FRBR, Before and After

ALA Editions purchases fund advocacy, awareness, and accreditation programs for library professionals worldwide.

FRBR

BEFORE AND AFTER

A Look at Our Bibliographic Models

KAREN COYLE

ala editions

An imprint of the American Library Association
Chicago 2016

KAREN COYLE is a librarian with over 30 years' experience with library technology, who serves as consultant on a variety of issues relating to digital libraries. She has published dozens of articles and reports, many of which are available at kcoyle.net. She has served on several standards committees, including the MARC standards group (MARBI) and the NISO committee AX for the OpenURL standard, and was an ALA representative on the e-book standards development team that contributed to the ePub standard. She writes and speaks on a wide range of policy areas, including intellectual property, privacy, and public access to information. Her January 2010 issue of Library Technology Reports, "Understanding the Semantic Web: Bibliographic Data and Metadata," was awarded the 2011 ALCTS Outstanding Publication Award.

ISBNs
978-0-8389-1345-1 (paper)
978-0-8389-1364-2 (PDF)
978-0-8389-1365-9 (ePub)
978-0-8389-1366-6 (Kindle)

Library of Congress Cataloging-in-Publication Data

Coyle, Karen, author.
FRBR, before and after : a look at our bibliographic models / Karen Coyle.
pages cm
Includes bibliographical references and index.
ISBN 978-0-8389-1345-1 (print : alk. paper) — ISBN 978-0-8389-1365-9 (epub) — ISBN 978-0-8389-1364-2 (pdf) — ISBN 978-0-8389-1366-6 (kindle) 1. FRBR (Conceptual model) I. Title.
Z666.6.C69 2016
025.3'2—dc23 2015020100

Cover design by Kimberly Thornton. Image © Shutterstock, Inc.
Design composition by Alejandra Diaz in the ITC Galliard Std, Gotham, and Miso typefaces.

♾ This paper meets the requirements of ANSI/NISO Z39.48–1992 (Permanence of Paper).

Printed in the United States of America
20 19 18 17 16 5 4 3 2 1

CONTENTS

FIGURES

ACKNOWLEDGMENTS

I want to thank Patrick Hogan of ALA Publishing for his willingness to experiment with an open access model for this book. A digital version is available at my web site, http://kcoyle.net. I hope, however, that many of you will find the convenience of hard copy worthwhile so that ALA Publishing will be encouraged to do this for other works in the future.

I thank Lynne Cameron for reading through the manuscript and making many highly valuable suggestions. (All remaining errors of fact are, of course, mine.) My conversations and collaboration with Tom Baker have been instrumental to my understanding of FRBR and its relation to RDF. This book began as a series of sketchy and perhaps poorly worded online postings; these brave souls read and commented on that: Kevin M. Randall, Lukas Koster, Heidrun Wiesenmüller, Amanda Cossham, and Joel Kalvesmaki. I hope I did not miss anyone.

Finally, I give thanks to Dewey the Decimal Dog, otherwise known as Dewey 636.7, whose job it was to get me out of my chair and away from the computer at regular intervals. There's nothing like an enthusiastic golden retriever with a tennis ball in his mouth to remind you that there is more to life than FRBR.

INTRODUCTION

Go to your bookshelf and pull off a book; any book. It may be one you have read many times, or it could be one that is still on your "to read" list. Take a look at it. It may be bound with the flimsy cardboard of a paperback sporting a slick, shiny cover. Or the pages could be held between the cloth-covered boards of a quality hardback. It is unlikely, however, unless you are either very wealthy or very lucky, that your fingers will be touching a fine leather binding.

It is probable that you did not purchase the book for its physical appearance, as appealing as that may be, but for its content. That's where things get complicated in our story: complicated because it is very hard to say what the content consists of. Words, yes, but you didn't buy just a set of words, unless your book is a dictionary. No, you bought this book for the story it tells or for the information it imparts. You may have been seeking entertainment, or to learn something new (and happy is the person who gets both!). Although the story or the information came to you as words, you may not be able to recite even a small passage verbatim. We read the words but we remember the meaning, another concept that is difficult to define.

If I ask you some questions about the book, some will be easy to answer, some more difficult. I could ask you for the title, and most likely you know that. The same for the author. You could surely tell me what the book is about, either with a topic ("it's a history of the Venetian Republic") or a story ("it's about a girl who lives on the prairie and what she and her family go through to survive"). Chances are, though, that if I ask you who published the book, you'll be taking a sneak peek at the title page or the spine to find that information. The place and date of publication will not only be less imprinted in your consciousness but they may actually be a bit hard to find. The precise number of pages is another undeniable fact about the book that may not be on the tip of your tongue.

As a reader, it is the reading experience and what it leaves behind in your memory that makes up the inherent value of the book. And we do know that readers value their books. There would be no other reason to use the bulk of the wall space of one's home for shelves for book storage, or, when moving to a new home, to pack, lug, and unpack untold pounds of what appears to be inert tree pulp.

