West sued for copyright infringement. The citation for the West Publishing vs. Mead Data Central case is 799 F.2d 1219 (8th Cir 1986), which means that the case begins on page 1219 of volume 799 in the set of opinions from the 8th Circuit Court of Appeals that West published in print form. West won the case and Mead Data had to pay substantial

Some resources have names that contain dates, years or other time indicators, most often to point to the future. The film studio named “20<sup>th</sup> Century Fox” took on that name in the 1930s to give it a progressive, forward-looking identity, but today a name with “20<sup>th</sup> Century” in it does the opposite.<sup>11</sup>

### *The Semantic Gap*

The *semantic gap* is the difference in perspective in naming and description when resources are described by automated processes rather than by people.<sup>12</sup>

The [semantic gap](#) is largest when computer programs or sensors obtain and name some information in a format optimized for efficient capture, storage, [decoding](#), or other technical criteria. The names—like `IMG20268.jpg` on a digital photo—might make sense for the camera as it stores consecutively taken photos but they are

royalties. Fortunately, this logic behind this decision was repudiated by the US Supreme Court a few years later in a case that West published as *Feist Publications, Inc., v. Rural Telephone Service Co.*, 499 U.S. 340 (1991), and West can no longer claim copyright on page numbers.

11. When George Orwell gave the title “[1984](#)” to a novel he wrote in 1949, he intended it as a warning about a totalitarian future as the Cold War took hold in a divided Europe, but today 1984 is decades in the past and the title does not have the same impact.
12. ([Dorai and Venkatesh 2002](#)).





a cousin, someone with whom they play basketball, and their professional mentor from work, how many places at the table will be set? Anywhere from two to five; it is possible all those relational descriptions refer to a single person, or to four different people, and because “friend,” “cousin,” “basketball teammate” and “mentor” do not name specific people you will have to guess who is coming to dinner.

If instead of descriptions you are told that the dinner guests are Bob, Carol, Ted, and Alice, you can count four names and you know how many people are having dinner. But you still cannot be sure exactly which four people are involved because there are many people with those names.

The uncertainty is eliminated if we use identifiers rather than names. [Identifiers](#) are names that refer unambiguously to a specific person, place, or resource because they are assigned in a controlled way. Identifiers are often strings of numbers or letters rather than words to avoid the biases and associations that words can convey. For example, a professor might grade exams that are identified by student numbers rather than names.

#### Names {and, or, vs} Identifiers

People change their names for many reasons: when they get married or divorced, because their name is often mispronounced or misspelled, to make a political or ethnic statement, or because they want to stand out. A few years a football player with a large ego named Chad Johnson, which is the second most common surname in the US, decided to change his name to his player number of 85, becoming Chad “Ochocinco.” He had an ochocinco.com website and used the ochocinco name on Facebook and Twitter. In a bit of irony, when

Ochocinco wanted to put Ocho Cinco on the back of his football jersey, the football league would not let him because his legal name does not have a space in it. That surely contributed to his decision to change his name back to Chad Johnson in 2012.

When you go to coffee shops, you are often asked your name, which the cashier writes on the empty cup so that your drink can be identified after the *barista* makes it. They do not actually need your name; just as some establishments use a receipt number to distinguish orders, what they need is an identifier. So even if your name is Joe, you can tell them it is Thor, Wotan, Mercurio, El Greco, Clark Kent, or any other name that is likely to be a unique identifier for the minute it takes to make your beverage.<sup>13</sup>

13. [\(Queenan 2011\)](#).

Most common US surnames;

[http://names.mongabay.com/most\\_common\\_surnames.htm](http://names.mongabay.com/most_common_surnames.htm).

Chad Ochocinco story: [http://en.wikipedia.org/wiki/Chad\\_Ochocinco](http://en.wikipedia.org/wiki/Chad_Ochocinco).

Fake names at Starbucks: <http://online.wsj.com/article/SB10001424053111904106704576582834147448392.html>.

Twitter on sports jerseys: <http://www.forbes.com/sites/alexknapp/2011/12/30/pro-lacrosse-team->

The distinction between [names](#) and [identifiers](#) for people is often not appreciated. (See the sidebar, [Names {and, or, vs} Identifiers](#).)

## *Make Names Informative*

The most basic principle of naming is to choose names that are informative, which makes them easier to understand and remember. It is easier to tell what a computer program or [XML](#) document is doing if it uses names like “ItemCost” and “TotalCost” rather than just “I” or “T.” People will enter more consistent and reusable address information if a form asks explicitly for “Street,” “City,” and “PostalCode” instead of “Line1” and “Line2.”

Identifiers can be designed with internal structure and semantics that conveys information beyond the basic aspect of pointing to a specific resource. An International Standard Book Number (ISBN) like “978-0-262-07261-8” identifies a resource (07261=“Document Engineering”) and also reveals that the resource is a book (978), in English (0), and published by The MIT Press (262).<sup>14</sup>

[replaces-names-with-twitter-handles-on-jerseys/?partner=technology\\_newsletter.](#)

14. Identifiers with meaningful internal structure are said to be structured or intelligent. Those that contain no additional information are sometimes said to be unstructured, opaque, or dumb. The 8 in the [ISBN](#) example is a check digit, not technically part of the identifier, that is algorithmically derived from the other digits to detect errors in entering the [ISBN](#).

The navigation points that mark intersections of radial signals from ground beacons or satellites that are crucial to aircraft pilots used to be meaningless five-letter codes that were changed to make them suggest their locations; semantic landmark names made pilots less likely to enter the wrong names into navigation systems. For example, some of the navigation points near Orlando, Florida—the home of Disney World—are MICKI, MINEE, and GOOFY.<sup>15</sup>

### *Use Controlled Vocabularies*

One way to encourage good names for a given resource domain or task is to establish a [controlled vocabulary](#). A controlled vocabulary is like a fixed or closed dictionary that includes the terms that can be used in a particular domain. A controlled vocabulary shrinks the number of words used, reducing synonymy and homonymy, eliminating undesirable associations, leaving behind a set of words with precisely defined meanings and rules governing their use.

A [controlled vocabulary](#) is not simply a set of allowed words; it also includes their definitions and often specifies rules by which the vocabulary terms can be used and combined. Different domains can create specific controlled vocabularies for their own purposes, but the important thing is that the vocabulary be used consistently throughout that domain.<sup>16</sup>

15. ([McCartney 2006](#)).
16. ([Svenonius 2000](#)) calls vocabulary control “the *sine qua non* of information organization” (p. 89). “The imposition of vocabulary control creates an artificial language out

For bibliographic resources important aspects of vocabulary control include determining the authoritative forms for author names, uniform titles of works, and the set of terms by which a particular subject will be known. In library science, the process of creating and maintaining these standard names and terms is known as *authority control*.

When evaluating what name to use for an author, librarians typically look for the name form that is used most commonly across that author's body of work while conforming to rules for handling prefixes, suffixes and other name parts that often cause name variations. For example, a name like that of Johann Wolfgang von Goethe might be alphabetized as both a "G" name and a "V" name, but using "G" is the authoritative way. "See" and "see also" references then map the variations to the authoritative name.<sup>17</sup>

Official authority files are maintained for many resource domains: a gazetteer associates names and locations and tells us whether we should be referring to Bombay or Mumbai; the Domain Name System(DNS) maps human-oriented domain and host names to

of a natural language" (p. 89), leaving behind an official, normalized set of terms and their uses.

17. This mapping is "the means by which the language of the user and that of a retrieval system are brought into sync" ([Svenonius 2000, p. 93](#)) and allows an information-seeker to understand the relationship between, say, Samuel Clemens and Mark Twain. The Library of Congress(LOC) maintains a list of standard, accepted names for authors, subjects, and titles called the Name Authority File. <http://id.loc.gov/authorities/names.html>.

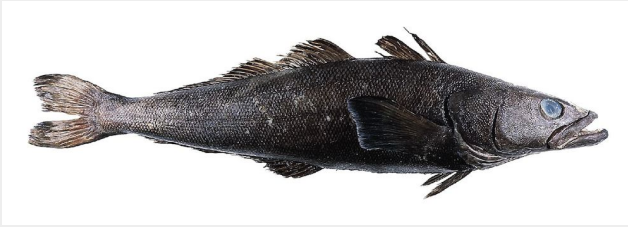
their IP addresses; the Chemical Abstracts Service Registry assigns unique identifiers to every chemical described in the open scientific literature; numerous institutions assign unique identifiers to different categories of animal species.<sup>18</sup>

In some cases, authority files are created or maintained by a community, as in the case of MusicBrainz, an “open music encyclopedia” to which users contribute information about artists, releases, tracks, and other aspects of music. Music metadata is notoriously unreliable; one study found over 100 variations in the description of the Knockin’ on Heaven’s Door song (written by Bob Dylan) as recorded by Guns N’ Roses.<sup>19</sup>

### *Allow Aliasing*

Aliasing: Bad for this Fish

18. Pan-European Species Directory Infrastructure (PESI): <http://www.eu-nomen.eu/pesi>; Consortium for the Barcode of Life (CBOL): <http://www.barcoding.si.edu/>; NatureServe: <http://services.natureserve.org/BrowseServices/getSpeciesData/getSpeciesListREST.jsp>.
19. [\(Hemerly 2011\)](#).



A fish once known as the Patagonian Toothfish because of its large and unattractive teeth became popular in American restaurants when a fish wholesaler began marketing it as the Chilean Sea Bass even though it is usually found farther south in cold Antarctic waters and it is not a sea bass. Unfortunately for the fish, this alias was so successful that it led to overfishing, threatening the survival of the species. Some environmentally-oriented chefs, restaurateurs, and seafood distributors organized a boycott to save the fish. ([Fabricant 2002](#))

(Photo published by the United States Government.  
Not protectable by copyright (17 USC Sec. 105).)

A [controlled vocabulary](#) is extremely useful to people who use it, but if you are designing an organizing system for other people who do not or cannot use it, you need to accommodate the variety of words they will actually use when they seek or describe resources. The authoritative name of a certain fish species is *Amphiprion ocellaris*, but most people would search for it as “clownfish,” “anemone fish,” or even by its more familiar film name of Nemo.

Furnas suggests “unlimited aliasing” to connect the uncontrolled or natural vocabularies that people use with the controlled one

employed by the organizing system. By this he means that there must be many alternate access routes to each word or function that a user is trying to find. For example, the birth name of the 42nd President of the United States of America is “William Jefferson Clinton,” but web pages that refer to him as “Bill Clinton” are vastly more common, and searches for the former are redirected to the latter. A related mechanism used by search engines is spelling correction, essentially treating all the incorrect spellings as aliases of the correct one (“did you mean California?” when you typed “Claifornia”).

### *Make Identifiers Unique or Qualified*

Even though an identifier refers to a single resource, this does not mean that no two identifiers are identical. One military inventory system might use stock number 99 000 1111 to identify a 24-hour, cold-climate ration pack, while another inventory system could use the same number to identify an electronic radio valve. Each identifier is unique in its inventory system, but if a supply request gets sent to the wrong warehouse hungry soldiers could be sent radio valves instead of rations.<sup>2021</sup>

20. This rations / radio confusion is described in ([Wheatley 2004](#)). In 2008 a similar mistake in managing inventory at a US military warehouse led to missile launch fuses being sent to Taiwan instead of helicopter batteries, causing a high-level diplomatic furor when the Chinese government objected to this as a treaty violation ([Hoffman 2008](#)).

21. Organizing systems in libraries, museums, and



We can prevent or reduce identifier collisions by adding information about the *namespace*, the domain from which the names or identifiers are selected, thus creating what are often called [qualified names](#). There are several dozen US cities named “Springfield” and “Washington,” but adding state codes to mail addresses distinguishes them. Likewise, we can add prefixes to XML element names when we create documents that reuse components from multiple document types, distinguishing `<book:Title>` from `<legal:Title>`.

We can fix problems like these by qualifying or extending the identifier, or by creating a [globally unique identifier \(GUID\)](#), one that will never be the same as another identifier in any organizing system anywhere else. One easy method to create a GUID is to use a URL you control and append a string to it, the same approach that gives every web page a unique address. GUIDs are often used to identify software objects, the resources in distributed systems, or data collections.<sup>22</sup>

Because they are not created by an algorithm whose results are

businesses often give sequential accession numbers to resources when they are added to a collection, but these identifiers are of no use outside of the context in which they are assigned, as when a union catalog or merged database is created.

22. A more general technique is to use the *Universally Unique Identifier (UUID)* standard, which standardizes some algorithms that generate 128-bit tokens that, for all practical purposes, will be unique for hundreds, if not thousands, of years.

provably unique, we do not consider fingerprints, or other biometric information, to be globally unique identifiers for people, but for all practical purposes they are.<sup>23</sup>

### *Distinguish Identifying and Resolving*

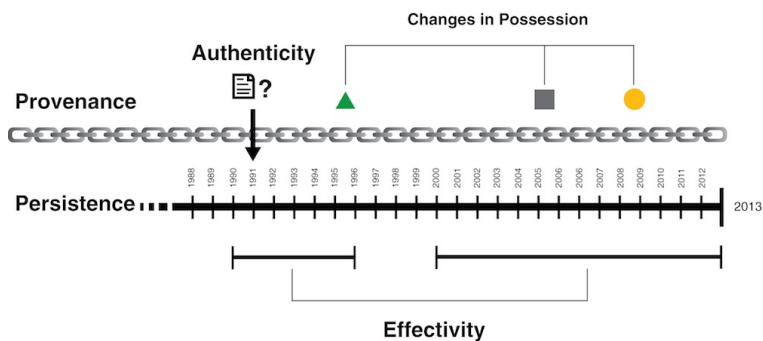
Library call numbers are identifiers that do not contain any information about where the resource can be found in the library stacks or in a digital repository. This separation enables this identification system to work when there are multiple copies in different locations, in contrast to URIs that serve as both identifiers and locations much of the time. When the identifier does not contain information about resource location, it must be “resolved” to determine the location. With physical resources, [resolution](#) takes place with the aid of signs, maps, or other associated resources that describe the resource arrangement in some physical environment; for example, “You are here” maps associate each resource identifier with a coordinate or other means of finding it on the map. With digital resources, the resolver is a directory system or service that interprets an identifier and looks up its location or directly initiates resource retrieval.

23. ([OASIS 2003](#)). The Organization for the Advancement of Structured Information Systems(OASIS) XML Common Biometric Format(XCBF) was developed to standardize the use of biometric data like DNA, fingerprints, iris scans, and hand geometry to verify identity ([https://www.oasis-open.org/committees/tc\\_home.php?wg\\_abbrev=xcbf](https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xcbf)).

## 29. Resources over Time

Problems of “what is the resource?” and “how do we identify it?” are complex and often require ongoing work to ensure they are properly answered as an organizing system evolves. We might need to know how a resource does or does not change over time (its [persistence](#)), whether its state and content come into play at a specified point in time (its [effectivity](#)), whether the resource is what it is said to be (its [authenticity](#)), and sometimes who has certified its authenticity over time (its [provenance](#)). A resource might have persistence, but only the provenance provided by an documented chain of custody enables questions about authenticity to be answered with authority. Effectivity describes the limits of a resource’s lifespan on the time line.

### Resources over Time



Four considerations that arise with respect to the maintenance of resources over time are their persistence, provenance, authenticity, and effectivity.

[Figure: Resources over Time](#) portrays the relationships among the concepts of Persistence, Provenance, Effectivity, and Authenticity.

## Persistence

The Great Sphinx at Giza



The Great Sphinx has persisted for over four thousand years. It has survived acts of vandalism, target practice by Napoleon's artillery, shoulder-deep burial in desert sandstorms, and eventual excavation in the early 20th century. (Its origin and that of the name "Sphinx" are debated among scholars, yet both are recognized worldwide as a persistent meme in our collective consciousness.)

(Photo by R. Glushko.)

Even if you have reached an agreement as to the meaning of "a thing" in your organizing system, you still face the question of the identity of the resource over time, or its *persistence*.

## *Persistent Identifiers*

How long must an identifier last? Coyle gives the conventional, if unsatisfying, answer: “As long as it is needed.”<sup>1</sup> In some cases, the time frame is relatively short. When you order a specialty coffee and the *barista* asks for your name, this identifier only needs to last until you pick up your order at the end of the counter. But other time frames are much longer. For libraries and repositories of scientific, economic, census, or other data the time frame might be “forever.”

The design of a scheme for persistent identifiers must consider both the required time frame and the number of resources to be identified. When the Internet Protocol(IP) was designed in 1980, it contained a 32-bit address scheme, sufficient for over 4 billion unique addresses. But the enormous growth of the Internet and the application of IP addresses to resources of unexpected types have required a new addressing scheme with 128 bits.<sup>2</sup>

Recognition that URIs are often not persistent as identifiers for web-based resources led the [Association of American Publishers\(AAP\)](#) to develop the *Digital Object Identifier(DOI)* system. The location and owner of a digital resource can change, but its DOI is permanent.<sup>3</sup>

1. ([Coyle 2006, p. 429](#)).
2. IP v6 for Internet addresses. The threat of exhaustion was the motivation for remedial technologies, such as classful networks, *Classless Inter-Domain Routing(CIDR)* methods, and *Network Address Translation(NAT)* that extend the usable address space.
3. Digital                      Object                      Identifier(DOI)                      system

## *Persistent Resources*

Even though persistence often has a technology dimension, it is more important to view it as a commitment by an institution or organization to perform activities over time to ensure that a resource is available when it is needed. Put another way, preservation (“[Preservation](#)”) and governance (“[Governance](#)”) are activities carried out to ensure the outcome of persistence.

### Stop and Think: The Paradox of Theseus

Every day that Theseus’s ship is in the harbor, a single plank gets replaced, until after a few years the ship is completely rebuilt: not a single original plank remains. Is it still the ship of Theseus? And suppose, meanwhile, the shipbuilders have been building a new ship out of the replaced planks? Is that the ship of Theseus?