Now let's leave books behind and look at other media. Just as many of us love our books, we also have among us many music lovers; people with towering racks of CDs or digital devices chock full of tunes. Here, though, we find some differences from our book story. Ask a music lover the "author" (composer) of a tune and you may be asking the obvious ("Beethoven's 5th symphony") or not ("Santa Claus Is Coming to Town"), even though both pieces of music are easily recognizable when heard. They are recognizable also because, unlike books, we listen to the same piece of music many times, and in different versions. This is a function of the fact that music is performed. Some performances are faithful interpretations of the music, and others, like jazz or digital sampling, are creative distortions of the original.

Music lovers with sufficient talent can reproduce a version of the music either by humming, singing, or playing the music on an instrument. We remember the notes of music in a way that we do not remember the words of a book. But if asked what the music is "about" we are in some difficulty in most cases. Unless the music has a specific story attached to it, such as Sergei Prokofiev's *Peter and the Wolf*, or the teen drama of "Dead Man's Curve," much music does not have a plot or a message that translates to "aboutness."

Other information that only dedicated aficionados of a music genre can relate about their listening choices are date of recording; names of all performers; date of composition; number and types of instruments. Asked what type of music we like, the answers are broad categories like rock, jazz, classical, or country; or sometimes a more specific category, still covering a wide swath: heavy metal; mostly Mozart; Irish folk music; Reggae.

Books and music are two common creative forms that many of us encounter in our everyday lives, and yet what we know about them and how we interact with them are quite different. Now let's look at another creative form: computer games. A player will know the name of the game, the general plot of the play (capture castle, defeat enemy, solve puzzle), and the names of characters. She will also know what capabilities she has as a player (running, jumping, opening doors). If it is a multi-player game, she will know the names of other players—that is, the names they are using in the game. She may not, however, be able to respond to the question "who wrote or created the game?" Games often do not have measurable durations although some have ending points, so asking "how long is it?" may not make sense.

With a movie, on the other hand, the running time for the film is a key element and moviegoers, unless they walk out in the middle of the film, will experience that actual duration. Movies have directors and producers, screenwriters, and hundreds of other participants from makeup artists to caterers. Some directors are famous, but what makes movies *The Movies!* are the stars: the people that you see on the screen. Having seen the film, most people will be able to relate the story and the names of the primary actors. Very few will remember the producer, although his name will have appeared briefly in big letters at the beginning of the film, and even fewer will have noted the screenwriter's name. The name of the studio that produced the film, analogous to the publisher of a book, is rarely noticed.

All of these above-mentioned creative forms are ones to which its users or participants have a certain emotional attachment. There are other kinds of created resources that we seek out but that are less enticing. I'm thinking of items like census figures, standards documents, technical reports, or court proceedings. If asked about authors of these materials, few people outside of librarianship would name courts or government as authors, although they might see them as responsible bodies of some kind. Users of these materials, however, may be keenly aware of the version of the material; a 1950 census is obviously not the same as a year 2000 census, and a version 0.7 of a standards document would be expected to differ from the 2.0 version. Having the latest version may be essential for some functions, although comparisons of figures across time make use of different versions of the data. Knowing that the copy that you have is authentic and has not been altered is another consideration for these materials. For, like census or economic data, a key factor is whether it is formatted for possible number-crunching.

The point of this brief walk through the various resource types is this: given how different these resources are, and how different our relationship to them is, making any general statement about the structure or data elements needed to describe all resources for all users of a library catalog is going to be difficult, if not impossible. And yet, that is exactly what we do on a routine basis: we create records that treat all resource types the same, and for only one definition of "user." We also ignore or downplay many of the characteristics that are important for users. We often place the names of film actors, when we provide them at all, in a note field that is barely searchable. We also give technical information about data sets and computer files in a note. We give book readers a place of publication and a number of pages but don't give them a clue to the story that the book holds. ("Mentally ill—Fiction" is a subject heading on *Moby Dick*.)

All of this is to point out how varied is our bibliographic universe, and this is without having looked at the differences among users: from novices to experts, children and adults, beach readers and researchers.

Quite clearly, in terms of bibliographic services, one size cannot possibly fit all.

This illustrates the difficulty we have in defining the fundamental nature of the bibliographic "thing," often called a "work." And it also illustrates that the users are an element in that definition. It provides an argument for a flexible treatment that can accommodate a range of user approaches and needs, perhaps a modular structure that can be modified to place emphasis on different information for different materials and different users. Why shouldn't a search on an author return information about the author, including the author's works? Where was the author born, when did she live, what is she known for? In library catalogs, there is no differentiation between Edgar Allen Poe and Barbara Cartland. This isn't neutrality, it's a lack of information. If an item is retrieved on title, there is clearly more that could be said about it than where and when that particular exemplar was published. We present a copy of Charles Darwin's *On the Origin of Species* with a publication date of 2003 without any further explanation, neither of the importance of the work, nor its own true origins. *On the Origin of Species* is meaningful only if you know what scientific thinking was before Darwin's discovery, and that this book is the beginning text for the entire science of evolution.

All of this is possible, but only if we can make some fundamental changes in our approach to bibliographic description. A new approach presupposes a redefining of bibliographic description from a fixed, immovable block of data to a set of interrelated information units that can be viewed from different vantage points.

The challenge for us lies in transforming what we can of our data into interrelated "things" without overindulging that metaphor. There are indeed things of interest to be defined for cultural heritage and creative objects, but our universe of operation lacks the precision of, for example, financial data, where every point of information is precisely known, or the calculation of tensile strength in the engineering task of bridge building. What we describe is not easily subject to quantitative testing, and the difference between success and failure is hard to measure. We are fortunate that errors in library catalogs rarely result in death of the user, but we are hindered by a lack of knowledge of our effect on learning and culture. In spite of the attempts in the 1960s to convince the world that one could add the word *science* to *library* and gain a modicum of status, describing information resources remains an art.

We do have some cold, hard facts in our data storehouse, but we also have some squishy bits—some areas where we simply cannot achieve the level of precision enjoyed by science and engineering. Part of the reason for our imprecision is the durability of our inventory. Unlike a warehouse of electronic gadgets, we don't discard last year's product when the latest offerings arrive. Some of us even keep the old, the ragged, and the unused materials. Our material lacks uniformity: we have books without authors, articles with citations to prior works that no longer exist, artworks without titles, and boxes of papers that we have not yet had time to open much less cogently catalog. There are works with authors whose real identity is hidden behind the mask of a pseudonym or a coy phrase like "Kind Gentlelady of Upper Norwich" as a way to evade censorship or skirt social norms, and thus to confound library users. We have parts of things that should be whole: scattered issues of a journal, volume two of a three-volume publication, the left side of a triptych.

Sometimes to be precise about what we have, we should be equally precise about what we do not, yet we may not know what we do not have. Some number of works are permanently lost due to war, conflagration, neglect, and low budgets. Creative works arise in a cultural and social context, and only an omniscient cataloger could place all of the items owned by the library in their proper place in the extended history of human thought. Omniscient catalogers are, however, in short supply.

Because we cannot achieve omniscience, we have to take advantage of the technologies available to us. At the same time, we need to retain a healthy skepticism against any promises that technologies, on their own, will solve all of the problems of connecting today's seekers to the wealth of recorded intelligence (and sometimes lack thereof) that may be available through a library.

This book looks at the ways that we define the things of the bibliographic world, and in particular how our bibliographic models reflect our technology and the assumed goals of libraries. There is, of course, a history behind this, as well as a present and a future. The first part of the book begins by looking at the concept of the "work" in library cataloging theory, and how that concept has evolved since the mid-nineteenth century to date. Next it talks about models and technology, two areas that need to be understood before taking a long look at where we are today. It then examines the new bibliographic model called Functional Requirements for Bibliographic Records (FRBR) and the technical and social goals that the FRBR Study Group was tasked to address. The FRBR entities are analyzed in some detail. Finally, FRBR as an entity-relation model is compared to a small set of Semantic Web vocabularies that can be seen as variants of the multi-entity bibliographic model that FRBR introduced.

PART I

WORK, MODEL, TECHNOLOGY

ONE

THE WORK

As librarians became increasingly aware of the concept of the work as a meaningful creative unit separate from the physical package, various members of the profession put forth their ideas on how to define this abstract concept. The best source of information on this aspect of librarianship is Richard Smiraglia's 2001 book, *The Nature of "A Work": Implications for the Organization of Knowledge.*

You might think that a key concept like "work" would be well-understood in libraries, and uncontroversial. You might also assume that libraries would have integrated this basic concept into their services and procedures. Instead, the integration of the work into library practices is, in this second decade of the twenty-first century, still in our future. As Smiraglia has concluded, "a catalog inventory of books must give way to an encyclopedic catalog of works. In this there is no dissent" (Smiraglia 2012).

I suspect that some dissent could always be found within the cataloging community, but it is true that the question of the work had planted itself fully within the cataloging theory of the mid- to late twentieth century, with Seymour Lubetzky and Patrick Wilson as the most influential theorists of that view.

CREATORS, WORKS, TOPICS

The bibliographic world has its own trinity, which consists of creators, their works, and the place of the works on some conceptual map. None of these concepts is simple, but they vary in their level of complexity. The easiest, from a bibliographic organization point of view, is creators: when neither deceptive nor anonymous, these can often be identified. Next in level of difficulty is the concept of "a work" which is nearly indefinable, yet most of us are quite comfortable with a practical everyday usage of the term. The most complex and difficult concept is that of the topics or subjects of a resource. This latter poses deep philosophical and practical issues, and we have made little change in our approach to subject analysis in the last half century, possibly because there isn't a clear direction for improving this aspect of our work.

I'm going to assume that the treatment of the creator, as well as other sentient beings who have some role in producing intellectual resources, is fairly well under control. The main activity in this area today is the development of broad and interconnected systems that identify the persons and institutions that are responsible for the production of the resources that are created, disseminated, and curated. None of the existing solutions is perfect—neither library name authority data nor the academic systems that allow researchers to create and maintain their own identities—but progress is being made.