The subtle relationship between preservation and persistence raises some interesting questions about what it means for a resource to stay the same over time. One way to think of persistence is that a persistent resource is never changed. However, physical resources often require maintenance, repair, or restoration to keep them accessible and usable, and we might question whether at some point these activities have transformed them into different

(<http://www.doi.org>). However, DOI has its issues too. It is a highly political, publisher-controlled system, not a universal solution to persistence.

resources.<sup>4</sup> Likewise, digital resources require regular backup and migration to keep them available and this might include changing their digital format.

Many resources like online newspapers or blog feeds continually change their content but still have persistent identifiers. This suggests we should think of persistence more abstractly, and consider as persistent resources any that remain functionally the same to support the same interactions at any point in their lifetimes, even if their physical properties or information values change.

Active resources implemented as computational agents or web services might be re-implemented numerous times, but as long as they do not change their interfaces they can be deemed to be persistent from the perspective of other resources that use them. Similarly, the dataset that defines a user or customer model in a recommendation system should be treated as a persistent resource; it includes information like name and date of birth that is persistent in the traditional sense; but it might also include “last purchase” and “current location,” which must change frequently to maintain the accuracy and usefulness of the customer model.

Some organizing systems closely monitor their resources and every interaction with them to prevent or detect tampering or other

4. This is called the Paradox of Theseus, a philosophical debate since ancient times. Every day that Theseus's ship is in the harbor, a single plank gets replaced, until after a few years the ship is completely rebuilt: not a single original plank remains. Is it still the ship of Theseus? And suppose, meanwhile, the shipbuilders have been building a new ship out of the replaced planks? Is that the ship of Theseus? ([Furner 2008, p. 6](#)).

unauthorized changes. Some organizing systems, like those for software or legal documents, explicitly maintain every changed version to satisfy expectations of persistence, because different users might not be relying on the same version. With digital resources, determining whether two resources are the same or determining how they are related or derived from one another are very challenging problems.<sup>5</sup>

## Effectivity

### In Which Country Do You Live?

Even if you always live in the same place, the answer to “what country do you live in?” can depend on when it is asked. Consider the case of an elderly woman born in 1929 in Zemun, a district in the eastern European city of Belgrade, who has never moved. The place she lives has been part of seven different countries during her lifetime: Kingdom of Yugoslavia (1929-1941); Independent State of Croatia (1941-1945); Federal People’s Republic of Yugoslavia (1945-1963); Socialist Federal Republic of Yugoslavia (1963-1992); Federal Republic of Yugoslavia (1992-2003); State Union of Serbia and Montenegro (2003-2006); Republic of Serbia (2007–present).

Many resources, or their properties, also have locative or temporal *effectivity*, meaning that they come into effect at a particular time

5. See [\(Renear and Dubin 2003\)](#), [\(Wynholds 2011\)](#).



and/or place; will almost certainly cease to be effective at some future date, and may cease to be effective in different places.

Temporal effectivity, sometimes known as “time-to-live,” is generally expressed as a range of two dates. It consists of a date on which the resource is effective, and optionally a date on which the resource ceases to be effective, or becomes stale. For some types of resources, the *effective date* is the moment they are created, but for others, the effective date can be a time different from the moment of creation. For example, a law passed in November may take effect on January 1 of the following year, and credit cards first need to be activated and then can no longer be used after their expiration date. An “effective date” is the counterpart of the “Best Before” date on perishable goods. That date indicates when a product goes bad, whereas an item’s effective date is when it “goes good” and the resource that it supersedes needs to be disposed of or archived.

Locative effectivity considers borders, security, roadways, altitude, depth and other geographic factors. Some types of resources, including people, are restricted as to where they may or may not be transported and/or used, such as hazardous cargo, explosives, narcotics, pharmaceuticals, alcohol, and cannabinoids. Jurisdictional issues concern borders, transportation corridors, weather stations, and geographic surveys. Parachutes are altitude-sensitive and scuba diving cylinders are depth-sensitive.

Effectivity concerns sometimes intersect with authority control for names and places. Name changes for resources often are tied to particular dates, events, and locations. Laws and regulations differ across organizational and geopolitical boundaries, and those boundaries often change. Some places that have been the site of

civil unrest, foreign occupation, and other political disruptions have had many different names over time, and even at the same time.<sup>6</sup>

Today these disputed borders cause a problem for Google Maps when it displays certain international borders. Because Google is subject to the laws of the country where its servers are located, it must present disputed borders to conform with the point of view of the host country when a country-specific Google site is used to access the map.<sup>7</sup>

6. See [http://www.nationsonline.org/oneworld/hist\\_country\\_names.htm](http://www.nationsonline.org/oneworld/hist_country_names.htm) for a list of formerly used country names and their respective effectivity.
7. See (Gravois 2010). One specific example of this effect of international geopolitics on an organizing system involves the northern border of the Crimean Peninsula. When running a query for “Ukraine” via [google.com/maps](http://google.com/maps) (USA), the border appears as a dotted line, which reflects a “neutral” perspective in the aftermath of recent political and military conflicts. Alternatively, when submitting the same query via [google.com.ua/maps](http://google.com.ua/maps) (Ukraine), there is no border at all, which is a reflection of a Ukrainian perspective that the Crimean Peninsula is part of Ukraine. Lastly, when the query is submitted via [google.ru/maps](http://google.ru/maps) (Russia), the border is represented as a solid line, which reflects a Russian perspective that the territory is part of Russia. A 2014 study of Google Maps found 32 situations where the answer to “what country

In most cases effectivity implies persistence requirements because it is important to be able to determine and reconstruct the configuration of resources that was in effect at some prior time. A new tax might go into effect on January 1, but if the government audits your tax returns what matters is whether you followed the law that was in effect when you filed your returns.<sup>8</sup>

## Authenticity

Do You Trust This?

is that on the map?” depended on where it was asked ([Yanovsky 2014](#))

8. Effectivity in the tax code is simple compared to that relating to documents in complex systems, like commercial aircraft. Because of their long lifetimes—the Boeing 737 has been flying since the 1960s—and continual upgrading of parts like engines and computers, each airplane has its own operating and maintenance manual that reflects changes made to the plane over time. Every change to the plane requires an update to the repair manual, making the old version obsolete. And while an aircraft mechanic might refer to “the 737 maintenance manual,” each 737 aircraft actually has its own unique manual.



Ustar.net sells photos autographed by celebrities, and each comes with a Certificate of Authenticity that includes a replica of the photo and a signature from a Ustar employee to guarantee that the autograph is an actual hand-signed one. But Ustar does not provide a certificate to guarantee that the employee signature is an authorized one.

(Screenshot by R. Glushko. Source: [ustar.net](http://ustar.net).)

In ordinary use we say that something is *authentic* if it can be shown to be, or has come to be accepted as what it claims to be. The importance and nuance of questions about authenticity can be seen in the many words we have to describe the relationship between “the real thing” (the “original”) and something else: copy,

reproduction, replica, fake, phony, forgery, counterfeit, pretender, imposter, ringer, and so on.

It is easy to think of examples where authenticity of a resource matters: a signed legal contract, a work of art, a historical artifact, even a person's signature.

The creator or operator of an organizing system, whether human or machine, can authenticate a newly created resource. A third party can also serve as proof of authenticity. Many professional careers are based on figuring out if a resource is authentic.<sup>9</sup>

There is a large body of techniques for establishing the identity of a person or physical resource. We often use judgments about the physical integrity of recorded information when we consider the integrity of its contents.

Digital authenticity is more difficult to establish. Digital resources can be reproduced at almost no cost, exist in multiple locations, carry different names on identical documents or identical names on different documents, and bring about other complications that do not arise with physical items. Technological solutions for ensuring digital authenticity include time stamps, watermarking, encryption, and digital signatures. However, while scholars generally trust technological methods, technologists are more skeptical of them because they can imagine ways for them to be circumvented or counterfeited. Even when a technologically sophisticated system for establishing authenticity is in place, we can still only assume the

9. A notary public is used to verify that a signature on an important document, such as a mortgage or other contract, is authentic, much as signet rings and sealing wax once proved that no one has tampered with a document since it was sealed.

constancy of identity as far back as this system reaches in the “chain of custody” of the document.

## Provenance

In [“Looking ‘Upstream’ and ‘Downstream’ to Select Resources”](#) we recommended that you analyze any evidence or records about the use of resources as they made their way to you from their headwaters to ensure they have maintained their quality over time. The concept of provenance transforms the passive question of “what has happened to this resource?” into actions that can be taken to ensure that nothing bad can happen to a resource or to enable it to be detected.

### Chinese Manuscript With Provenance Seals



This beautiful manuscript, preserved in the National Palace Museum in Taipei, was created by Zhao Ji (赵佶), Emperor Huizong, the 8th Emperor of the Chinese Song

Dynasty about a thousand years ago. He was famous for his skills in poetry, painting, and calligraphy. There are two poems here; the one on the right describes the techniques for Chinese landscape paintings, while the left one expresses the Emperor's appreciation of plum blossoms, which signal the onset of spring.

The red seals are those of several Ching Dynasty emperors over many generations, with the oldest being at least five hundred years after Huizong created the poems. Stamping your personalized red seal on a resource is analogous to but vastly more elegant and informative than "Liking" a web page today.

(Photo by R. Glushko.)

The idea that important documents must be created in a manner that can be authenticated and then preserved, with an unbroken chain of custody, goes back to ancient Rome. Notaries witnessed the creation of important documents, which were then protected to maintain their integrity or value as evidence. In organizing systems like museums and archives that preserve rare or culturally important objects or documents this concern is expressed as the principle of *provenance*. This is the history of the ownership of a collection or the resources in it, where they have been and who has had access to the resources.

A uniquely Chinese technique in organizing systems is the imprinting of elaborate red seals on documents, books, and paintings that collectively record the provenance of ownership and the review and approval of the artifact by emperors or important officials.

However, it is not only art historians and custodians of critical documents that need to be concerned with provenance. If you are planning to buy a used car, it is wise to check the vehicle history (using the Vehicle Identification Number, the car's persistent identifier) to make sure it hasn't been wrecked, flooded, or stolen.



## 30. Key Points in Chapter Four

- What is the relationship between a resource and a category?

We can consider a resource to be one of many members of a very broad category, as the unique instance of a category with only one member, or anywhere in between.

(See [“What Is a Resource?”](#))

- What factors affect the size of a category?

The size of the category—the number of resources that are treated as equivalent—is determined by the properties or characteristics we consider when we examine the resource.

(See [“What Is a Resource?”](#))

- What is metadata?

Organizing systems for physical information resources emphasize description resources or surrogates like bibliographic records that describe the information content rather than their physical properties.

(See [“Bibliographic Resources, Information Components, and “Smart Things” as Resources”](#))

- What is an identifier and what design goals must it satisfy?

An identifier is a special kind of name assigned in a controlled manner and governed by rules that define possible values and naming conventions. The design of a scheme for persistent

identifiers must consider both the required time frame and the number of resources to be identified.

(See [“Identity, Identifiers, and Names”](#))

- What are active resources?

[Active resources](#) create effects or value on their own, sometimes when they initiate interactions with passive resources. Active resources can be people, other living resources, computational agents, active information sources, web-based services, self-driving cars, robots, appliances, machines or otherwise ordinary objects like light bulbs, umbrellas, and shoes that have been made “smarter.”

(See [“Active or Operant Resources”](#))

- What is the recall/precision tradeoff?

More fine-grained organization reduces [recall](#), the number of resources you find or retrieve in response to a query, but increases the [precision](#) of the recalled set, the proportion of recalled items that are relevant.

(See [“Identity and Information Components”](#))

- What is agency?

Agency is the extent to which a resource can initiate actions on its own. We can define a continuum between completely passive resources that cannot initiate any actions and active resources that can initiate actions based on information they sense from their environments or obtain through interactions with other resources.

(See [“Resource Agency”](#))

- What are active resources?

Resources become active resources when they contain sensing and communication capabilities.

(See [“Resource Agency”](#))

- Is there a fundamental difference between a primary resource and metadata associated with it?

Which resources are primary and which are metadata is often just a decision about which resource is the [focus](#) of our attention.

(See [“Resource Focus”](#))

- What is the Document Type Spectrum?

It can be useful to view domains of information resources on the Document Type Spectrum from weakly-structured narrative content to highly structured transactional content.

(See the sidebar, [The Document Type Spectrum](#))

- What is the FRBR four-level abstraction hierarchy?

The concept of identity for bibliographic resources has evolved into a four-level abstraction hierarchy between the abstract [work](#), an [expression](#) in multiple formats or genres, a particular [manifestation](#) in one of those formats or genres, and a specific physical [item](#).

(See [“Identity and Bibliographic Resources”](#) and [Figure: The FRBR Abstraction Hierarchy](#).)

- What is the Internet of Things?

If the resource has an IP address, it is part of the “Internet of Things.”

(See [“Identity and Active Resources”](#).)

- What is the vocabulary problem?

Every natural language offers more than one way to express any thought, and in particular there are usually many words that can be used to refer to the same thing or concept.

(See [“The Problems of Naming”](#))

- What is a potential problem with basing names on a resource attribute?

Many resources are given names based on attributes that can be problematic later if the attribute changes in value or interpretation.

(See [“Names that Assume Impermanent Attributes”](#))

- What is the Semantic Gap?

The *semantic gap* is the difference in perspective in naming and description when resources are described by automated processes rather than by people.

(See [“The Semantic Gap”](#))

- What is the most basic principle of naming?

The most basic principle of naming is to choose names that are informative.

(See [“Make Names Informative”](#))

- What is a controlled vocabulary?

One way to encourage good names for a given resource domain or task is to establish a [controlled vocabulary](#). A *controlled vocabulary* is like a fixed or closed dictionary that includes the terms that can be used in a particular domain. A controlled vocabulary shrinks the number of words used, reducing synonymy and homonymy, eliminating undesirable

associations, leaving behind a set of words with precisely defined meanings and rules governing their use.

(See [“Use Controlled Vocabularies”](#))

- What is authority control?

For bibliographic resources important aspects of vocabulary control include determining the authoritative forms for author names, uniform titles of works, and the set of terms by which a particular subject will be known. In library science, the process of creating and maintaining these standard names and terms is known as *authority control*.

(See [“Use Controlled Vocabularies”](#))

- Which organizing system activities promote persistence?

Preservation and governance are activities carried out to ensure that resources will last as long as they are needed.

(See [“Persistence”](#))

- What is effectivity?

Many resources, or their properties, also have locative or temporal *effectivity*, meaning that they come into effect at a particular time and/or place; will almost certainly cease to be effective at some future date, and may cease to be effective in different places.

(See [“Effectivity”](#))

- What guarantees the authenticity of a resource?

The only guarantee of a resource’s authenticity is having total oversight over the “chain of custody” from its creation to the present.

(See [“Authenticity”](#) and [“Provenance”](#))

PART V

# RESOURCE DESCRIPTION AND METADATA

Robert J. Glushko

Kimra McPherson

Ryan Greenberg

Robyn Perry

Matthew Mayernik

Graham Freeman

Carl Lagoze



## 3I. Introduction (V)

This chapter is a turning point in the book. The earlier chapters have discussed the key ideas of the discipline of organizing: identifying and selecting the resources to organize, and then organizing and maintaining them and their organizing system. We have emphasized that finding things later is the most important reason for organizing them. This can be surprisingly hard to do. People know things by different names or remember different aspects of them.

“Arrangement in Grey and Black No. 1”



“Arrangement in Grey and Black No. 1” (1871). Oil on canvas, by James Abbott McNeill Whistler. Alternative



titles: “Portrait of the Artist’s Mother” and “Whistler’s Mother.” The painting is in Musée d’Orsay in Paris.

(Photo by Jean-Gilles Berizzi Source: [Wikimedia Commons](#).)

The famous painting here by the 19th century American painter James Whistler is exhibited in the Musée d’Orsay in Paris, and has been described as a Victorian-era [Mona Lisa](#). What name do you know it by? How should it be described?

Resource descriptions for art usually contain the name of the artist, the medium, the year of its completion, and, of course, its title. Most of these map fairly obviously to the properties they describe; the title, owing to its prominence and expressive power, is often an exception.

Most often, a painting’s title describes its subject. If you recognize the previous painting, you most likely know it by its colloquial name, Whistler’s Mother. While it is a portrait of Anna McNeill Whistler, mother of painter James Abbott McNeill Whistler, the artist gave it a radically different title, Arrangement in Grey and Black, No. 1, because he believed the most important property of a painting was not the subject it depicted, but its purely aesthetic properties and their effect on the viewer. So Whistler named his paintings, which were mostly landscapes and portraits, in the manner of musical compositions: Nocturne in Black and Gold; Symphony in White; Arrangement in Pink, Red, and Purple; and so on.

If Whistler’s title surprises you, because you would have described it as a portrait of an elderly woman, this helps reinforce how wildly different names of the same resource can be. Resource descriptions and metadata provide meaning, but to whom? What is salient about

a resource can depend on the context in which it is experienced, and thus may change over time. Descriptions that make sense to some people might not make sense to others. People searching on the “wrong descriptions” or the “wrong metadata” will not find what they are looking for.

#### Mt. St. Helens Before and After



Before 1980, Mt. St. Helens was a “postcard-like” snow-covered mountain. Afterward, the mile-wide crater where its mountaintop once was reminds us of its violent volcanic eruption.

(Credit: Public domain images from US Forest Service and USGS.)

Mt. St. Helens, in the southwest corner of Washington State, was usually just described as a mountain until 1980. Then, the deadliest and most economically destructive volcanic event in the history of the United States blew away the top of the mountain, killing 57 people, and leaving a mile-wide crater. Today almost every description of Mt. St. Helens mentions the volcanic eruption.