Taking a short digression here, it is worth mentioning that the management of personal identity is hardly a new phenomenon, but it has exploded quantitatively with the advent of social media that puts identity management in the hands of the individual. We still have passports and school records and other identities that are not under our control and which in some cases can represent the unwelcome intrusion of social and political powers. The ability for persons to create, manage, and augment their own identities is a revolution that would have been unimaginable to a small-town dweller just decades ago. In a very short while we have gone from "everyone knows everyone else's business" to "on the Internet no one knows you are a dog." We've also gone from a limited scope of relationships to being able to broadcast our thoughts around the world. Unfortunately, that doesn't mean that there are millions who want to listen to us, except perhaps the giant yet impersonal surveillance systems that we now know are hoovering up our bits and bytes, if not actually paying attention to what we have to say.

Socially engineered identity abounds in the modern cultural world. Social and political commentary often takes place in online environments where the authors are pseudonymous. Performers of many types often have a separate public identity from their private identity. In the avant-garde music world, especially where money is not the object and there are few legal contracts that bind relationships, individuals may pass through identities as often as they change their hair color.

Other creative areas have a different approach to identity. Commercial authors' identities are a strong part of their bankability. The best example of this was the attempt by J. K. Rowling, author of the Harry Potter series, to write in a different genre for a different audience, pseudonymously. Sales were modest for the book under the pen name Robert Galbraith. When the true identify of Galbraith was revealed, sales of the book leaped to best seller status immediately. No less a thinker than Michel Foucault suggested that the rise of the author in Western society was precipitated by the need to know who to pay for works, as well as who was to be blamed for them.

Academic writers rely heavily on being properly identified as a work's author so that they will be credited with all of the output upon which their careers depend. This unfortunately has been hindered by the practices of publishers and indexing services, which until recently have not interested themselves in establishing identities, but have been content to record author names without concern for disambiguation. The same person can appear on publications or in bibliographic citations as "John H. Smith," "JH SMITH," "Smith, JH," and so on. Libraries do establish identities for persons, but libraries focus on individually published works, like books, and therefore do not fully cover those academic works that appear in journals.

Returning to subject access to resources, the heyday of library interest in subject access solutions is now quite distant, nearly a century or so past. The development of a combined shelving and classification system in the late nineteenth century by Melvil Dewey was possibly the last great invention in the area of subject access. At the very least, it still informs the methods we use today. Dewey was not alone in his interest in organizing the world of letters topically—that century saw the development of various systems, created by great thinkers such as Paul Otlet, who was responsible for the development of the Universal Decimal System, and Charles A. Cutter, whose Expansive Classification became the basis for the system still in use today in the Library of Congress and other large libraries. In the twentieth century we had S. R. Ranganathan, the Indian mathematician and librarian who promoted the first fully faceted classification system, and also the members of the British Classification Society of the 1960s and 70s in London. Yet in terms of implementation and innovation in subjects, there has been only a slow evolution of the existing systems like the Dewey Decimal Classification, the Library of Congress Classification, and the Universal Decimal Classification. Ranganathan's brilliant Colon Classification seems to have been too complex to find practical adherents. Limited faceting has been implemented in some library systems, but a fully faceted classification was never employed in Western libraries.

The potential revolution in terms of bibliographic models that is the focus of this book has no effect on subject access. No new subject approaches have been suggested along with the new models for bibliographic description. The proposed descriptive models, from FRBR (Functional Requirements for Bibliographic Records) to BIBFRAME to RDA (Resource Description and Access), each contain a small blank spot where subject access of an undefined nature will presumably be attached to the bibliographic record. We can only speculate on the reasons behind this, but it is abundantly clear that the library descriptive cataloging community has a coherence that is not found in the related subject access area. This may be some accident of history, or it could be related to the feasibility of the tasks that the different groups face. Whatever the reason, we find our profession in the midst of an active discussion of descriptive bibliography, with very little attention going to the task of facilitating access by topic.

WORK: THE WORD, THE MEANING

Words are so beautifully and yet frustratingly meaningful, and the word *work* is a key one in our story. The word has many different uses, and some are relatively precise. You work, she works. A work of art. The works of Shakespeare.

Discussions—or arguments—about the meaning of "work" are part of our philosophical history. Notoriously employed by the post-modern literary critics, the conflict of work versus creator has spawned numerous schools of thought. None of this would matter to those of us involved in public services around works except for that element of "public," meaning anyone and everyone. A small group of scientists in a tightly-defined research area can agree on a specific use of terminology, or even invent new terms to communicate amongst themselves, but anyone who intends to serve a liberally defined "public" cannot limit her communication to a small group of cognoscenti. There is danger in making use of a term that is already in wide circulation and that has well-established meaning(s), and yet it often is not possible to do otherwise. That is the situation with "work."

Philosophers, linguists, and cultural critics speak frequently about the meaning of words, but cognitive psychologists actually perform tests. Their focus, however, is less on the individual word but on the concept conveyed and understood by one or more terms. One of the theories that has been the subject of tests in cognitive science is that of degrees of belonging. The easiest way to explain this is to give an example. In an experiment recounted in Gregory L. Murphy's *The Big Book of Concepts* (2004), the subjects are given a list of terms and are asked to put them in order based on the degree to which they answer the question "Is this a fruit?" Although the exact ranking varies, the average ranking comes out something like:

1. orange
2. apple
3. banana
4. peach
5. pear
6. apricot
7. plum
8. grapes
9. strawberry
10. grapefruit
11. pineapple
12. blueberry
13. lemon
14. watermelon
15. honeydew
16. pomegranate
17. date
18. coconut
19. tomato
20. olive

The purpose of this experiment is to show that our categories are not binary; the world is not divided up into fruit/not-fruit, but into a concept of "degrees of fruitness." Few of us would argue with the first couple of items as being high on the "fruitness" scale, and some of us would be surprised to see tomato and olive on the list at all, but not surprised at seeing them at the bottom. How we do this in our brains, and what it means is still an open question. Whether it is subject to some discernable logic, such as commonality of attributes—like sweetness for fruits—is also an open question.

Nor does this ability to categorize bend itself predictably to acquired knowledge. In one experiment, users were asked to rank a group of even numbers based on which they considered the "best" even numbers. Numbers 2, 4, and 8 came out ahead of 34 and 106 (Armstrong 1999). That some even numbers are somehow more even than others is obviously false to anyone with even a minimum background in mathematics, yet the wonderful flexibility of the human brain makes this kind of thinking possible, albeit not necessarily predictable.

If this is a difficult problem with fruits and even numbers, it is an even more difficult problem with less precise concepts. No less an intelligence than Ludwig Wittgenstein set out to prove, in his *Philosophical Investigations,* that we cannot really define unambiguously the concept behind the simple word *game.* That pretty much knocks the wind out of the sails of anyone wanting to use words to communicate anything specific.

We do, however, communicate our ideas and desires and orders using words that represent concepts, and generally our communication is correct. Precision is provided by the context, which also allows us to use terms like *that, this,* and *there*. George Kingsley Zipf, who was an early researcher into the statistical analysis of natural language text, showed that there are a relatively few multipurpose words that we use frequently, and presumably in a variety of contexts. These he likens to the general-purpose tools that we keep close to us on our workbench: a hammer, a screwdriver, some pliers. (And it is no coincidence that the saying begins "if all you have is a hammer . . . .") These we can use in many ways. Further out on our workbench, and in the statistical curve that he derived from natural language texts, we find the specialist tools; these are the ones that we use only occasionally, when the general purpose tools are not adequate. Essentially, Zipf provided a logical explanation for the linguistic long tail. The word *bird* will be in the high use area, while *passerine* will be in the long tail (Zipf 1949).

The word *work* is a hammer-like tool, using Zipf's analogy; it has an imprecise but highly utile meaning. Like many common words in English, it is both a noun and a verb, so to begin with we have to make clear that we are only interested in the noun form. Even with that restriction you can "have work" (meaning employment), "do some hard work" (meaning to labor), or "create a work" (produce a result of some kind). My garden can be a "work of art," as can a Van Gogh painting. My house is near the "public works" offices of my town, and my bookshelf holds the works of many authors. The word *work* is one of those multipurpose words that supports George Kingsley Zipf's Principle of Least Effort: it is a word with multiple meanings that, however, makes sense in context.

SOME HISTORY

We live today with an abundance of "product"—there are more books than readers who want them, as evidenced by the copious piles on remainder racks at our bookstores. It wasn't always thus, of course. Before the advent of printing, each copy was unique and there were few of them. Printing brought exact copies, but it also brought editions, as printers throughout Europe produced their own versions of texts. One European intellectual of the 1500s, Conrad Gessner, felt a need to gain some control over this tsunami of works; he set out to create a universal bibliography of all works in print, but not all of the various editions of the works. Gessner's *Bibliotheca Universalis* was in part a response to what he saw as wasteful duplication among printers, and he hoped that a list of available works would lead them to concentrate on new works rather than reprinting works already on the market (Serrai and Serrai 2005). Here it can be said that Gessner obviously did not understand the economics of the book trade.

Libraries, some private, some public, also took advantage of the increased printed book production to grow their collections. One such collection was that of the British Museum Library. In the early 1800s, Anthony Panizzi found himself as head of the British Museum Library with the wonderful title "Keeper of the Printed Books." This means that there was a parallel position for the other kind of books—manuscripts—and therefore it was necessary to state that "printed books" was a distinct department. We can see this as a kind of microcosm of the transition from precious objects to an abundance that required, as it was later called, "bibliographic control."

Panizzi had some major problems on his hands. The library's catalog had been long neglected to the extent that the library had no inventory of its holdings and users could not be sure if the library had the book they sought. The library also had many works in multiple editions coming from the very active English presses. Clearly, Gessner's goal of stemming the tide of multiple printings of the same work had failed.

The library board had allocated funds for the creation of a new catalog, but not enough to create the catalog that Panizzi felt was needed. This led to the famous showdown between Panizzi and the board as Panizzi explained that a mere "finding list" of authors and titles would not be sufficient for the library to serve its users, nor to efficiently continue to build its collection. The cataloging rules devised by Panizzi specified in each case that the edition be noted by the place of publication and the date, as well as a numbered edition if so stated. (Interestingly, the names of the printers—whom today we would call publishers—were only to be included in his catalog if the printer itself had achieved some level of eminence.)