It would seem impossible to search using the wrong description if the descriptions of a resource were kept current to include all the latest information, but search engines are already too powerful,

usually producing too much information. Technology improvements in search and retrieval do not eliminate the cognitive effort to remember what things are, how they are best described, and where they might be found. The design of resource descriptions and metadata depends on why we need to find the information later. This chapter is about how and why.

Stop and Think:

These Places Have Their Moments

Our description of Mt. St. Helens forever changed after its volcanic eruption. Surely there are times and places that you remember differently because of their part in an important event. A family wedding? The Olympic Games? A natural disaster? The Twin Towers?

It is easy to find before and after images of Mt. St. Helens doing a web search. What information might be associated with these images? Modern cameras assign an identifier to the stored photograph and they also capture the technical description of the image's production: the type of camera, lens, shutter speed, light sensitivity, aperture, and other settings.<sup>1</sup> Many modern cameras also record information about the geographic and temporal circumstances surrounding the image's creation: the date, time and location on Earth where the photograph is taken. When the image is transferred out of the camera and is published for all to see,

1. Most digital cameras use the Exchangeable Image File Format(EXIF). The best source of information about it looks like its Wikipedia entry. [http://en.wikipedia.org/wiki/Exchangeable\\_image\\_file\\_format](http://en.wikipedia.org/wiki/Exchangeable_image_file_format).

it might be useful to record biographical information about the photographer to help viewers relate to the photographer and better understand the photograph's context. There may also be different licenses and copyright information to associate with the picture—who owns it and how it can be used.

Consider a completely different context. Four 7-year old boys are selecting Lego blocks to complete their latest construction. The first boy is looking for “cylinder one-ers,” another for “coke bottles,” the third for “golder wipers,” and the final boy is looking for “round one-bricks”? It turns out, they are all the same thing; each boy has devised his own set of descriptive terms for the tiny building blocks. Some of their many descriptions are based on color alone (“redder”), some on color and shape (“blue tunnel”), some on role (“connector”), some on common cultural touchstones (“light saber”). Others, like “jail snail” and “slug,” seem unidentifiable—unless, of course, you happen to be inside the mind of a particular 7-year-old kid. It doesn't matter if the boys use different description vocabularies when they play by themselves, but they will have to agree if they play together.<sup>2</sup>

2. This is much more than just a “kids say the darnedest things” story (see [http://en.wikipedia.org/wiki/Kids\\_Say\\_the\\_Darndest\\_Things](http://en.wikipedia.org/wiki/Kids_Say_the_Darndest_Things)). Giles Turnbull (Turnbull 2009) noticed that his kids never used the official names for Lego blocks (e.g., Brick 2x2). He then asked other kids what their names were for 32 types of Lego blocks. His survey showed that the kids mostly used different names, but each created names that followed some systematic principles. The most standard name was the “light saber,” used by every kid in Turnbull's sample.

Paintings, digital photos, and Lego blocks are all very different, but together these scenarios raise important questions about describing resources that we attempt to answer in this chapter:

- What is the purpose of resource description?
- What resource properties should be described?
- How are resource descriptions created?
- What makes a good resource description?

## Navigating This Chapter

We begin with an overview of [\*resource description\*](#) (“[\*An Overview of Resource Description\*](#)”), which we propose as a broad concept that includes the narrower concepts of [\*bibliographic descriptions\*](#) and [\*metadata\*](#). “[\*The Process of Describing Resources\*](#)” describes a 7-step process of describing resources that includes determining scope, focus and purposes, identifying resource properties, designing the description

vocabulary, designing the description form and implementation, and creating and evaluating the descriptions. Because many principles and methods for resource description were developed for describing text resources in physical formats, in [“Describing Non-text Resources”](#) we briefly discuss the issues that arise when describing museum and artistic resources, images, music, video, and contextual resources.



## 32. An Overview of Resource Description

We describe resources so that we can refer to them, distinguish among them, search for them, manage access to them, preserve them, and make predictions about what might happen to them or what they might do. Each purpose may require different [resource descriptions](#). We use [resource descriptions](#) in every communication and conversation; they are the enablers of organizing systems.

### Naming {and, or, vs.} Describing

[Resources in Organizing Systems](#) discussed how to decide what things should be treated as resources and how names and identifiers distinguish one resource from another. Names can suggest the properties and principles an organizing system uses to arrange its resources. We can see how societies organize their people by noting that among the most common surnames in English are descriptions of occupations (Smith, Miller, Taylor), descriptions of kinship relations (Johnson, Wilson, Anderson), and descriptions of appearance (Brown, White).<sup>1</sup>

1. [\(Reaney and Wilson 1997\)](#) classify surnames as local, surnames of relationship, surnames of occupation or office, and nicknames. The dominance of occupational names reflects the fact that there are fewer occupations than places. While there are only a handful of kinship



In many cultures, one spouse or the other takes a name that describes their marital relationship. In many parts of the English-speaking world, married women have often referred to themselves using their husband's name.<sup>2</sup>

Similarly, many other kinds of resources have names that are property descriptions, including buildings (Pentagon, White House), geographical locations (North America, Red Sea), and cities (Grand Forks, Baton Rouge).

Every resource can be given a name or identifier. Identifiers are especially efficient [resource descriptions](#) because, by definition, identifiers are unique over some domain or collection of resources. Names and identifiers do not typically describe the resource in any ordinary sense because they are usually assigned to the resource rather than recording a property of it.

However, the arbitrariness of names and identifiers means that they do not serve to distinguish resources for people who do not already know them. This is why we use what linguists call referring expressions or definite descriptions, like “the small black dog”

relationships used in surnames (patronymic or father-based names are most common), because the surname includes the father's name there is more variation than for occupations.

2. This odd convention is preserved today in wedding invitations, causing some feminist teeth gnashing ([Geller 1999](#)).

rather than the more efficient “Blackie,” when we are talking to someone who does not know that is the dog’s name.<sup>3</sup>

Similarly, when we use a library catalog or search engine to locate a known resource, we query for it using its name, or some specific information we know about it, to make it easier to find. In contrast, when we look for resources to satisfy an information need but do not have specific resources in mind, we query for them using descriptions of their content or other properties. In general, information retrieval can be characterized as comparing the description of a user’s needs with descriptions of the resources that might satisfy them.

## “Description” as an Inclusive Term

Up to now we have used the concept of “description” in its ordinary sense to mean the labeling or explaining of the visible or important features that characterize or represent something. However, the concept is sometimes used more precisely in the context of organizing systems, where [resource description](#) is often more formal, systematic, and institutional. In the library science context of [bibliographic description](#), a *descriptor* is one of the terms in a carefully designed language that can be assigned to a resource to designate its properties, characteristics, or meaning, or its relationships with other resources. In the contexts of conceptual modeling and information systems design, the terms in [resource descriptions](#) are also called “keywords,” “index terms,” [attributes](#),

3. See ([Donnellan 1966](#)). A contemporary analysis from the perspective of cognitive science is ([Heller, Gorman, and Tanenhaus 2012](#)).

[attribute values](#), [elements](#), “data elements,” “data values,” or “the vocabulary.” In business intelligence, predictive analytics or other data science contexts these are called “variables,” “features,” [properties](#), or “measurements.” In contexts where descriptions are less formal or more personal the description terms are often called “labels” or “tags.” Rather than attempt to make fine distinctions among these synonyms or near-synonyms, we will use “description” as an inclusive term except where conventional usage overwhelmingly favors one of the other terms.

Many of these terms come from a narrow semantic scope in which the purpose of description is to identify and characterize the essence, or [aboutness](#), of a resource. However, as it becomes trivial to associate computationally generated information with resources, many additional kinds of information beyond strict “aboutness” can support additional interactions. We describe many of these purposes and the types of information needed to enable them in [“Determining the Purposes”](#). We apply [resource description](#) in an expansive way to accommodate all of them.

[Resources in Organizing Systems](#) introduced the distinction of [“Resource Focus”](#) to contrast primary resources with resources that describe them, which we called Description Resources. We chose this term as a more inclusive and more easily understood alternative to two terms that are well established in organizing systems for information resources: [bibliographic descriptions](#) and [metadata](#). We will also distinguish resource description as a general concept from the narrower senses of statistical description, tagging of web resources, and the *Resource Description Framework*(RDF) language used to make statements about web resources and physical resources that can be identified on the Web.

## *Bibliographic Descriptions*

The purposes and nature of bibliographic description are the foundation of library and information science and have been debated and systematized for nearly two centuries. *Bibliographic descriptions* characterize information resources and the entities that populate the bibliographic universe, which include works, editions, authors, and subjects.<sup>4</sup>

4. Despite the “biblio-” root, bibliographic descriptions are applied to all of the resource types contained in libraries, not just books. Note also that this definition includes not just the information resources being described as distinct instances, but also as sets of related instances and the nature of those relationships.

An excellent source for both the history and theory of bibliographic description is [\*The Intellectual Foundation of Information Organization\*](#) by Elaine Svenonius ([Svenonius 2000](#)). She divides bibliographic descriptions into “those that describe information from those that describe its documentary embodiments,” contrasting conceptual or subject properties from those that describe physical properties (p. 54). A more radical contrast was proposed by ([Wilson 1968, p. 25](#)), who distinguished descriptions according to the kind of bibliographic control they enabled. *Descriptive control* is objective and straightforward, *lining up a population of writings in any arbitrary order*. *Exploitative control*, defined as the ability

A bibliographic description of an information resource is typically realized as a structured record in a standardized format that describes a specific resource. The earliest bibliographic records in the nineteenth century were those in book catalogs, which organized for each author a list of his authored books, with separate entries for each edition and physical copy. Relationships between books by different authors were described using cross-references.

The nature and extent of bibliographic descriptions were highly constrained by the book catalog format, which also made the process of description a highly localized one because every library or collection of resources created its own catalog. The adoption of printed cards as the unit of organization for bibliographic descriptions around the turn of the twentieth century made it easier to maintain the catalog, and also enabled the centralized creation of the records by the Library of Congress.

The computerization of bibliographic records made them easier to use as aids for finding resources. However, digitizing legacy printed card-oriented descriptions for online use was not a straightforward task because the descriptions had been created according to cataloging rules designed for collections of books and other physical resources and intended only for use by people.

*to make the best use of a body of writings, requires descriptions that evaluate resources for their suitability for particular uses. Wilson argued that descriptive control was a poor substitute for exploitative control, but recognized that evaluative descriptions were more difficult to create.*

## *Metadata*

*Metadata* is often defined as “data about data,” a definition that is nearly as ubiquitous as it is unhelpful. A more content-full definition of metadata is that it is structured description for information resources of any kind. Metadata is more useful when supported by a metadata schema that defines the elements in the structured description.<sup>5</sup>

The concept of metadata originated in information systems and database design in the 1970s, so it is much newer than that of bibliographic description. The earliest metadata schemas, called data dictionaries, documented the arrangement and content of data fields in the records used by transactional applications on mainframe computers. A more sophisticated type of metadata emerged as the documentation of the data models in database management systems, called database schemas, which described the structure of relational tables, attribute names, and legal data types and values for content.

In 1986, the *Standard Generalized Markup Language*(SGML) formalized the *Document Type Definition*(DTD) as a metadata form for describing the structure and content elements in hierarchical and hypertextual document models. SGML was superseded in 1997 by eXtensible Markup Language(XML), whose purpose was structured and computer-processable web content.<sup>6</sup>

5. ([Gill 2008](#))
6. ([Rubinsky and Maloney 1997](#)) capture this transitional perspective. A more recent text on XML is ([Goldberg 2008](#)).

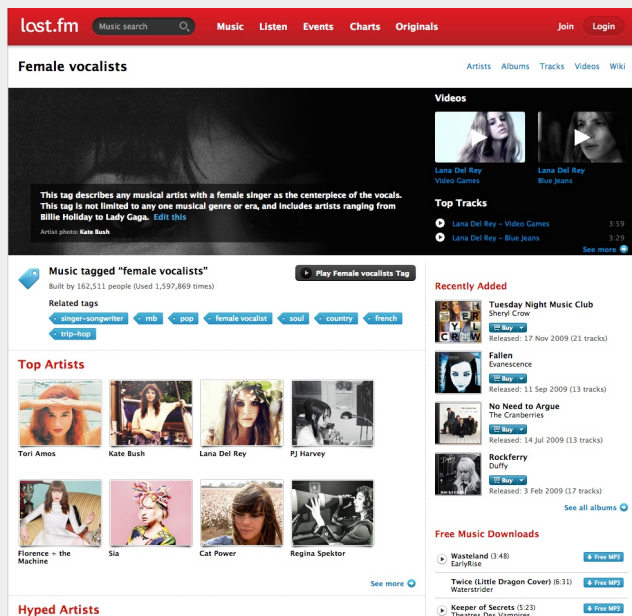
Today, XML schemas and other web- and compute-friendly formats for resource description have broadened the idea of [resource description](#) far beyond that of bibliographic description to include the description of software components, business and scientific datasets, web services, and computational objects in both physical and digital formats. The resource descriptions themselves serve to enable discovery, reuse, access control, and the invocation of other resources needed for people or computational agents to effectively interact with the primary ones described by the metadata.<sup>7</sup>

### *Tagging of Web-based Resources*



Tags on Last.fm

7. See ([Sen 2004](#)), ([Laskey 2005](#)).



Last.fm analyzes tags and other metadata to create rich multimedia “discovery” pages that bring together artist catalogs, new songs, free downloads, and music videos that its algorithm predicts will satisfy a user’s taste. This allows users to browse for new music in a more intuitive manner than searching by artist or genre.

(Screenshot by Ian MacFarland.)

The concept of metadata has been extended to include the tags, ratings, bookmarks or other types of descriptions that individuals apply to individual photos, blog or news items, or any other resource with a web presence. The practice of tagging has emerged as a way to apply labels to content in order to describe and identify it. Sets of tags are useful in managing one’s collection of websites



or digital media, in sharing them with others, and enabling new types of interactions and services.<sup>8</sup> For example, users of Last.fm tag music with labels that describe its nature, era, mood, or genre, and Last.fm uses these tags to generate radio stations that play music similar to that tag and related tags.

But tagging has a downside. The tendency for users to tag intuitively and spontaneously revives the vocabulary problem ([“Naming Resources”](#)) because one photographer’s “tree” is another’s “oak.” Likewise, unsystematic word choice leads to morphological inconsistency ([“Relationships among Word Forms”](#)); the same photo might be tagged with “burning” and “trees” and also with “burn” and “tree” by another. This disparity in the descriptors people use to categorize the same or similar resources can turn systems that use tagging into a “tag soup” lacking in structure.<sup>9</sup>

Some social media sites have incorporated mechanisms to make the tagging activity more systematic and to reduce [vocabulary problems](#).

8. See [\(Marlow, Naaman, Boyd, and Davis 2006\)](#). These authors propose a conceptual model of tagging that includes (1) tags assigned to a specific resource, (2) connections or links between resources, and (3) connections or links between users and explain how any two of these can be used to infer information about the other.
9. [\(Hammond, Hanney, Lund, and Scott 2004\)](#) coined the phrase “tag soup” in an review of social bookmarking tools written early in the tagging era that remains insightful today. Many of the specific tools are no longer around, but the reasons why people tag are still the same.

For example, on Facebook, users can indicate that a specific person is in an uploaded picture by clicking on the faces of people in photographs, typing the person's name, and then selecting the person from a list of Facebook friends whose names are formatted the way they appear on the friend's profile. Some social media systems suggest the most popular tags, perform morphological normalization, or allow users to arrange tags in bundles or hierarchies.<sup>10</sup>

### *Resource Description Framework (RDF)*

The Resource Description Framework(RDF) is a standard model for making computer-processable statements about web resources; it is the foundation for the vision of the [Semantic Web](#).<sup>11</sup> We have been using the word “resource” to refer to anything that is being organized. In the context of [RDF](#) and the web, however, “[resource](#)” means something more specific: a resource is anything that has been given a Uniform Resource Identifier(URI). [URIs](#) can take

10. Making tagging more systematic leads to “tag convergence” in which the distribution of tags for a particular resource stabilizes over time ([Golder and Huberman 2006](#)). Consider three things a user might do if his tag does not match the suggested tags; (1) Change the tag to conform? (2) Keep the tag to influence the group norm? (3) Add the proposed tag but keep his tag as well?
11. ([RDF Working Group 2004](#)). The official source for all things [RDF](#) is the [W3C RDF](#) page at <http://www.w3.org/RDF/>.

various forms, but you are probably most familiar with the URIs used to identify web pages, such as `http://springfield-elementary.edu/`. (You are probably also used to calling these URLs instead of URIs.) The key idea behind RDF is that we can use URIs to identify not only things “on” the web, like web pages, but also things “off” the web like people or countries. For example, we might use the URI `http://springfield-elementary.edu/` to refer to Springfield Elementary itself, and not just the school’s web page.

RDF models all descriptions as sets of “triples,” where each triple consists of the resource being described (identified by a URI), a property, and a value. Properties are resources too, meaning they are identified by URIs. For example, the URI <http://xmlns.com/foaf/0.1/schoolHomepage> identifies a property defined by the *Friend of a Friend*(FOAF) project for relating a person to (the web page of) a school they attended. Values can be resources too, but they do not have to be: when a property takes simple values like numbers, dates, or text strings, these values do not have URIs and so are not resources.

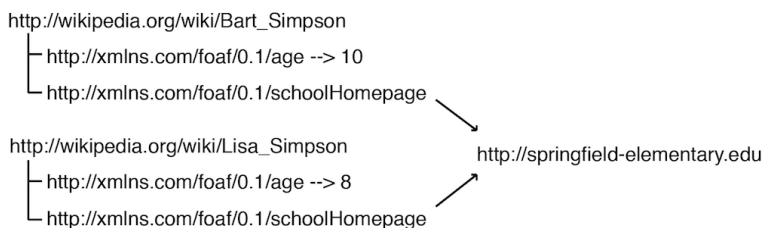
>

Because RDF uses URIs to identify described resources, their properties, and (some) property values, the triples in a description can be connected into a network or graph. [Figure: RDF Triples Arranged as a Graph](#) shows four triples that have been connected into a graph. Two of the triples describe Bart Simpson, who is identified using the URI of his Wikipedia page.<sup>12</sup> The other two

12. Some argue that the resource being described is thus Bart Simpson’s Wikipedia page, not Bart Simpson himself. Whether or not that is an important distinction is a controversial question among RDF architects and users.

describe Lisa Simpson. Two of the triples use the property `age`, which takes a simple number value. The other two use the property `schoolHomepage`, which takes a resource value, and in this case they happen to have the same resource (Springfield Elementary's home page) as their value.

### RDF Triples Arranged as a Graph



Two RDF triples can be connected to form a graph when they have a resource, property, or value in common. In this example RDF triples that make a statement about the home page of the elementary school attended by Bart Simpson and Lisa Simpson can be connected because they have the same value, namely the URI for Springfield Elementary.

Using URIs as identifiers for resources and properties allows descriptions modeled as RDF to be interconnected into a network of “linked data,” in the same way that the web enabled information to be interconnected into a massive network of “linked documents.” Proponents of RDF claim that this will greatly benefit knowledge discovery and inference.<sup>13</sup> But the benefits of RDF’s highly

13. ([Heath and Bizer 2011](#)) and <http://linkeddata.org> are excellent sources.

prescriptive description form must be weighed against the costs; turning existing descriptions into RDF can be labor-intensive.

RDF can be used for bibliographic description, and some libraries are exploring whether RDF transformations of their legacy bibliographic records can be exposed and integrated with resource descriptions on the open web. This activity has raised technical concerns about whether the RDF model of description is sophisticated enough and more fundamental concerns about the desirability of losing control over library resources.<sup>14</sup>

### *Aggregated Information Objects*

In the pre-digital age, information objects came with explicit tangible boundaries. Books consisted of pages bound within a cover, a vinyl record album physically bound together a set of songs (you could even see the groove pattern separating the songs), a movie was delivered on a strip of film spooled onto a reel, and a collection was (usually) demarcated as a designated shelf or room in a library.

Boundaries of information objects in the digital realm are neither tangible nor obvious. Consider the simple notion of a web page. Our cognitive notion of that which is rendered in our browser window

14. ([Byrne and Goddard 2010](#)) present a balanced analysis of the cultural and technical obstacles to the adoption of RDF and linked data in libraries. ([Yee 2009](#)) is a highly specific technical demonstration of converting bibliographic descriptions to RDF. A detailed analysis / rebuttal of Yee's article is at <http://futurelib.pbworks.com/w/page/13686677/YeeRDF>.

(e.g., some formatted text with an associated image) is actually, in web architecture terms ([Jacobs & Walsh, 2004](#)), three information objects (aka resources); the HTML encoding the text, CSS that defines the formatting rules, and the JPEG that encodes the image. All three have URLs and can independently be retrieved and linked. The situation is even more ambiguous for the common notion of a web site, the boundaries of which are not defined technically and are cognitively difficult to express.

Aggregations can be convenience methods for simplifying dissemination or organization, but they can also be transformative; resources can derive nearly all their value from their inclusion in an aggregation. On a web page, the CSS file is virtually useless on its own, since its role is to style the HTML file. In iTunes, the playback and organization functions are optimized for pop music, where individual songs can usually stand on their own when separated from the rest of an album. Classical music fans often struggle with this, because the individual “tracks” of a recording, split up to reduce file size and facilitate navigation through long works, are not separable; pieces are meant to be listened to in their entirety, and it can be difficult to ensure that they are aggregated together and have the proper metadata assigned to their aggregations. In other words: you can’t listen to symphonies on shuffle.<sup>15</sup>

The problem here is how to architecturally and technically express the notion of an aggregation, a set of information objects that, when considered together, compose another named information object. Aggregations are prevalent all over our digital information space: the web page and site mentioned above; a scholarly publication consisting of text, figures, and data; a dataset that is the composition of multiple data files. Notably the notion is both

15. ([Pancake 2012](#))

recursive and non-exclusive. An object that is itself an aggregation may be aggregated into another object. Information objects included in one aggregation may also be included in other aggregations, allowing reuse and re-factoring of existing information objects. A solution to this problem is a critical aspect of organizing digital information because, without well-defined boundaries we cannot deterministically identify, reference, or describe information objects.

The following are a number of technical approaches to the aggregation problem.

**Kahn-Wilensky Digital Object Framework.** Robert Kahn and Robert Wilensky coined the term “digital object” in a paper describing the core components of digital library infrastructure; the content, naming scheme, repository configuration, and access protocol. The digital object they describe is a uniquely-identified container packaging multiple data and metadata components. The model incorporates the recursive notion of a container of containers. The Kahn/Wilensky framework was the inspiration for the Fedora (Flexible Extensible Digital Object Repository Architecture) system, open source software deployed worldwide in information applications that leverage the container model.<sup>16</sup>

**Warwick Framework.** An early result of work by the Dublin Core Metadata Initiative, the Warwick Framework was motivated by the

16. [\(Kahn & Wilensky, 1995\) \(Lagoze, Payette, Shin, & Wilper, 2005\)](#)

desire to associate multiple metadata packages with content (e.g., descriptive, rights, administrative). It specifies a container architecture with distinct metadata packages that could be included directly or by-reference, allowing reuse of individual packages. The Warwick Framework was the inspiration for METS, described next.<sup>17</sup>

**METS.** The [Metadata Encoding and Transmission Specification](#) is a widely deployed XML-based container format that packages together multiple metadata components and content, both either directly or by-reference. Metadata packages are classified into one of a set of pre-defined types; descriptive, administrative, rights, and structural. METS is specified by an XML schema that has been extended for a variety of specialized applications.<sup>18</sup>

**OAI-ORE.** The Open Archives Initiative's [Object Reuse and Exchange](#) specification was motivated by a desire for an aggregation architecture fully congruent with web architecture principles, thus exposing aggregations to standard web tools, e.g. browsers, crawlers, HTTP servers. OAI-OARE introduces the notion of two types of URI-identified web resources; a resource map and an aggregation. When de-referenced through its respective URI, the former returns an REF/XML formatted description that establishes the identity of the aggregation, the resources that are included in the aggregation, and the semantics of the relationships among them.<sup>19</sup>

17. ([Lagoze, 1996](#))

18. ([McDonough, 2006](#))

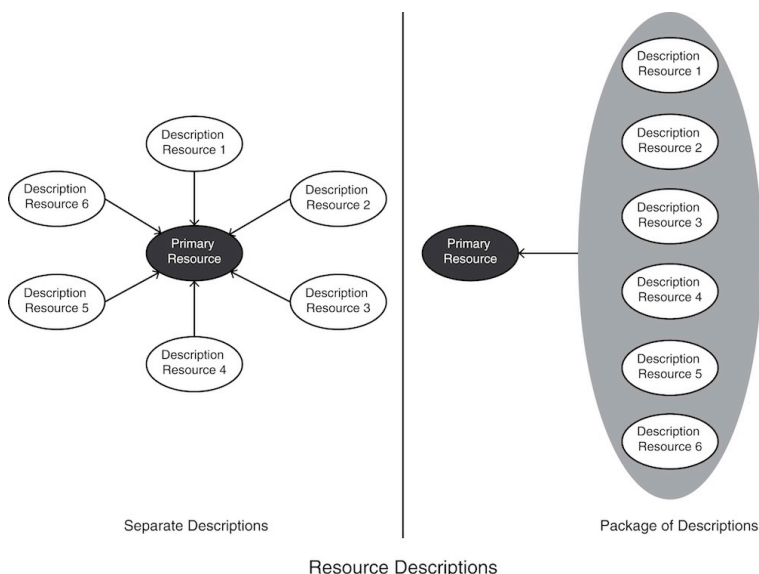
19. ([Lagoze et al., 2008](#))



# Frameworks for Resource Description

The broad scope of resources to which descriptions can be applied and the different communities that describe them means that many [frameworks](#) and classifications have been proposed to help make sense of resource description.

## Architectures for Resource Description



Two contrasting architectures for resource descriptions are separate descriptions versus packaged descriptions, which were dominant in library catalogs with printed cards containing descriptions about a resource.

The dominant historical view treats resource descriptions as a package of statements; this view is embodied in the printed library card catalog and its computerized analog in the [MARC21](#) format

(an exchange format for library catalog records), which contains many fields about the bibliographic characteristics of an object like author, title, publication year, publisher, and pagination. An alternate architecture for resource description focuses on each individual description or assertion about a single resource, as the RDF and linked data approaches do. These two alternatives are contrasted in [Figure: Architectures for Resource Description](#).

In either case, these common ways of thinking about resource description emphasize—or perhaps even overemphasize—two implementation decisions:

- The first is whether to combine multiple resource descriptions into a structural package or to keep them as separate descriptive statements.
- The second is the choice of syntax in which the descriptions are encoded.

Both of these implementation decisions have important implications, but are secondary to the questions about the purposes of resource description, how resource properties are selected as the basis for description, how they are best created, and other logical or design considerations. In keeping with a fundamental idea of the discipline of organizing (introduced in [“The Concept of “Organizing Principle”](#)”), it is imperative to distinguish design principles from implementation choices. We treat the set of implementation decisions about character notations, syntax, and structure as the *form* of resource description and we will defer them as much as we can until [The Forms of Resource Descriptions](#).

In library and information science, it is very common to discuss resource descriptions using a classification proposed by Arlene Taylor, which distinguishes administrative, structural, and

descriptive metadata.<sup>20</sup> A similar typology proposed by Gilliland breaks metadata down into five types: administrative, descriptive, preservation, use, and technical.<sup>21</sup>

Resource description is not an end in itself. Its many purposes are all means for enabling and using an organizing system for some collection of resources. As a result, our framework for resource descriptions aligns with the activities of organizing systems we discussed in [Activities in Organizing Systems](#): selecting, organizing, interacting with, and maintaining resources.

20. ([Taylor and Joudrey 2009](#)) Taylor's book on [The Organization of Information](#), now in its 3<sup>rd</sup> edition (with co-author Daniel Joudrey), has been widely used in library science programs for over a decade.

21. ([Gilliland-Swetland 2000](#)).

## 33. The Process of Describing Resources

We prefer the general concept of resource description over the more specialized ones of bibliographic description and metadata because it makes it easier to see the issues that cut across the domains where those terms dominate. In addition, it enables us to propose more standard process that we can apply broadly to the use of resource descriptions in organizing systems. A shared vocabulary enables the sharing of lessons and best practices.

The process of describing resources involves seven interdependent and iterative steps. We begin with a generic summary of the process to set the stage for a detailed step-by-step discussion.

1. Identifying resources to describe is the first step; this topic is covered in detail in [“Resource Identity”](#). The resource [domain](#) and [scope](#) circumscribe the describable properties and the possible purposes that descriptions might serve. The resource [focus](#) determines which are primary information resources and which ones are treated as the corresponding resource descriptions. Two important decisions at this stage are [granularity](#) of description—are we describing individual resources or collections of them?—and the *abstraction* level—are we describing resource instances, parts of them, or resource types?
2. Generally, the purpose of resource description is to support the activities common to all organizing systems: selecting, organizing, interacting with, and maintaining resources, as we saw in [Activities in Organizing Systems](#). The particular resource domain and the context in which descriptions are created and

used imposes more specific requirements and constraints on the purposes that resource description can serve.

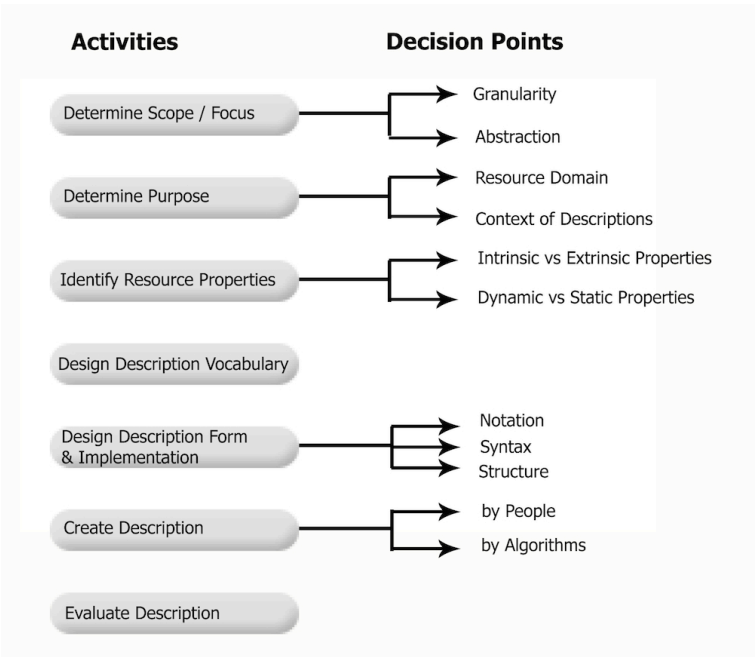
3. Once the purposes of description in terms of activities and interactions have been determined, the specific properties of the resources that are needed to enable them can be identified. The goal of description is not to be exhaustive; there are always more possible properties than can be reasonably described. Instead, the challenge is to use the properties that are most robust and reliable for supporting the desired interactions.
4. This step includes several logical and semantic decisions about how the resource properties will be described. What terms or element names should be used to identify the resource properties we have chosen to describe? Are there rules or constraints on the types of data or values that the property descriptions can assume? When dealing with numerical descriptions, their data types and levels of measurement constrain the kinds of processing to which they may submit. Nominal, ordinal, interval, and ratio data each are limited to particular transformations based on what they represent. A good description vocabulary will be easy to assign when creating resource descriptions and easy to understand when using them.
5. The logical and semantic decisions about the description vocabulary are reified by decisions about the notation, syntax, and structure of the descriptions. Taken together, these decisions collectively determine what we call the *form* or *encoding* of the resource descriptions. The implementation of the descriptions involves decisions about how and where they are stored and the technology used to create, edit, store, and retrieve them.
6. Resource descriptions are created by individuals, by informal or formal groups of people, or by automated or computational means. Some types of descriptions can only be created by

people, some types of descriptions can only be created by automated or algorithmic techniques, and some can be created in either manner.

- 7. The resource descriptions must be evaluated with respect to their intended purposes. The results of this evaluation will help determine which or the preceding steps need to be redone.

The next seven sub-sections discuss each of these steps in detail. A quick reference guide is [Figure: The Process of Describing Resources](#).

The Process of Describing Resources



The process of describing resources consists of seven steps: Determining the scope and focus, determining the purpose,

identifying resource properties, designing the description vocabulary, designing the description form and implementation, creating the descriptions, and evaluating the descriptions.

How explicit and systematic each step needs to be depend on the resource domain and scope, and especially on the intended users of the organizing system. If we look carefully, we can see most of these steps taking place even in very informal contexts, like the kids playing with Lego blocks with which we started this chapter. The goal of building things with the blocks leads the boys to identify which properties are most useful to analyze. They develop descriptions of the blocks that capture the specific values of the relevant properties. Finally, they evaluate their descriptions by using them when they play together; it becomes immediately obvious that a description is not serving its purpose when one boy hands a block to another that was not the one he thought he had asked for.

In contrast, a picture-taking scenario involves a much more explicit and systematic process of resource description. The resource properties, description vocabulary, and description form used automatically by a digital camera were chosen by an industry association and published as a technical specification implemented by camera and mobile phone manufacturers worldwide.

The resource descriptions used by libraries, archives, and museums are typically created in an even more explicit and systematic manner. Like the descriptions of the digital photo, the properties, vocabulary, and form of the descriptions used by their organizing systems are governed by standards. However, there is no equivalent to the digital camera that can create these descriptions automatically. Instead, highly trained professionals create them meticulously.

A great many resources and their associated descriptions in business and scientific organizing systems are created by automated or computational processes, so the process of describing individual resources is not at all like that in libraries and other

memory institutions. However, the process for designing the data models or schemas for the class of resources that will be generated is equally systematic and is typically performed by highly skilled data analysts and data modelers.

## Determining the Scope and Focus

Which resources do we want to describe? As we saw in [Resources in Organizing Systems](#), determining what will be treated as a separate resource is not always easy, especially for resources with component parts and for information resources where the most important property is their content, which is not directly perceivable. Identifying the thing you want to describe as precisely as practical is the first step to creating a useful description.

In “[Resource Focus](#)”, we introduced the contrast between primary resources and description resources, which we called resource focus. Determining the resource focus goes hand in hand with determining which resources we intend to describe; these often arbitrary decisions then make a huge difference in the nature and extent of resource description. One person’s metadata is another person’s data.

- For a librarian, the price of a book might be just one more attribute that is part of the book’s record.
- For an accountant at a bookstore, the price of that book—both the cost to buy the book and the price at which it is then sold to customers—is critical information for staying in business.
- In a medical records context, a patient’s insurance provider isn’t of much concern to the doctor, but to the person responsible for billing, it is central. For the nurse, the patient’s current vital signs may be of most importance, while for the doctor, it may



be most important to understand how those data in aggregate serve to indicate a longer-term prognosis of the patient's health.

- A scientist studying comparative anatomy preserves animal specimens and records detailed physical descriptions about them, but a scientist studying ecology or migration discards the specimens and focuses on describing the context in which the specimen was located.

### *Describing Instances or Describing Collections*

It is simplest to think of a resource description as being associated with another individual resource. As we discussed in [Resources in Organizing Systems](#), it is challenging to determine what to treat as an individual resource when resources are themselves objects or systems composed of other parts or resources. For example, we sometimes describe a football team as a single resource and at other times we focus on each individual player. However, after deciding on resource granularity, the question remains whether each resource needs a separate description.