Some forty years later, when Cutter presented his Rules for a Dictionary Catalog in 1876, one of his objects was for the catalog "to assist the user in the choice of a book (G) as to its edition (bibliographical)."

During the decades from 1840 to 1870, the time between Panizzi and Cutter, distinguishing different editions of the same work had become the norm in bibliographic control. Cutter did not discuss whether some users might not care precisely which edition they received, although he did provide an example of the user for whom editions would matter: "for the student, who often wants a particular edition and cares no more for another than he would for an entirely different work." Cutter's rules, though, still placed an emphasis on places and dates, and not the publishers themselves: "Print publishers' names, when it is necessary to give them, in italics after the place" (Cutter 1875).

The rules also acknowledged that the same catalog that served the users also served the library's collection development function, in that the recording of editions was also needed "in the library service, to prevent the rejection of works which are not really duplicates." Duplicate, in 1875, meant the same edition, not the same work.

In my research I have not uncovered the tipping point that led library thinkers like Seymour Lubetzky and Eva Verona to take up the question of the work versus the edition. Yet somehow between the latter part of the nineteenth century and the first half of the twentieth century, it appears that the number of different editions in libraries had become burdensome to users. Although it was still essential to distinguish between editions, it also became important to inform the user that a certain group of editions represented the same work. In just a little over one hundred years we had come full swing from presenting users solely with works, then solely with editions, to needing to gather editions back into their work groups.

THE WORK IN BIBLIOGRAPHIC CONTEXT

We've seen that the term *work* covers a number of different concepts. The difficulty that we have is not with the word, however, but with the meaning that we ascribe to it. Eva Verona, who could be regarded as an early twentieth-century philosopher in the area of cataloging, chose to refer to the focus within the cataloging context as the "bibliographic unit" (Verona 1985). That would distinguish the "item in hand" that is being described from the abstract concept that some wish to be called a "work." Indecs, the metadata model developed in the late twentieth century for digital commerce, referred to "stuff" in its basic diagram, which reads: "People make stuff; people make deals about stuff." This is an interesting punt on defining the exchange of value for labor. (One wonders how Karl Marx would have reacted to such a definition.)

The question of defining the work in the context of library catalogs is multifold. Its meaning must be functional, that is, it should serve a purpose. Defining that purpose is not a simple matter. It also needs to communicate readily to the broad and heterogeneous population that both creates catalogs and uses those catalogs. Without dwelling overly on the choice of terms, we can look at the desired functionality expressed by thinkers in the library arena.

Lubetzky's Work View

Seymour Lubetzky was arguably the most influential force in cataloging theory in the twentieth century. He began working at the Library of Congress (LC) in 1943, and one of his first assignments was to do a study of the descriptive cataloging rules used by LC at the time, the second edition of the A.L.A. Cataloging Rules, published in 1941. Lubetzky's analysis led to a revision of the rules, issued in 1949. By 1955 he was awarded the Margaret Mann Citation for his contributions to cataloging. He continued to study, publish, and teach as a professor at the School of Library Service at the University of California, Los Angeles. Even after retirement in 1975 he spoke at meetings and participated in discussions. He published his last work in 1999. In the year 1998 the library world feted Lubetzky's one-hundredth birthday with a special symposium. Lubetzky was there. He died in 2003 at the age of 104.

Lubetzky's analysis of the principles of cataloging, published in 1969, became the groundwork for all cataloging rules that have followed. This work greatly influenced the revision of the Anglo-American Cataloguing Rules (AACR) in 1978. Although clearly erudite and studious, Lubetzky's approach to the catalog had a large dose of common sense. In particular, he insisted that the cataloging rules be derived from the functions they were to serve. This was not the case with the 1941 ALA rules that he was first asked to study, which resembled, according to Julia Pettee, "an encyclopedia of pedantic distinctions." (Lubetzky 2001, xiv) Some of Lubetzky's ideas would be considered heretical even today. For example, he decried the repetition of the author between the heading and the statement of responsibility. He also criticized the fact that the information on the card was not placed in order of importance, causing users to scan through unwanted information to look for what served them.

There are two threads in Lubetzky's work that came to the fore at the end of the twentieth century when new bibliographic models were proposed. The first is that the content of the book is not represented by a physical description of the book. This seems obvious, but descriptive cataloging does focus on physicality, and sometimes solely on physicality. Lubetzky argued that the physical "is only a medium through which the work of an author, the product of his mind or skill, is present . . . and that, consequently, the material and the work presented by it are not, and should not be treated as one thing" (Lubetzky 2001,). This is the separation of content (the work) and carrier (the physical medium), although the implementation of this in the library catalog remained (and remains) vague. The second thread is that these physical books (or other media) can be editions of the same work. This establishes a relationship between bibliographic items based on their "workness." Unfortunately exactly how one determines workness was neither defined nor explained. As we know from later efforts, this raises a number of awkward questions about where one work ends and another begins, and whether there are degrees of workness.