Libraries and museums specialize in curating resource descriptions about the instances in their collections. Resource descriptions are also applied to classes or collections of resources, because a collection is also a resource ([“The Concept of “Collection””](#)). Archives and special collections of maps are typically assigned resource descriptions, but each document or map contained in the collection does not necessarily have its own bibliographic description. Similarly, business and scientific datasets are invariably described at collection-level granularity because they are often analyzed in their entirety.

Furthermore, the granularity of description for a collection of resources tends to differ for different users or purposes. An investor

who owns many different stocks focuses on their individual prices, while other investors put their money in index funds that combine all the separate prices into a single value.

Many web pages, especially e-commerce product catalogs and news sites, are dynamically assembled and personalized from a large number of information resources and services that are separately identified and described in content management and content delivery systems. However, a highly complex collection of resources that comes together in a single page is treated as a single resource when that page appears in a list of search engine results. Moreover, all of the separately generated pages can be given a single description when a user creates a bookmark to make it easy to return to the home page of the site.

### *Abstraction in Resource Description*

We can also associate resource descriptions with an entire type or domain of resources. (See [“Preserving Resource Types”](#) and [“Resource Domain”](#).) A collection of resource descriptions is vastly more useful when every resource is described using common description elements or terms that apply to every resource. A *schema* (or model, or metadata standard) specifies the set of descriptions that apply to an entire resource type. Sometimes this schema, model, or standard is inferred from or imposed on a collection of existing resources to ensure more consistent definitions, but more often, it is used as a specification when the resources are created or generated in the first place. (See [What about “Creating” Resources?](#) in [“Introduction”](#).)

A relational database, for example, is easily conceptualized as a collection of records organized as one or more tables, with each record in its own row having a number of fields or attributes that contain some prescribed type of content. Each record or row in

the database table is a description of a resource—an employee, a product, anything—and the individual attribute values, organized by the columns and rows of the table, are distinct parts of the description for some particular resource instance, like employee 24 or product 8012C.<sup>1</sup>

1. Because the relational database schema serves as a model for the creation of resource descriptions, it is designed to restrict the descriptions to be simple and completely regular sets of attribute-value pairs. The database schema specifies the overall structure of the tables and especially their columns, which will contain the attribute values that describe each resource. An employee table might have columns for the attributes of employee ID, hiring date, department, and salary. A date attribute will be restricted to a value that is a date, while an employee salary will be restricted according to salary ranges established by the human resources department. This makes the name of the attribute and the constraints on attribute values into resource descriptions that apply to the entire class of resources described by the table.

It is often necessary to associate some descriptions with individual resources that are specific to that instance and other kinds of descriptions that reflect the abstract class to which the instance belongs. When a typical car comes off the assembly line, it has only one instance-level description that differentiates it from its peers: its vehicle identification number (VIN). Specific cars have

The information resources that we commonly call documents are, by their nature, less homogeneous in content and structure than those that can be managed in databases. Document schemas, commonly represented in SGML or XML, usually allow for a mixture of data-like and textual descriptive elements.

XML schema languages have improved on SGML and XML by expressing the description of the document schema in XML itself, making it easy to create resources using the metadata as a template or pattern. XML schemas are often used as the specifications for XML resources created and used by information-intensive applications; in this context, they are often called XML vocabularies. XML schemas can be used to define web forms that capture resource instances (each filled-out form). XML schemas are also used to describe the interfaces to web services and other computational resources.<sup>2</sup>

individualized interior and exterior colors and installed options, and they all have a date and location of manufacture. Other description elements have values that are shared with many other cars of the same model and year, like suggested price and the additional option packages, or configurations that can be applied to it before it is delivered to a customer. Alternatively, any descriptive information that applies to multiple cars of the same model year could be part of a resource description at that level that is referred to rather than duplicated in instance descriptions.

2. Web services are generally implemented using XML documents as their inputs and outputs. The interfaces

## *Scope, Scale, and Resource Description*

If we only had one thing to describe, we could use a single word to describe it: “it.” We would not need to distinguish it from anything else. A second resource implies at least one more term in the description language: “not it.” However, as a collection grows, descriptions must become more complex to distinguish not only between, but also among resources.

Every element or term in a description language creates a dimension, or axis, along which resources can be distinguished, or it defines a set of questions about resources. Distinctions and questions that arise frequently need to be easy to address, such as:

- What is the name of the resource?
- Who created it?
- What type of content or matter does it contain?

Therefore, as a collection grows, the language for describing resources must become more rigorous, and descriptions created when the collection was small often require revision because they are no longer adequate for their intended purposes.<sup>3</sup>

to web services are typically described using an XML vocabulary called *Web Services Description Language*(WSDL). See ([Erl 2005b](#)), especially Ch. 3, *Introduction to Web Services Technologies*.

3. Creating descriptions that can keep pace with the growth of a collection has been an issue for librarians for years, as libraries moved away from describing simply

This co-evolution of descriptive scope and description complexity is easy to see in the highly complex bibliographic descriptions created by professional catalogers. The commonly used Anglo-American Cataloguing Rules(AACR2) cataloging standards distinguish 11 different categories of resources and specify several hundred descriptive elements. AACR2 has recently been superseded by the *Resource Description and Access*(RDA) standards, which make fine-grained distinctions about content, media type, and carrier (technology).<sup>4</sup>

Because the task of library resource description has been standardized at national and international levels, cataloging work is distributed among many describers whose results are shared. The principle of standardization has been the basis of centralized bibliographic description for a century.

“whatever came across a cataloger’s desk” to cataloging resources for a national and even international audience ([Svenonius 2000, p. 31](#)).

4. The [AACR2](#) includes rules for books, pamphlets, and printed sheets; cartographic materials; manuscripts and manuscript collections; music; sound recordings; motion pictures and video recordings; graphic materials; electronic resources; 3-D artifacts; microforms; and continuing resources. The Concise [AACR2](#) (4th Edition) is the most accessible treatment of these very complex rules ([Gorman 2004](#)). The *Resource Description and Access*(RDA) vocabularies are the successor to [AACR2](#) and make even finer distinctions among resource types. See the RDA [content](#), [media](#), and [carrier](#) value lists.

Centralized resource description by skilled professionals works for libraries, but even in the earliest days of the web many library scientists and web authoring futurists recognized that this approach would not scale for describing web resources. In 1995, the *Dublin Core*(DC) metadata element set with only 15 elements was proposed as a vastly simpler description vocabulary that people not trained as professional catalogers could use. Since then, the Dublin Core initiative has been highly influential in inspiring numerous other communities to create minimalist description vocabularies, often by simplifying vocabularies that had been devised by professionals for use by non-professionals.<sup>5</sup>

Of course, a simpler description vocabulary makes fewer distinctions than a complex one; replacing “author,” “artist,” “composer” and many other descriptions of the person or non-human resource responsible for the intellectual content of a resource with just “creator” (as Dublin Core does) results in a substantial loss of precision when the description is created and can cause misunderstanding when the descriptions are reused.<sup>6</sup>

5. We can also view the Dublin Core as part of the intellectual foundations for the “crowdsourcing” or “community curation” of resource descriptions by non-professionals (“[Social and Web Curation](#)”). See the *Dublin Core Metadata Initiative* (DCMI) at <http://dublincore.org/>.

6. The semantic “bluntness” of a minimalist vocabulary is illustrated by the examples for use of the “creator” element in an official Dublin Core user guide ([Hillmann 2005](#)) that shows “Shakespeare, William” and “Hubble Telescope” as creators.

The negative impacts of growing scope and scale on resource description can sometimes be avoided if the ultimate scope and scale of the organizing system is contemplated when it is being created. It would not be smart for a business with customers in six US states to create an address field in its customer database that only handled those six states; a more extensible design would allow for any state or province and include a country code. In general, however, just as there are problems in adapting a simple vocabulary as scope and scale increase, designing and applying resource descriptions that will work for a large and continuously growing collection might seem like too much work when the collection at hand is small.

The challenges that arise with large description vocabularies are transformed when resource descriptions are created and assigned by computer algorithms. A large dataset might contain many thousands of descriptions for each resource, but clearly the computer does not have cognitive difficulty generating or using them. However, computer models with this many features can be hard for people to understand and trust.

## Determining the Purposes

Resource description serves many purposes, and the mix of purposes and the resulting kinds of descriptions in any particular organizing system depends on the scope and scale of the resources being organized. We can identify and classify the most common purposes using the four activities that occur in every organizing system: selecting, organizing, interacting with, and maintaining resources (see [Activities in Organizing Systems](#)). Resource description also has a more open-ended purpose in [sensemaking](#) and science (see [“Resource Description for Sensemaking and Science”](#)); we observe and describe the world to make sense of our experiences and to predict future observations.



## *Resource Description to Support Selection*

Defining selection as the process by which resources are identified, evaluated, and then added to a collection in an organizing system emphasizes resource descriptions created by someone other than the person who is using them. We can distinguish several different ways in which resource description supports selection:

### **Discovery**

What available resources might be added to a collection? New resources are often listed in directories, registries, or catalogs. Some types of resources are selected and acquired automatically through subscriptions or contracts.

### **Capability and Compatibility**

Will the resource meet functional or interoperability requirements? Technology-intensive resources often have numerous specialized types of descriptions that specify their functions, performance, reliability, and other “-ilities” that determine if they fit in with other resources in an organizing system.<sup>7</sup> Some services have qualities of service levels, terms and conditions, or interfaces documented in resource

7. The Intel Core 2 Duo Processor has detailed specifications (<http://www.intel.com/products/processor/core2duo/specifications.htm>) and seven categories of technical documentation: application notes, datasheets, design guides, manuals, updates, support components, and white papers (<http://www.intel.com/design/core2duo/documentation.htm>).

descriptions that affect their compatibility and interoperability. Some resources have licensing or usage restrictions that might prevent the resources from being used effectively for the intended purposes. Decisions about “people selection” are becoming more data-driven, and sports teams, business employers, and dating sites now rely on predictive statistics to find the best person.

### **Authentication**

Is the resource what it claims to be? ([“Authenticity”](#)) Resource descriptions that can support authentication include technological ones like time stamps, watermarking, encryption, checksums, and digital signatures. The history of ownership or custody of a resource, called its provenance ([“Provenance”](#)), is often established through association with sales or tax records. Import and export certificates associated with the resource might be required to comply with laws designed to prevent the theft of antiquities or the transfer of technology or information with national security or foreign policy implications.

### **Appraisal**

What is the value of this resource? What is its cost? At what rate does it depreciate? Does it have a shelf life? Does it have any associated ratings, rankings, or quality measures? Moreover, what is the quality of those ratings, rankings, and measures?

We also consider the perspective of the person creating the resource description and his or her primary purpose, which is often to encourage the selection of the resource by someone else. Product marketing is about devising names and descriptions to make a resource distinctive and attractive compared to alternatives. For many years prunes were promoted as a dietary supplement that people (especially old ones) need to “maintain regularity.” But after the California Prune Board (the world’s biggest supplier) re-branded them as “dried plums” and started marketing them as a snack food

(and simultaneously renaming itself as the California Dried Plum Board) sales increased significantly.<sup>8</sup>

Many countries require that imported goods are labeled with their country or origin. Consumers often use this property in resource descriptions as an indicator of high quality, as they might with Swiss watches, French or Italian fashions, or Canadian bacon. Alternatively, consumers might want to buy domestic or locally-sourced goods out of economic patriotism or to comply with procurement regulations. Not surprisingly, when consumers view origin in a positive light, this information is conspicuous and easy to read. In contrast, when consumers view origin less positively, perhaps as a warning of low-quality goods, the supplier is likely to make the origin information as inconspicuous as legally possible, or might even misrepresent the goods as domestic ones.<sup>9</sup>

8. Real estate advertisements are notorious for their creative descriptions; a house “convenient to transportation” is most likely next to a noisy highway, and a house in a “secluded location” is in a remote and desolate part of town.

9. In its early days, when US consumers were generally unaware that Sony was a Japanese company and the quality of Japanese products was viewed in a negative light, Sony would make the “Made in Japan” label as inconspicuous as it could get away with. ([John 1999](#))

In the summer of 2015, the consumer advocacy organization Truth in Advertising reported finding on Walmart’s website over 100 product descriptions

This misrepresentation is also ubiquitous in online dating, though the amount of misrepresentation must be balanced with the goals of the relationship and chances of the deception being discovered.<sup>10</sup>

### *Resource Description to Support Organizing*

We have defined [organizing](#) as specifying the principles for describing and arranging resources to create the capabilities upon which interactions are based. This definition treats the creation of resource descriptions and their use to organize resources for interactions as separate and sequential activities. This is easiest to see when people assign keywords and classifications to documents, or when sensors produce data, and these resource descriptions are later used to enable document retrieval or data analysis. A department store clerk might sort dress shirts on a display table

inaccurately presenting the products as being made in the United States. (See <https://www.truthinadvertising.org/walmart-made-in-usa/>)

10. Findings from a study of four online dating services ([Toma et al 2008](#)) found that 81% of people lied about at least one characteristic. Men were more likely to lie about height, while women lied more about weight, and the further their actual heights and weights were from the mean, the more they lied. A later study ([Hall et al 2010](#)) confirmed the finding for women and weight but also found that men are highly likely to misrepresent their personal assets.

using labels that describe their brands, sizes, and other properties. Rules governing the collection, integration, and analysis of personal information are also resource descriptions that influence the organization of information resources.

However, even if resource description and resource organization are logically separable, at times they are intertwined. When you arrange your own clothes, you don't use explicit resource descriptions and instead rely on implicit ones about easily perceived properties like color, shape, and material of composition. When algorithms rather than people analyze texts to identify descriptive features for applications like information retrieval, spam classification, and sentiment analysis, resource descriptions and resource organization co-evolve, often continuously as the algorithm adapts and learns with each new resource it describes. This tight connection between resource description and resource organization is also exploited in organizing systems that use usage records from session logs, browsing, or downloading activities as interaction resources, tying them to payments for using the resources or analyzing them to influence the selection and organizing of resources in future personalized interactions. (See ["The Concept of "Interaction Resource"](#))

### *Resource Description to Support Interactions*

Most discussions of the purposes of resource descriptions and metadata emphasize the interactions that are based on resource descriptions that have been intentionally and explicitly assigned. The Functional Requirements for Bibliographic Records (FRBR), defined by library scientists, specifies the four interactions of

Finding, Identifying, Selecting, and Obtaining resources, but these apply generically to organizing systems, not just those in libraries.<sup>11</sup>

11. We encountered FRBR several times in previous chapters, especially in [“Identity and Bibliographic Resources”](#), where we asked “What is this thing we call ‘Macbeth?’” and which describes FRBR four-level abstraction hierarchy of the work, depicted in [Figure: The FRBR Abstraction Hierarchy](#).

The FRBR framework is the most recent formalization of the purposes of resource description that started in nineteenth-century libraries. This long history means it is not surprising that how we think about resource description still shows some bias toward interactions with physical bibliographic resources and the descriptions needed to obtain them.

There are three members of the “FRBR family.” One member is the “Group 1 entities” — [Work](#), [Expression](#), [Manifestation](#), and [Item](#)—which are used to define classes of intellectual products and their relationships to each other. The “Group 2 entities” are responsible for the creation or custodianship of the Group 1 entities — Person, Family, and Corporate Body. The Group 2 model and requirements have been developed further as the *Functional Requirements for Authority Data*(FRAD) to enable catalogs to answer questions about the relationships between corporate entities or

## **Finding**

What resources are available that “correspond to the user’s stated search criteria” and thus can satisfy an information need? Before there were online catalogs and digital libraries, we found resources by referencing catalogs of printed resource descriptions incorporating the title, author, and subject terms as access points into the collection; the subject descriptions were the most important finding aids when the user had no particular resource in mind. Modern users accept that computerized indexing makes search possible over not only the entire description resource, but often over the entire content

collaborators; see <http://www.ifla.org/publications/functional-requirements-for-authority-data> and <http://archive.ifla.org/VII/d4/wg-franar.htm#Authority>. The “Group 3 entities” are the subject descriptions for Group 1 and Group 2; see the *Functional Requirements for Subject Authority Data*(FRSAD) at <http://www.ifla.org/node/1297>.

For bibliographic resources these interactions and the models of resource descriptions needed to support them have been formalized as the Functional Requirements for Bibliographic Records(FRBR). We encountered FRBR several times in previous chapters (especially in [“Bibliographic Resources, Information Components, and “Smart Things” as Resources](#)”) where we asked “What is this thing we call ‘Macbeth?’” and described the FRBR four-level abstraction hierarchy of the [work](#).

of the primary resource. Businesses search directories for descriptions of company capabilities to find potential partners, and they also search for descriptions of application interfaces (APIs) that enable them to exchange information in an automated manner.

### **Identifying**

Another purpose of resource description is to enable a user to confirm the identity of a specific resource or to distinguish among several that have some overlapping descriptions. In bibliographic contexts, this might mean finding the resource that is identified by its citation. Computer processable resource descriptions like bar codes, QR codes, or RFID tags are also used to identify resources. In Semantic Web contexts, URIs serve this purpose. Color can be used as resource descriptions when physical resources need to be identified quickly.<sup>12</sup>

12. In the very busy and dangerous environment of an aircraft carrier flight deck, the sailors wear vests and shirts that are color-coded to their jobs. For example, red shirts handle munitions, purple shirts handle fuel, green shirts run the catapults and hooks that launch and land the jets, and yellow shirts manage the flights. Color makes it faster and takes less attention for people to see if the right people are where they are supposed to be

The official Navy color chart for aircraft carrier personnel is available at <http://www.navy.mil/navydata/ships/carriers/rainbow.asp>

A similar principle is used in some sports; goalies wear



## Selecting

Selecting in this context means the user activity of using resource descriptions to support a choice of resource from a collection, not the institutional activity of selecting resources for the collection in the first place. Search engines typically use a short “text snippet” with the query terms highlighted as resource descriptions to support selection. People often select resources with the least restrictions on uses as described in a Creative Commons license.<sup>13</sup> A business might select a supplier or distributor that uses the same standard or industry reference model to describe its products or business processes

different color jerseys to make it easier to enforce position-specific rules, and football quarterbacks wear distinctive practice jerseys to remind defensive players not to tackle them and possibly injure them.

It is worth noting that color blindness affects approximately 7% of the population.

13. The Creative Commons nonprofit organization defines six kinds of copyright licenses that differ in the extent they allow commercial uses or modifications of an original resource (see <http://creativecommons.org/licenses/>). The Flickr photo sharing application is a good example of a site where a search for reusable resources can use the Creative Commons licenses to filter the results (<http://www.flickr.com/creativecommons/>).

because it is almost certain to reduce the cost of doing business with that business partner.<sup>14</sup>

### **Obtaining**

Physical resources often require significant effort to obtain after they have been selected. Catching a bus or plane involves coordinating your current location and time with the time and location the resource is available. With information resources in physical form, obtaining a selected resource usually meant a walk through the library stacks. With digital information resources, a search engine returns a list of the identifiers of resources that can be accessed with just another click, so it takes little effort to go from selecting among the query results to obtaining the corresponding primary resource.<sup>15</sup>

14. Using the same standards to describe products or to specify the execution of business processes can facilitate the implementation and operation of information-intensive business models because information can then flow between services or firms without human intervention. In turn this enables the business to become more demand or event-driven rather than forecast driven, making it a more “adaptive,” “agile,” or “on-demand” enterprise. See ([Glushko and McGrath 2005](#)), especially Ch. 5, How Models and Patterns Evolve.
15. For new resources, the labor-intensive cost of traditional bibliographic description is less justifiable when you can follow a link from a resource description to the digital resource it describes and quickly decide its relevance.

Elaine Svenonius proposed that a fifth task called Navigation be added to the [FRBR](#) list, and in 2016 that happened but it was renamed as “Explore”:<sup>16</sup>

### ***Navigation or Explore***

If users are not able to specify their information needs in a way that the [finding](#) functionality requires, they should be able to use relational and structural descriptions among the resources to navigate from any resource to other ones that might be better. Svenonius emphasizes generalization, aggregation, and derivational relationships.<sup>17</sup> But in principle, any relationship or property could serve as the navigation “highway” between resources.

What some authors call “structural metadata” can be used to support the related tasks of moving within multi-part digital resources like electronic books, where each page might have associated information about previous, next, and other related pages. Documents described using [XML](#) models can use *Extensible Stylesheet Language Transformations*(XSLT) and

That is, web search engines demonstrate that algorithmic analysis of the content of information resources can make them self-describing to a significant degree, reducing the need for bibliographic description.

16. See the [FRBR-Library Reference Model](#)

17. [\(Svenonius 2000, pages 18-19\).](#)

XPath to address and select data elements, sub-trees, or other structural parts of the document.<sup>18</sup>

### *Resource Description to Support Maintenance*

Many types of resource descriptions that support selection (“[Resource Description to Support Selection](#)”) are also useful over time to support maintenance of specific resource and the collection to which they belong. In particular, technical information about resource formats and technology (software, computers, or other) needed to use the resources, and information needed to ensure resource integrity is often called “preservation metadata” in a maintenance context.<sup>19</sup>

18. Ken Holman’s [Definitive XSLT and XPath \(Holman 2001\)](#) is the book to get started on with XPath, and no one has taught more people about XPath than Holman. The first five hours of a 24-hour video course on Practical Transformation Using XSLT and XPath is available for free at <http://www.udemy.com/practical-transformation-using-xslt-and-xpath>.
19. ([PREMIS 2011](#)). The Preservation Metadata Implementation Strategies(PREMIS) standard for preservation metadata is maintained by the US Library of Congress at <http://www.loc.gov/standards/premis/>. A good place to start is the 2011 PREMIS Data Dictionary(<http://www.loc.gov/standards/premis/v2/premis-2-1.pdf>).

Resource descriptions that are more exclusively associated with maintenance activities include version information and effectivity, or useful life information. Equipment maintenance schedules are typically related to the number of miles driven (indicated by a car's odometer), number of hours operated (stored by many engines), number of pages printed, or other easily recorded information about resource use or interactions. With smart resources now capable of capturing, analyzing, and communicating more data about real-time performance, more sophisticated prediction and scheduling of maintenance work is now possible. It is also easier to identify resources that are not being used as much as expected, which might imply that they are no longer needed and can thus be safely archived or discarded.

### *Resource Description for Sensemaking and Science*

Up to now in [“Determining the Purposes”](#), we have discussed how resource descriptions are used to perform well-defined tasks within an existing organizing system. However, there is a broader and less well-defined purpose of resource description that is older and more fundamental: the use of resource descriptions as the raw material for making sense of the world.

For thousands of years, even before the invention of written language, people have systematically collected things, information about those things, and observations of all kinds to understand how their world works. Paleolithic humans made cave paintings depicting the results of hunts and animal migrations; ancient Egyptians recorded the annual floods of the Nile River in stone carvings; and Babylonian, Egyptian, Chinese, and Mesoamerican astronomers organized lunar, solar, and planetary observations as calendars starting about five thousand years ago.

These diverse efforts to impose meaning on experience by

recording, analyzing, organizing, and reorganizing observations can be collectively described as [sensemaking](#). (See the sidebar, [Sensemaking and Organizing](#).)

### Sensemaking and Organizing

People organize to make sense of equivocal inputs and enact this sense back into the world to make it more orderly ([Weick 2005](#))

[Sensemaking](#) and [organizing](#) are intertwined. Ancient cultures recorded time-based observations and analyzed patterns among crop cycles, commodity prices, weather conditions, and astronomical sightings. Think back to the early astronomers, who oriented temple buildings to align with astronomical events and who decorated temple walls with zodiac imagery.

- Which of the planets and stars in the night sky should they observe and how should they record the details of those observations?
- What mathematical and statistical techniques should be used to analyze and describe these observations?
- What subset of observations are most useful in predicting the onset of the Nile River floods, caused by unobserved rainfall thousands of miles away?

Every choice about what to observe and how to describe it reflects a set of assumptions and potential

hypotheses that are often implicit and unstated. Choices that increase understanding are built upon, and those that fail to provide insight are abandoned, but there is no guarantee that the iterative process of choosing what to observe and describe will lead to a correct understanding.

The principle that an accurate or comprehensive dataset is insufficient on its own to yield a correct model is exemplified in the interlocking efforts of Tycho Brahe and Johannes Kepler. Brahe was a 16th-century Danish nobleman astronomer who spent decades collecting data about the positions of hundreds of stars and the planets. However, because of prevailing religious and scientific biases, Brahe accepted the incorrect assumptions that the sun and planets revolved around the earth in circular orbits. After Brahe died in 1601, Kepler spent a decade analyzing Brahe's data, and then rejected the idea of earth-centric and circular planetary orbits in favor of elliptical ones with the sun at one focus. These new organizing principles for Brahe's data made the model of the solar system vastly simpler, and Kepler was able to discover laws of planetary motion that are part of the foundation of modern astronomy and physics.<sup>20</sup>

20. [\(Lockyer 1893\)](#) and [\(Bell 1970\)](#)

The joint story of Brahe's data collecting and Kepler's analysis and theorizing is told in an entertaining manner in [\(Ferguson 2002\)](#). An equally fascinating analysis that

Some aspects of sensemaking are hard-wired by evolution, which has given our brains powerful mechanisms that automatically simplify and organize the perceptual data we obtain from the world (see the sidebar [Gestalt Principles](#)). But this automatic sensemaking is dominated and amplified by intentional sensemaking.

Intentional sensemaking takes place when systematic statistical, experimental, and scientific methods are consciously followed to extract and organize knowledge from collections of samples, observations, or measurements. It is critical to recognize here that the contents of these collections represent choices made about what to collect, because most things and most phenomena have a great many descriptions or properties that could be recorded about them.

After things or data have been collected, statistical methods summarize the values of properties in a collection or dataset and the relationships among them. Making sense of a single collection or dataset by determining the properties that contrast and classify the instances is the start toward the more important goal of understanding the larger set or population from which the initial collection is just a sample. There is no better example of this than the periodic table of elements developed by Mendeleev in 1869, who organized known elements on the basis of their common chemical properties and then successfully predicted some properties of yet undiscovered ones.

Computational models developed from the initial dataset can predict future observations. Classification models assign a new instance to a category (e.g., spam or not spam message, Madison or Hamilton as author, outdoor or indoor scene); regression models

interprets Kepler's conceptual shifts with a model of analogical reasoning is ([Gentner et al., 1997](#)).



predict a specific value of some measurement (given a description of a new movie, how much money will it make?); ordinal regression models predict values for non-metric measures (how much will you like the movie?). Experimental methods for hypothesis testing help develop and refine models of any type by systematically varying the conditions under which observations are made to discover how the results change in different situations.

A fundamental challenge in sensemaking and modeling is finding a balance between the competing goals of understanding a particular collection or dataset and being able to apply that understanding to new instances. Models can differ in the number of resource descriptions they use as parameters, and it is easy and tempting to overfit a model by using more parameters that capture random variations in observations. Overfitting produces spurious accuracy in reproducing the original observations, but it makes models less generalizable.

The highest level of sensemaking is the creation of scientific models or theories that propose interpretable and causal mechanisms for the observations. And just as automatic sensemaking creates simple explanations, scientists generally prefer simpler theories, a heuristic known as Occam's Razor, or the law of parsimony. Even though complex theories can sometimes be more accurate, simpler theories produce more testable predictions, making it easier to verify or refine the theory. Occam's famous principle, expressed eight centuries ago, is to prefer models that make the fewest

assumptions, often measured in terms of the number of parameters or variables needed to make a prediction.<sup>212223</sup>

21. The concept of [sensemaking](#) originated from business school research in management and organizational theory ([Weick 1995](#)) but has been widely employed by ethnographers in many contexts including emergency rooms, classrooms with minority students, airline safety inspections, and crime investigation. See ([Weick 2005](#)) and ([Chater 2016](#))
22. Occam's Razor has a long tradition in scientific philosophy, but some people have argued that it is overrated as a heuristic for choosing among alternative explanations or theories, particularly because it depends on how you define simplicity.
23. One way to make simplicity more useful as a guide for choosing between mathematical models is to explicitly penalize those that are more complex by adding error to the predictions, a technique that computer scientists have given the non-intuitive name of regularization. This penalty requires complex models to be significantly better at explaining the data than simpler ones because they have to overcome the added errors.

# Identifying Properties

Once the purposes of description have been established, we need to identify the specific properties of the resources that can satisfy those purposes. There are four reasons why this task is more difficult than it initially appears.

- First, any particular resource might need many resource descriptions, all of which relate to different properties, depending on the interactions to be supported and the context in which they take place. Selecting people for a basketball team focuses on their physical properties such as height, strength, leaping ability, and coordination. Selections for a debate team will be more concerned with their verbal and intellectual properties.
- Second, different types of resources need to incorporate different properties in their descriptions. For resources in a museum, these might include materials and dimensions of pieces of art; for files and services managed by a network administrator, these include access control permissions; for electronic books or DVDs, they would include the digital rights management (DRM) code that expresses what you can and cannot do with the resource.
- Third, as we briefly touched on in [“Scope, Scale, and Resource Description”](#), which properties participate in resource descriptions depends on who is doing the describing. It makes little sense to expect fine-grained distinctions and interpretations about properties from people who lack training in the discipline of organizing. We will return to this tradeoff in [“Creating Resource Descriptions”](#) and again in [“Describing Museum and Artistic Resources”](#).
- Fourth, what might seem to be the same property at a

conceptual level might be very different at an implementation level. Many resources have a resource description that is a surrogate or summary of the primary resource. For photos, paintings, and other resources whose appearance is their essence, an appropriate summary description can be a smaller, reduced resolution photo of the original. This surrogate is simple to create and easy for users to relate to the primary resource. On the other hand, distilling a text down to a short summary or abstract is a skill unto itself. Time-based resources provide greater challenges for summary. Should the summary of a movie be a textual summary of the plot, a significant clip from the movie, a video summary, or something else altogether?

This implementation gap is often very large for properties about people because people are not as easy to measure as most types of resources. Businesses need to quantify a person's interest in their products to predict what price they would be willing to pay, but "interest" cannot be measured directly. Instead, predictions rely on proxy measures for "interest" like how long the customer looked at the product web page and whether they also looked at a competitor's web page.

Two important dimensions for understanding and contrasting resource properties used in descriptions and organizing principles are: property essence—whether the properties are intrinsically or extrinsically associated with the resource, and; property persistence—whether the properties are static or dynamic. Taken together these two dimensions yield four categories of properties, as illustrated in [Figure: Property Essence x Persistence: Four Categories of Properties](#). These four categories provide a useful framework for thinking about resource properties, even if, at times, the classification of properties is debatable.<sup>24</sup>

24. For example, the composition of a chair is presented here

Property Essence x Persistence:  
Four Categories of Properties

as a static intrinsic property but in fact a wooden chair might deteriorate over time as a result of exposure to sunlight, heat, or biological agents that attack it. A skill can be considered intrinsic and dynamic, but it might also be highly dependent on context, making it extrinsic. The subject category assigned to a book is extrinsic and static, but if the classification system is revised the book might be reclassified. Finally, while the location of a resource can be extrinsic and dynamic, the location history of the resource at some specific point in time is a fact, an intrinsic and static property.

|                      |         | Property Essence                                                                                                                                                                                                                                                                      |                                                                                                                                                                                                                                                      |
|----------------------|---------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                      |         | Intrinsic                                                                                                                                                                                                                                                                             | Extrinsic                                                                                                                                                                                                                                            |
| Property Persistence | Static  | <p><b>Intrinsic Static</b></p> <p><b>Definition:</b> Directly experienced, subject matter, implicit, inherent properties.</p> <p><b>Examples:</b> Size, color, shape, author, date of creation.</p>  | <p><b>Extrinsic Static</b></p> <p><b>Definition:</b> Assigned to resource, name, identifier.</p> <p><b>Examples:</b> Dewey decimal</p> <p>704.00 → </p>             |
|                      | Dynamic | <p><b>Intrinsic Dynamic</b></p> <p><b>Definition:</b> Inherent properties; change over time.</p> <p><b>Examples:</b> Skills, experience</p>                                                          | <p><b>Extrinsic Dynamic</b></p> <p><b>Definition:</b> Behavioral and contextual properties</p> <p><b>Examples:</b> Current owner, location, best seller lists.</p>  |

The distinctions of property persistence and property essence combine to distinguish four categories of properties: intrinsic static, extrinsic static, intrinsic dynamic, and extrinsic dynamic properties.

### *Intrinsic Static Properties*

Intrinsic or implicit properties are inherent in the resource and can often be directly perceived or experienced. Static properties do not change their values over time. The species of an animal, the material of composition of a wooden chair, and the diameter of a wheel are all static properties that do not change their values over time. Static properties like color or shape are often used to describe and organize physical resources.

### Intrinsic Static Properties Define a Dalmatian

The spots on a Dalmatian dog are intrinsic static properties that appear shortly after birth, and they are so distinctive that it is impossible to describe the breed without acknowledging the spots.



This particular Dalmatian is the “greeter” at the Viader Winery in Deer Park, California. The dog is nice and the wines are excellent.

(Photo by R. Glushko.)

Intrinsic physical properties are usually just part of resource descriptions. In many cases, physical properties describe only the surface layer of a resource, revealing little about what something is or its original intended purpose, what it means, or when and why it was created. The author of a song and the context of its creation are other examples of intrinsic and static resource properties that are not directly perceivable.

Intrinsic descriptions are often extracted or calculated by computational processes. For example, a computer program might calculate the frequency and distribution of words in some particular document. Similarly, visual signatures or audio fingerprints are intrinsic descriptions ([“Describing Non-text Resources”](#)).

Some relationships among resources are intrinsic and static, like the parent-child relationship or the sibling relationship between two children with the same parents. Part-whole or compositional relationships for resources with parts are also intrinsic static properties often used in resource descriptions. However, it is better to avoid treating resource relationships as properties, and instead express them as relations. [Describing Relationships and Structures](#) discusses part-whole and other semantic relationships in great detail.

### *Extrinsic Static Properties*

Extrinsic or explicit properties are assigned to a resource rather than being inherent in it. The name or identifier of a resource is often arbitrary but once assigned does not usually change. Arranging resources according to the alphabetical or numerical order of their descriptive identifiers is a common organizing principle. Classification numbers and subject headings assigned to bibliographic resources are extrinsic static properties, as are the serial numbers stamped on or attached to manufactured products.

For information resources that have a digital form, the properties of their printed or rendered versions might not be intrinsic. Some text formats completely separate content from presentation, and as a result, style sheets can radically change the appearance of a printed document or web page without altering the primary resource in any way. For example, were a different style applied to this paragraph to



highlight it in bold or cast in 24-point font, its content would remain the same.

### *Intrinsic Dynamic Properties*

Intrinsic dynamic properties change over time. Developmental personal characteristics like a person's height and weight, skill proficiency, or intellectual capacity, for example. Because these properties are not static, they are usually employed only to organize resources whose membership in the collection is of limited duration. Sports programs or leagues that segregate participants by age or years of experience are using intrinsic dynamic properties to describe and organize the resources.