Lubetzky did take up the question of books versus works. In his *Principles of Cataloging, Phase I,* issued in 1969 (and never completed), he recognizes that the book itself is a complex entity:

> In summary, then, it must be recognized that, genetically, a book is not an independent entity but represents a particular edition of a particular work by a particular author; and that, consequently, it may be of interest to different users either as a particular edition, or as a representation of a particular work, or as a representation of the work of a particular author. (Lubetzky 2001, 272)

The lack of a definition for works means that some assumptions of the time are not necessarily ones that would be accepted today. Lubetzky was one of the first cataloging theorists to attempt to address the wide range of new media in the cataloging rules, treating non-books as first-class bibliographic entities in their own right, no less worthy of being entered into the catalog than books. In this quote, he allows the concept of "work" to cross the boundaries of physical media, saying "that the same work may be presented in different *media,*" a view that would be greatly qualified today as changes in medium of the type he lists here are considered changes in work.

> Beginning then, with the material cataloged, it is recognized in the revision from the outset that a book, phonorecord, motion picture, or other material is only a medium through which the work of an author, the product of his mind or skill, is present; that the same work may be presented through different media, and in each medium by different editions; and that, consequently, the material and the work presented by it are not, and should not be treated as one thing. (Lubetzky 2001, 199)

Writing in the time of the card catalog, Lubetzky's solutions to the work/edition question are limited to the collocation of works through the use of a "main entry" that consists of the author and the title, or, in the case of editions of a work, the uniform title. Although Lubetzky is considered to have brought the work question to the attention of the library cataloging community, his cataloging rules had little to say about workness, although they did provide significant new approaches to authorship.

In that same 1960 publication, Lubetzky defined a two-part set of primary objectives for the catalog:

> (1) to facilitate the location of a particular publication, and (2) to relate and bring together the editions of a work and the works of an author.

Relating of editions of a work became known as the "second objective," and it was this issue that was addressed by Patrick Wilson not long afterward. The second objective and what it means for the bibliographic model will be covered in a later chapter.

Wilson's Bibliographic Families

Patrick Wilson, professor of Library Science in the University of California at Berkeley School of Library and Information Science, published his book *Two Kinds of Power* in 1968. Although not a focus of the book, he addressed the meaning of the term *work* in the first chapter, "The Bibliographical Universe," in which he defines what he sees as the inhabitants of that universe. It is interesting that by referring to "inhabitants," and not "things," he creates an atmosphere of living beings.

Wilson focuses on texts, and describes the world of letters thus: a person composes a *work,* by ordering letters and words into a *text,* and setting these within an *exemplar.* He makes the point that "these three descriptions are not independent, for he could have produced no work without producing some text, and could have produced no text without producing some permanent or transitory exemplar of the text" (Wilson, 1968, 6). Although they are not independent, each has its own distinct qualities. This may be the first elaboration of the model underlying Group 1 of the Functional Requirements for Bibliographic Records (FRBR), although, as we'll discuss in the modeling section, no two approaches to the inhabitants of the bibliographic biome create exactly the same division of that body.

What Wilson contributes in particular is his own unique definition of the work. He defines a work not as an aspect of a single text, but "a work simply is a group or family of texts." In keeping with the view of beings that inhabit the bibliographic universe, Wilson's works are not static, but the work families develop over time as texts are reproduced or republished in the same or modified form:

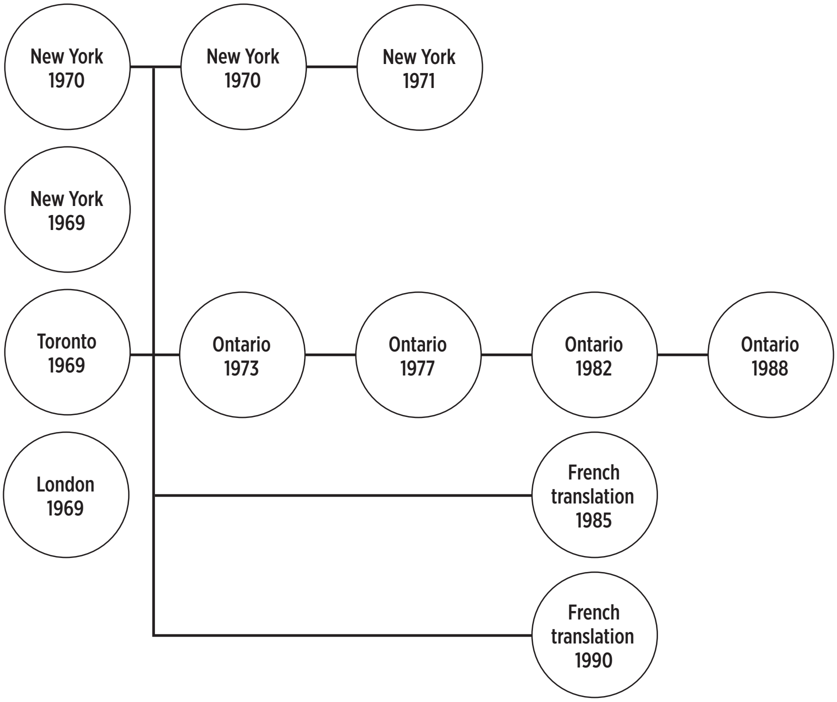
> The production of a work is clearly not the writing down of all the members of the family, but is rather the starting of a family, the composing of one or more texts that are the ancestors of later members of the family. (Wilson 1968, 9)