### *Extrinsic Dynamic Properties*

Extrinsic dynamic properties are in many ways arbitrary and can change because they are based on usage, behavior, or context. The current owner or location of a resource, its frequency of access, the joint frequency of access with other resources, its current popularity or cultural salience, or its competitive advantage over alternative resources are typical extrinsic and dynamic properties that are used in resource descriptions. A topical book described as a best seller one year might be found in the discount sales bin a few years later. A student's grade point average is an extrinsic dynamic property.

Why are Ottoman Carpets  
Named After a German Painter?



An example of a cultural category that has far outlasted its motivation is that of the Holbein carpet. A particular type of geometrically patterned Ottoman rug came to be known as a “Holbein carpet” after the German Renaissance painter Hans Holbein, who often depicted the rugs in his work (probably to show off his extremely meticulous technique). Holbein was famous in his time, and his commissioned paintings of the English King Henry VIII have Henry standing on such rugs. This painting, called “The Ambassadors,” was painted in 1533 and now hangs in The National Gallery, London.

(Source: Google Art Project)

Extrinsic dynamic properties are useful features for data scientists making prediction or classification models. Your current location, the thing you just bought, and the place you bought it can be viewed as manifestations of unobservable preferences and values. Fingerprints found on a doorknob at a crime scene are an extrinsic dynamic property associated with the door, and clever detectives would analyze them along with other interaction resources they discovered with the goal of identifying the person for whom the fingerprints are intrinsic static properties.

Many relationships between resources are extrinsic and dynamic properties, like that of best friend.

*Contextual properties* are those related to the situation or context in which a resource is described. Dey defines [context](#) as “any information that characterizes a situation related to the interactions between users, applications, and the surrounding environment.”<sup>25</sup> This open-ended definition implies a large number of contextual properties that might be used in a description; crisper definitions of context might be “location + activity” or “who, when, where, why.” Since context changes, context-based descriptors might be appropriate when assigned but can have limited persistence and effectivity ([“Resources over Time”](#)); the description of a document as “receipt of a recent purchase” will not be useful for very long.

25. ([Dey 2001](#)) further defines the “environment” of [context](#) as places, people, and things, and for each of “entities” there are four categories of context information: location, identity, status (or activity), and time. This [framework](#) thus yields 12 dimensions for describing the context of an environment.

### Latent Feature Creation and Netflix Recommendations

Recent advances in computing technology and [data science](#) techniques are making it possible to discover or create resource properties that are called “latent” because they are inferred rather than observed. Many such features are used by businesses to segment customers or make recommendations to them based on their recent behavior, so these features are also extrinsic and dynamic.

Your own movie preferences prove that easy to identify properties like sex and age do not differentiate movie watchers enough to make good recommendations, even if you combine them to create a category like “single male students between 18 and 25.” Netflix found that it was necessary to combine demographic properties, viewing history, and browsing behavior with very detailed ratings of dozens of movie properties to make good recommendations. It takes enormous computing power to discover a category of Netflix users who typically like action movies, yet consistently hover their mouse over romance movies, and to use this latent feature to recommend a sub-genre of western movies (one of nearly 100,000) that it calls “Romantic Action Adventure Movies.”<sup>26</sup>

26. A fascinating story about the Netflix's design and use of tens of thousands of movie sub-genres in its recommendation system is ([Madrigal 2014](#)).

Citations of one information resource by another are extrinsic static descriptions when they are in print form, but when they are published in digital libraries it is usually the case that “cited by” is a dynamic resource description. Similarly, any particular link from one web page to another is an extrinsic static description, but because many web pages themselves are highly dynamic, we can also consider links as dynamic as well. Citations and web links are discussed in more detail in [Describing Relationships and Structures](#).

Resources are often described with [cultural properties](#) that derive from conventional language or culture, often by analogy, because they can be highly evocative and memorable.<sup>27</sup>

Sometimes a cultural description outlives its salience, losing its power to evoke anything other than puzzlement about what it might mean.<sup>28</sup>

For the Lego boys, current with the latest Star Wars movies, “light

27. Consider how many events are named by appending a “-gate” suffix to imply that there is something scandalous or unethical going on that is being covered up. This cultural description is not immediately meaningful to anyone who does not know about the break-in at the headquarters of the Democratic National Committee headquarters at the Watergate hotel and subsequent cover-up that led to the 1974 resignation of US President Richard Nixon. A list of “-gate” events is maintained at [http://en.wikipedia.org/wiki/List\\_of\\_scandals\\_with\\_%22-gate%22\\_suffix](http://en.wikipedia.org/wiki/List_of_scandals_with_%22-gate%22_suffix).

28. ([http://en.wikipedia.org/wiki/Holbein\\_carpet](http://en.wikipedia.org/wiki/Holbein_carpet)).

saber” was just the obvious description for a long, neon tube with a handle. However, someone unfamiliar with the Star Wars franchise might not understand “light saber,” and would describe the piece some other way.

## Designing the Description Vocabulary

After we have determined the properties to use in resource descriptions, we need to design the description vocabulary: the set of words or values that represent the properties. [“Naming Resources”](#) discussed the problems of naming and proposed principles for good names, and since names are a very important resource description, much of what we said there applies generally to the design of the description vocabulary.

However, because the description vocabulary as a whole is much more than just the resource name, we need to propose additional principles or guidelines for this step. In addition, some new design questions arise when we consider all the resource descriptions as a set whose separate descriptions are created by many people over some period of time.

### *Principles of Good Description*

In [The Intellectual Foundation of Information Organization](#), Svenonius proposes a set of principles or “directives for design” of a description language.<sup>29</sup> Her principles, framed in the narrow

29. [\(Svenonius 2000, Ch. 5\)](#).

context of bibliographic descriptions, generally apply to the broad range of resource types we consider in this book.

### **User Convenience**

Choose description terms with the user in mind; these are likely to be terms in common usage among the target audience.

### **Representation**

Use descriptions that reflect how the resources describe themselves; assume that self-descriptions are accurate.

### **Sufficiency and Necessity**

Descriptions should have enough information to serve their purposes and not contain information that is not necessary for some purpose; this might imply excluding some aspects of self-descriptions that are insignificant.

### **Standardization**

Standardize descriptions to the extent practical, but also use aliasing to allow for commonly used terms.

### **Integration**

Prefer the same properties and terms for all types of resources.

Any set of general design principles faces two challenges.

- The first is that implementing any principle requires many additional and specific context-dependent choices for which the general principle offers little guidance. For example, how does the principle of Standardization apply if multiple standards already exist in some resource domain? Which of the competing standards should be adopted, and why?
- The second challenge is that the general principles can

sometimes lead to conflicting advice. The User Convenience recommendation to choose description terms in common use fails if the user community includes both ordinary people and scientists who use different terms for the same resources; whose “common usage” should prevail?

### *Who Uses the Descriptions?*

Focus on the user of the descriptions. This is a core idea that we cannot overemphasize because it is implicit in every step of the process of resource description. All of the design principles in the previous section share the idea that the design of the description vocabulary should focus on the user of the descriptions. Are the resources being organized personal ones, for personal and mostly private purposes? In that case, the description properties and terms can be highly personal or idiosyncratic and still follow the design principles.

Similarly, when resource users share relevant knowledge, or are in a context where they can communicate and negotiate, if necessary, to identify the resources, their resource descriptions can afford to be less precise and rigorous than they might otherwise need to be. This helps explain the curious descriptions in the Lego story with which we began this chapter. The boys playing with the blocks were talking to each other with the Legos in front of them. If they had not been able to see the blocks the others were talking about, or if they had to describe their toys to someone who had never played with Legos before, their descriptions would have been quite different.

More often, however, resource descriptions can not assume this degree of shared context and must be designed for user categories rather than individual users: library users searching for books, business employees or customers using part and product catalogs, scientists analyzing the datasets from experiments or simulations.



In each of these situations resource descriptions will need to be understood by people who did not create them, so the design of the description vocabulary needs to be more deliberate and systematic to ensure that its terms are unambiguous and sufficient to ensure reliable context-free interpretation. A single individual seldom has the breadth of domain knowledge and experience with users needed to devise a description vocabulary that can satisfy diverse users with diverse purposes. Instead, many people working together typically develop the required description vocabulary. We call the results institutional vocabularies, to contrast them with individual or cultural ones. (We will discuss this contrast more fully in [Categorization: Describing Resource Classes and Types](#))

#### Stop and Think: Description and Expertise

Everyone knows something about trees, but some people know more than others, and their particular experience and perspective influences how they describe trees. What kind of properties and descriptions would be used by university students? By research botanists? By landscape designers? By park maintenance workers? By indigenous people who live in tropical rain forests?

Some resource descriptions are designed for use by machines, which seemingly reduces the importance of design principles that consider user preferences or common uses. However, even if resources are described and organized by algorithms, when people need to explain the classifications and predictions that the algorithms produce, resource descriptions that are comprehensible and easily communicated are preferable to statistically optimal ones. Moreover, standardization and integration principles become more important for inter-machine communication to enable

efficient processing, reuse of data and software, and increased interoperability among organizing systems.<sup>30</sup>

### *Controlled Vocabularies and Content Rules*

As we defined in [“Use Controlled Vocabularies”](#), a [controlled vocabulary](#) is a fixed or closed set of description terms in some domain with precise definitions that is used instead of the vocabulary that people would otherwise use. For example, instead of the popular terms for descriptions of diseases or symptoms, medical researchers and teaching hospitals can use the National Library of Medicine’s *Medical Subject Headings*(MeSH) controlled vocabulary.<sup>31</sup>

We can distinguish a progression of vocabulary control: a glossary is a set of allowed terms; a thesaurus is a set of terms arranged in a hierarchy and annotated to indicate terms that are preferred, broader than, or narrower than other terms; an ontology expresses the conceptual relationships among the terms in a formal logic-based language so they can be processed by computers. We will say more about ontologies in [Describing Relationships and Structures](#).

*Content rules* are similar to controlled vocabularies because they also limit the possible values that can be used in descriptions. Instead of specifying a fixed set of values, content rules typically restrict descriptions by requiring them to be of a particular data type (integer, Boolean, Date, and so on). Possible values are constrained by logical expressions (e.g., a value must be between

30. [\(Laskey 2005\)](#).

31. <http://www.ncbi.nlm.nih.gov/mesh/>.

0 and 99) or [regular expressions](#) (e.g., must be a string of length 5 that must begin with a number). Content rules like these are used to ensure valid descriptions when people enter them in web forms or other applications.

### *Vocabulary Control as Dimensionality Reduction*

In most cases, a controlled vocabulary is a subset of the natural or uncontrolled vocabulary, but sometimes it is a new set of invented terms. This might sound odd until we consider that the goal of a controlled vocabulary is to reduce the number of descriptive terms assignable to a resource. Stated this way the problem is one of *dimensionality reduction*, transforming a high-dimensional space into a lower-dimensional one. Reducing the number of components in a multidimensional description can be accomplished by many different statistical techniques that go by names like “feature extraction,” “principle components analysis,” “orthogonal decomposition,” “latent semantic analysis,” “multidimensional scaling,” and “factor analysis.”<sup>32</sup>

32. We cannot cite all of mathematical statistics in one short endnote, but if you are inclined to learn more, ([Mardia, Kent, and Bibby 1980](#)) and ([Lee and Verleysen 2007](#)) are the kindest and gentlest resources. If we look very generously at “dimensionality reduction” we might even consider the indexing step of eliminating “stop words” to be a form of dimensionality reduction. Stop words appear with such high frequency that they have no discriminating power, so they are discarded from queries

These techniques might sound imposing and they are computationally complex, but they all have the same simple concept at their core, that the features or properties that describe some resource are often highly correlated. For example, a document that contains the word “car” is more likely to contain the words “driver” and “traffic” than a document that does not. Similar correlations exist among the visual features used to describe images and the acoustic features that describe music. Dimensionality reduction techniques analyze the correlations between resource descriptions to transform a large set of descriptions into a much smaller set of uncorrelated ones. In a way this implements the principle of Sufficiency and Necessity we mentioned in [“Principles of Good Description”](#) because it eliminates description dimensions or properties that do not contribute much to distinguishing the resources.

Here is an oversimplified example that illustrates the idea. Suppose we have a collection of resources, and every resource described as “big” is also described as “red,” and every “small” resource is also “green.” This perfect correlation between color and size means that either of these properties is sufficient to distinguish “big red” things from “small green” ones, and we do not need clever algorithms to figure that out. But if we have thousands of properties and the correlations are only partial, we need the sophisticated statistical approaches to choose the optimal set of description properties and terms, and in some techniques the dimensions that remain are called “latent” or “synthetic” ones because they are statistically optimal but do not map directly to resource properties.

and not part of the description of the indexed documents.

## Designing the Description Form

By this step in the process of resource description we have made numerous important decisions about which resources to describe, the purposes for which we are describing, them, and the properties and terms we will use in the descriptions. As much as possible we have described the steps at a conceptual level and postponed discussion of implementation considerations about the notation, syntax, and deployment of the resource descriptions separately or in packages. Separating design from implementation concerns is an idealization of the process of resource description, but is easier to learn and think about resource description and organizing systems if we do. We discuss these implementation issues in [\*The Forms of Resource Descriptions\*](#).

Sometimes we have to confront legacy technology, existing or potential business relationships, regulations, standards conformance, performance requirements, or other factors that have implications for how resource descriptions must or should be implemented, stored, and managed. We will take this more pragmatic perspective in [\*The Organizing System Roadmap\*](#), [\*The Organizing System Roadmap\*](#), but until then, we will continue to focus on design issues and defer discussion of the implementation choices.

## Creating Resource Descriptions

Resource descriptions can be created by professionals, by the authors or creators of resources, by users, or by computational or automated means.

From the traditional perspective of library and information science with its emphasis on bibliographic description, these modes of

creation imply different levels of description complexity and sophistication; Taylor and Joudrey suggest that professionals create *rich* descriptions, untrained users at best create *structured* ones, and automated processes create *simple* ones.

This classification reflects a disciplinary and historical bias more than reality. “Simple” resource descriptions are “no more than data extracted from the resource itself... the search engine approach to organizing the web through automated indexing techniques.”<sup>33</sup>

It might be fair to describe an inverted index implementation of a Boolean information retrieval model as simple, but it is clearly wrong to consider what Google and other search engines do to describe and retrieve web resources as simple.<sup>34</sup>

A better notion of levels of resource description is one based on the amount of interpretation imposed by the description, an approach that focuses on the descriptions themselves rather than on their methods of creation. We will discuss this sort of approach in [“Describing Museum and Artistic Resources”](#) in the context of describing museum and artistic resources.

Professionally-created resource descriptions, author- or user-created descriptions, and computational or automated descriptions each have strengths and limitations that impose tradeoffs. A natural solution is to try to combine desirable aspects from each in hybrid

33. [\(Taylor and Joudrey 2009\)](#), p. 91.

34. See Ch. 4 of [\(Buttcher, Clarke, and Cormack 2010\)](#) for a description of a simple Boolean information retrieval model and Chapter 14 and 15 for descriptions of Google-scale ones. For a popular discussion of the Google algorithm see [\(Levy 2010\)](#).

approaches. For example, the vocabulary for a new resource domain may arise from tagging by end users but then be refined by professionals, lay classifiers may create descriptions with help from software tools that suggest possible terms, or software that creates descriptions can be improved by training it with human-generated descriptions, a form of [supervised learning](#) (see [“Categories Created by Clustering”](#)).

Often existing resource descriptions can or must be transformed or enhanced to meet the ongoing needs of an organizing system, and sometimes these processes can be automated. We will defer further discussion of those situations to [Interactions with Resources](#). In the discussion that follows we focus on the creation of new resource descriptions where none yet exist.

### *Resource Description by Professionals*

Before the web made it possible for almost anyone to create, publish, and describe their own resources and to describe those created and published by others, resource description was generally done by professionals in institutional contexts. Professional indexers and catalogers described bibliographic and museum resources after having been trained to learn the concepts, controlled descriptive vocabularies, and the relevant standards. In information systems domains professional data and process analysts, technical writers, and others created similarly rigorous descriptions after receiving analogous training. We have called these types of resource descriptions institutional ones to highlight the contrast between those created according to standards and

those created informally in *ad hoc* ways, especially by untrained or undisciplined individuals.<sup>35</sup>

35. Many institutional organizing systems are subject to a single centralized or governmental authority that can impose principles for describing and arranging resources. Examples of organizing systems where resources are described using standard centralized principles are:

Libraries that use national bibliographic standards to satisfy requirements set by industry associations or other accreditation bodies such as the Association of College and Research Libraries (ACRL). (<http://ala.org/acrl>)

Companies that follow industry standards for information or process models, product classification or identification to be eligible for government business ([Shah and Kesan 2006](#)).

Legislative documents that conform to National or European Community standards for structure, naming, and description ([Biasiotti 2008](#)).

The Internet Corporation for Assigned Names and Numbers(ICANN) and its policies for operating the Domain Name System(DNS) make it possible for every website to be located using its logical name (like “berkeley.edu” rather than using an IP address like 169.229.131.81). (<http://www.icann.org/>)



## *Resource Description by Authors or Creators*

The author or creator of a resource can be presumed to understand the reasons why and the purposes for which the resource can be used. And, presumably, most authors want to be read, so they will describe their resources in ways that will appeal to and be useful to their intended users. However, these descriptions are unlikely

In other domains multiple organizations or institutions have the authority to impose principles of resource description. Sometimes this authority derives from the voluntary collaboration of multiple autonomous parties who set and conform to standards because they benefit from being able to share resources or information about resources. Examples of organizing systems where resources are described using standardized decentralized principles are:

Firms that establish company-wide standards for their information resources, typically including the organization and management of source content, document type models, and a style guide that applies to print and web documents.

Firms that participate in the OASIS (<http://www.oasis-open.org/>) or the W3C (<http://www.w3.org/>) industry consortia to establish specifications or technical recommendations for their information systems or web services).

to use the controlled vocabularies and standards that professional catalogers would use.

### *Resource Description by Users*

Today's web contains a staggering number of resources, most of which are primary information resources published as web content, but many others are resources that stand for "in the world" physical resources. Most of these resources are being described by their users rather than by professionals or by their authors. These "at large" users are most often creating descriptions for their own benefit when they assign tags or ratings to web resources, and they are unlikely to use standard or controlled descriptors when they do so.<sup>36</sup> The resulting variability can be a problem if creating the description requires judgment on the tagger's part. Most people can agree on the length of a particular music file but they may differ wildly when it comes to determining to which musical genre that file belongs. Fortunately most web users implicitly recognize that the potential value in these "Web 2.0" or "user-generated content" applications will be greater if they avoid egocentric descriptions. In addition, the statistics of large sample sizes inevitably leads to some agreement in descriptions on the most popular applications

36. Many organizing systems describe and arrange their physical or information resources in *ad hoc* ways because the person or institution determining the arrangement is completely autonomous. This is the domain of organizing systems embraced by David Weinberger in [\*Everything is Miscellaneous\* \(Weinberger 2007\)](#).

because idiosyncratic descriptions are dominated in the frequency distribution by the more conventional ones.<sup>37</sup>

We are not suggesting that professional descriptions are always of high quality and utility, and socially produced ones are always of low quality and utility.<sup>38</sup> Rather, it is important to understand the limitations and qualifications of descriptions produced in each way. Tagging lowers the barrier to entry for description, making organizing more accessible and creating descriptions that reflects a variety of viewpoints. However, when many tags are associated with a resource, it increases [recall](#) while decreasing [precision](#). (See [“Resource Description by Users”](#))

37. ([Sen et al. 2006](#)) analyze the effects of four tag selection algorithms used in sites that allow user tags on vocabulary evolution (more often called “tag convergence” in the literature), tag utility, tag adoption, and user satisfaction.

38. But in an often-cited essay ([Doctorow, 2001](#)) provocatively titled [Metacrap: Putting the torch to seven straw-men of the meta-utopia](#), Cory Doctorow argues that much human-created metadata is of low quality because “people lie, people are lazy, people are stupid, mission impossible—know thyself, schemas are not neutral, metrics influence results, (and) there is more than one way to describe something.”

## *Automated and Computational Resource Description*

A picture's EXIF file created by a digital camera records properties associated with the camera and its settings, as well as some properties of the photo-taking context. (See [Figure: Contrasting Descriptions for a Work of Art](#), for an example.) Creating this highly detailed description by hand would be nearly impossible. The downside, however, is that the automated description does not capture the meaning of the photo; an automated picture description captures the time and place, but not that it is a picture of a honeymoon vacation. The difference between automated and human description is called the *semantic gap* ("[The Semantic Gap](#)").

Any resource that is smart enough to collect data about its state or environment is creating resource descriptions automatically (See "[Active or Operant Resources](#)"). Resources with computational capabilities can process the raw sensor data to identify important events and create more interpretable descriptions.

Some computational approaches create resource descriptions that are similar in purpose to those created by human describers. Text mining and summarization systems for customer comments about products can reduce thousands of comments to a list of the most important features.<sup>39</sup> People shopping for books at Amazon.com get insights about a book's content and distinctiveness from the statistically improbable phrases that it has identified by comparing all the books for which it has the complete text.<sup>40</sup>

Computational descriptions can use any observable or latent

39. ([Hu and Lui 2004](#)).

40. <http://www.amazon.com/gp/search-inside/sipshelp.html/>.

variable (see the sidebar, [Latent Feature Creation and Netflix Recommendations](#)) except some that are prohibited by law, such as race, religion, national origin, and marital status, to prevent discrimination. In practice, however, this prohibition is easily circumvented because these properties can usually be predicted using other ones. For example, race can often be reliably predicted using residence address and surname.<sup>41</sup>

Of course, all information retrieval systems compare a description of a user's needs with descriptions of the resources that might satisfy them. IR systems differ in the resource properties they emphasize; word frequencies and distributions for documents in digital libraries, links and navigation behavior for web pages, acoustics for music, and so on. These different property descriptions determine the comparison algorithms and the way in which relevance or similarity of descriptions is determined. We say a lot more about this in ["Describing Non-text Resources"](#) and in [Interactions with Resources](#).

## Evaluating Resource Descriptions

Evaluation is implicit in many of the activities of organizing systems we described in [Activities in Organizing Systems](#) and is explicit when we maintain a collection of resources over time. In this section, we focus on the narrower problem of evaluating resource descriptions.

41. The title says it all: [Predictive analytics: The power to predict who will click, buy, lie, or die. \(Siegel 2013\)](#). The Bayesian Surname and Geocoding technique for predicting race is described by [\(Elliott et al. 2008\)](#).

Evaluating means determining [quality](#) with respect to some criteria or dimensions. Many different sets of criteria have been proposed; for repositories of digital resources, the most commonly used ones are accuracy, completeness, and consistency.<sup>42</sup> Other typical criteria are timeliness, interoperability, and usability. It is easy to imagine these criteria in conflict; efforts to achieve accuracy and completeness might jeopardize timeliness; enforcing consistency might preclude modifications and personalizations that would enhance usability.

#### Stop and Think: Defining Quality

What characteristics or criteria would you use to determine the quality of a car? Of food? Of clothing? Of a place to live? Which of these criteria are domain-specific, and which ones apply more generally to many types of resources?

The [quality](#) of the outcome of the multi-step process proposed in this chapter is a composite of the [quality](#) created or squandered at each step. A scope that is too granular or abstract, overly ambitious or vague intended purposes, a description vocabulary that is hard to use, or giving people inadequate time to create good descriptions can all cause quality problems, but none of these decisions is visible at the end of the process where users interact with resource descriptions.

42. ([Park 2009](#)).

## *Evaluating the Creation of Resource Descriptions*

When professionals create resource descriptions in a centralized manner, which has long been the standard practice for many resources in libraries, there is a natural focus on [quality](#) at the point of creation to ensure that the appropriate controlled vocabularies and standards have been used. However, the need for resource description generalizes to resource domains outside of the traditional bibliographic one, and other [quality](#) considerations emerge in those contexts.

Resource descriptions in private sector firms are essential to running the business and in interacting efficiently with suppliers, partners, and customers. Compared to the public sector, there is much greater emphasis on the economics and strategy of resource description.<sup>43</sup> What is the value of resource description? Who will bear the costs of producing them? Which of the competing industry standards will be followed? Some of these decisions are not free choices as much as they are constraints imposed as a condition of doing business with a dominant economic partner, which is sometimes a governmental entity.

For example, a firm like Wal-Mart with enormous market power can dictate terms and standards to its suppliers because the long-term benefits of a Wal-Mart contract usually make the initial accommodation worthwhile. Likewise, governments often require

43. However, these concerns are rapidly becoming more important in the public sector. In particular, many public universities in the US are struggling with cuts in state and federal funding that are affecting library services and practices.

their suppliers to conform to open standards to avoid lock-in to proprietary technologies.<sup>44</sup>

In both the public and private sectors there is increased use of computational techniques for creating resource descriptions because the number of resources to be described is simply too great to allow for professional description. A great deal of work in text data mining, web page classification, semantic enrichment, and other similar research areas is already under way and is significantly lowering the cost of producing useful resource descriptions. Some museums have embraced approaches that automatically create user-oriented resource descriptions and new user interfaces for searching and browsing by transforming the professional descriptions in their internal collections management systems.<sup>45</sup> Google's ambitious project to digitize millions of books has been criticized for the quality of its algorithmically extracted resource descriptions, but we can expect that computer scientists will put

44. More generally, economists use the concept of the “mode of exchange” in a business relationship to include the procedures and norms that govern routine behavior between business partners. An “exit” mode is one in which the buyer makes little long-term commitment to a supplier, and problems with a supplier cause the buyer to find a new one. In contrast, in “voice” mode there is much greater commitment and communication between the parties, usually leading to improved processes and designs. See [\(Helper and McDuffie 2003\)](#).

45. [\(Schmitz and Black 2008\)](#).



the Google book corpus to good use as a research test bed to improve the techniques.<sup>46</sup>

Web 2.0 applications that derive their value from the aggregation and interpretation of user-generated content can be viewed as voluntarily ceding their authority to describe and organize resources to their users, who then tag or rate them as they see fit. In this context the consistency of resource description, or the lack of it, becomes an important issue, and many sites are using technology or incentives to guide users to create better descriptions.

### *Evaluating the Use of Resource Descriptions*

Regardless of, or in addition to, any quality criteria applied to the creation and selection of resource descriptions, at some point the resource descriptions meet their intended users. The most important quality criterion at that point is whether the resource descriptions satisfy their intended purposes in a usable way. In many ways, the answer is a disappointing no.

For example, in one of the earliest revisions to the original [HTML](#) specification, a <META> tag was added to allow creators of web resources to define a set of key terms to describe a website or web page. This well-motivated resource description was to be used by search engines to improve the relevance of retrieved pages. However, it soon became obvious that it was possible to “game” the META tag by adding popular terms even though they did not

46. [\(Nunberg 2009\)](#) called the quality of Google’s metadata “a disaster for scholars,” but [\(Sag 2012\)](#) argues that the otherwise neglected “orphan works” in the Google corpus are “grist for the data mill.”

accurately describe the page. Today search engines ignore the <META> tag for ranking pages, but many other techniques that use false [resource](#) descriptions continue to plague web users. (See [“Social and Web Curation”](#).)

The design of a description vocabulary circumscribes what can be said about a resource, so it is important to recognize that it implicitly determines what cannot be said as well, with unintended negative consequences for users. The [resource](#) description schema implemented in a physician’s patient management system defines certain types of recordable information about a patient’s visit—the date of the visit, any tests that were ordered, a diagnosis that was made, a referral to a specialist. The schema, and its associated workflow, impose constraints that affect the kinds of information medical professionals can record and the amount of space they can use for those descriptions. Moreover, such a schema might also eliminate vital unstructured space that paper records can provide, where doctors communicate their rationale for a diagnosis or decision without having to fit it into any particular box.

However, when resource descriptions are the data used to train models for prediction or classification, the focus of evaluation is not on the descriptions, which are often assumed to be accurate observations about the world. Instead, evaluation focuses on the model, and “model selection” is the task of choosing which of several competing models best fit the original data while also generalizing well to new data. In any event, any quality problems or selection biases with the original data will undermine the value of whatever model is selected.

### *The Importance of Iterative Evaluation*

The inevitable conflicts between [quality](#) goals mean that there will be compromises among the [quality](#) criteria. Furthermore,

increasing [scale](#) in an organizing system and the steady improvements of computational techniques for resource description imply that the nature of the compromise will change over time. As a result, a single evaluation of resource descriptions at one moment in time will not suffice.

This makes usage records, navigation history, and transactional data extremely important kinds of resource descriptions because they enable you to focus efforts on improving [quality](#) where they are most needed. Furthermore, for organizing systems with many types of resources and user communities, this information can enable the tailoring of the nature and extent of resource description to find the right balance between “rich and comprehensive” and “simple and efficient” approaches. Each combination of resource type and user community might have a different solution.

The idea that [quality](#) is a property of an end-to-end process is embodied in the “quality movement” and statistical process control for industrial processes but it applies equally well to resource description. The central idea is that [quality](#) cannot be tested in by inspecting the final products. Instead, [quality](#) is achieved through process control—measuring and removing the variability of every process needed to create the products.<sup>47</sup> Explicit feedback from users or implicit feedback from the records of their resource interactions needs are essential as we iterate through the design process and revisit the decisions made there.

47. The modern “quality movement” grew out of the efforts of the US to rebuild Japan after the Second World War and its “Bible” was Juran’s 1951 [Quality Control Handbook \(Juran 1951\)](#).

## 34. Describing Non-text Resources

Many of the principles and methods for resource description were developed for describing text resources in physical formats. Those principles have had to evolve to deal with different types of resources that people want to describe and organize, from paintings and statues to MP3s, JPEGs, and MPEGs.

Some descriptions for non-text resources are text-based, and are most often assigned by people. Other descriptions are in non-text formats are extracted algorithmically from the content of the non-text resource. These latter content-based resource descriptions capture intrinsic technical properties and in some domains are able to describe [aboutness](#) with some accuracy, thanks to breakthroughs in machine learning.

### Describing Museum and Artistic Resources

The problems associated with describing multimedia resources are not all new. Museum curators have been grappling with them since they first started to collect, store, and describe artifacts hundreds of years ago. Many artifacts may represent the same work (think about shards of pottery that may once have been part of the same vase). The materials and forms do not convey semantics on their own. Without additional research and description, we know nothing about the vase; it does not come with any sort of title page or tag that connects it with a 9th-century Mayan settlement. Since museums can acquire large batches of artifacts all at once, they have to make decisions about which resources they can afford to describe and how much they can describe them.

German art historian Erwin Panofsky first codified one approach to these problems of description. In his classic [Studies in Iconology](#), he defined three levels of description that can be applied to an artistic work or museum artifact. [Figure: Contrasting Descriptions for a Work of Art](#) contrasts these three levels in the descriptions of a marble statue. It also shows the striking differences between the EXIF description in a digital photo of the statue and those created by people.<sup>1</sup>

1. [\(Panofsky 1972\)](#) proposes these three levels of description:

#### **Primary subject matter**

At this level, we describe the most basic elements of a work in a generic way that would be recognizable by anyone regardless of expertise or training. The painting *The Last Supper*, for example, might be described as “13 people having dinner.”

#### **Secondary subject matter**

Here, we introduce a level of basic cultural understanding into a description. Someone familiar with a common interpretation of the Bible, for example, could now see *The Last Supper* as representing Jesus surrounded by his disciples.

#### **Intrinsic meaning or interpretation**

At this level, context and deeper understanding come into play—including what the creator of the

Professionals who create descriptions of museum and artistic resources, architecture and other cultural works typically use the VRA Core from the Library of Congress, or the Getty Trust *Categories for the Description of Works of Art*(CDWA), a massive controlled vocabulary with 532 categories and subcategories. A CDWA-Lite has been developed to create a very small subset for use by non-specialists.<sup>2</sup>

### Contrasting Descriptions for a Work of Art

description knows about the situation in which the work was created. Why, for example, did this particular artist create this particular depiction of The Last Supper in this way? Panofsky posited that professional art historians are needed here, because they are the ones with the education and background necessary to draw meaning from a work.

In other words, Panofsky saw the need for many different types of descriptors—including physical, cultural, and contextual –to work together when making a full description of an artifact.

2. For CDWA, see ([Harpring2009](#)) at [http://www.getty.edu/research/publications/electronic\\_publications/cdwa/](http://www.getty.edu/research/publications/electronic_publications/cdwa/).

For CDWA-Lite, see ([Getty2006](#)) at [http://www.getty.edu/research/publications/electronic\\_publications/cdwa/cdwalite.pdf](http://www.getty.edu/research/publications/electronic_publications/cdwa/cdwalite.pdf).



### EXIF Summary

|                        |                                                     |
|------------------------|-----------------------------------------------------|
| <b>Make</b>            | NIKON CORPORATION                                   |
| <b>Model</b>           | NIKON D90                                           |
| <b>Aperture</b>        | 9                                                   |
| <b>Exposure Time</b>   | 1/320 (0.003125 sec)                                |
| <b>Lens</b>            | ID AF-S DX VR Zoom-Nikkor<br>18-105mm f/3.5-5.6G ED |
| <b>Focal Length</b>    | 21.0 mm                                             |
| <b>Flash</b>           | Auto, Did not fire                                  |
| <b>File Size</b>       | 4.7 MB                                              |
| <b>File Type</b>       | JPEG                                                |
| <b>Image Height</b>    | 4288                                                |
| <b>Image Width</b>     | 2848                                                |
| <b>Date &amp; Time</b> | 2012:12:03 10:31:14                                 |

### 3 Levels

#### Primary

Marble statue of nude woman standing on a seashell.

#### Secondary

Statue made in 2005 by Lucio Carusi of Carrara, Italy, titled "Venus", made of local marble.

#### Interpretive

This is a 3d transformation of the 1486 painting by Italian painter Sandro Botticelli, titled "The Birth of Venus", now in the Uffizi Gallery in Florence. Carusi's Venus is substantially slimmer in proportions than Botticelli's because of changing notions of female beauty.

Descriptions for works of art can contrast a great deal, especially between those captured by a device like a digital camera and those created by people. Furthermore, the descriptions created by people differ according to the expertise of the creator and the amount of subjective interpretation applied in the description.

(Photo by R. Glushko. The statue, titled "Venus," was made by Lucio Carusi, of Carrara, Italy, and is currently part of a private collection.)

## Describing Images

Digital cameras, including those in cell phones, take millions of photos each day. Unlike the images in museums and galleries, most of these images receive few descriptions beyond those created by the device that made them. Nevertheless, a great many of them end up with some limited descriptions in Facebook, Instagram, Flickr, Picasa, DeviantArt, or others of the numerous places where people share images, or in professional image applications like Light Room. All of these sites provide some facilities for users to assign tags to images or arrange them in named groups.

Many different computational approaches have been used to describe or classify images. One approach uses the visual signature of an image extracted from low-level features like color, shape, texture, and luminosity, which are then used to distinguish significant regions and objects. Image similarity is computed to create categories of images that contain the same kinds of colors, objects, or settings, which makes it easy to find duplicate or modified images.<sup>3</sup>

For computers to identify specific objects or people in images, it is logically necessary to train them with images that are already identified. In 2005 Luis van Ahn devised a clever way to collect large amounts of labeled images with a web-based game called ESP that randomly paired people to suggest labels or tags for an image. The obvious choices were removed from contention, so a photo of a bird

3. See ([Datta et al. 2008](#)). The company Idée is developing a variety of image search algorithms, which use image signatures and measures of visual similarity to return photos similar to those a user asks to see.