Wilson's view is one possible interpretation of S. R. Ranganathan's statement that "a library is a growing organism." In Wilson's view, the library grows not only in the number of volumes, but with the addition of volumes families grow in a variety of ways. Each addition to the library potentially adds to the familial relationships that are there, and thus each may alter the nature of the bibliographic family that exists. Works are groups that grow and change over time as new editions or new related works come into being. This of course is a challenge for cataloging because it suggests that catalog entries may not be immutable if relationships are to be included in the catalog. There are relationships from newer resources to older, which could be represented in the description of the newer item only, but the family may grow in different directions. Because items are not necessarily added to the catalog in their order of publication or relation, introduction of new relationships could be disruptive.

In figure 1.1, the "progenitor" is a hardback published in 1969, with a close kin being a paperback in the same year from the same publisher in New York, a Canadian version published in Toronto, and a version published in London. Reprintings of the New York and Toronto versions become children of their respective progenitors. Translations follow, each with the original as "parent" and potentially with children of their own if there are republications of those.

FIGURE 1.1

"The Studhorse Man" as a Wilsonian family, based on Smiraglia 2001



There is no precise definition in Wilson's text to tell us what makes one text a member of a particular family. He considers translations to generally be members of the same family as the progenitor work, but doesn't exclude the possibility that some translations may go so far as to overcome their cultural genealogy and start their own families. It also appears that Wilson did not exclude the idea of a family including adapted works, such as films derived from books. Not being confined by the need to codify his ideas in cataloging rules, he leaves the topic of the work without pinning down a functional definition, and seems to relish the remaining ambiguity: "While there is good reason to distinguish work from text, it is necessary to recognize that the notion of a work is an incorrigibly vague one" (Wilson 1968).

In a 1989 article entitled "Interpreting the Second Objective of the Catalog," Wilson points out something that is obvious once mentioned but often overlooked: that the catalog generally only includes separately published works. Those separate publications often include multiple works, from the prefatory material to the main content, to photographs or illustrations that accompany a text (or to text that accompanies a publication of photographs or illustrations). "By no stretch of the imagination can the author/title catalog be said to give information about all the works available in the library" (Wilson 1989). This of course complicates the study of works, as well as the development of any solutions based on how "works" are defined in the library catalog.

Leaving the work without sharp boundaries is consistent with the remaining theme of his book, in particular his description of the exploitative power, which is individual and contextual and therefore cannot be defined with absolute precision. It is probably his training as a philosopher that allowed him to be comfortable with "incorrigibly vague" concepts; it should come as no surprise that these concepts, then, did not find their way into rules for bibliographical control, where catalogers can't easily sit on the fence over the relationship between a text and a work.

Smiraglia's Semiotic View

Richard Smiraglia has written perhaps the only book on the work question: *The Nature of "A Work": Implications for the Organization of Knowledge* (2001). He covers the various definitions that have arisen in librarianship, more than I include here, but also adds his own, based on the branch of philosophy known as semiotics. Semiotics is a study of meaning, and how meaning is created using signs and symbols. Semiotics is also a study of communication, and therefore touches ever so slightly on the communication theories that have been born out of mathematics and computation. However, the two strike out in very different directions, with semiotics remaining unquantifiable.

Smiraglia calls works “vehicles for communication” and says that “works contain representations of recorded knowledge.” Their role is social because they “transport ideas along a human continuum.” Works are born as works, both in Smiraglia and Wilson’s definitions, yet both allow the “workness” to grow to include new instances as more of the (presumably) same ideas are brought forth as publications.

Smiraglia includes both the ideas and the symbols in his definition of work, whereas Wilson speaks separately of work and text. This speaks to the abstractness of the concept of work; for Smiraglia the work must have been expressed in order to exist. This separation between ideas and expressions is an area where the philosophers of this area diverge.

By taking a semiotic view, Smiraglia includes the reader in his view of the work, and affords the work itself with a cultural and communicative role that changes with each reading (or viewing, or listening). The work is in the eye of the beholder.

> Thus we replace the arbitrariness of the abstract concept of the work with a definitive changeling. Works change over time, they take on new meanings as they are assimilated in cultures, they reflect their perceptions, and they evolve in content and tangibility. (Smiraglia 2001)

Because his view includes communication and culture, his theory can take into account some of the particular characteristics of different kinds of works, such as music, which has the added facet of performance.

Unlike the pure theorists in this summary, Smiraglia conducted quantitative research to discover the extent of work relations in libraries. Using Wilson’s concepts of family and progenitor, he sampled the OCLC WorldCat database, New York University’s Bobst library, the Georgetown University library, and the Burke Theological Library. Note that these studies were done in 1992 and 1999 and the nature of WorldCat changed considerably after that time, increasing tenfold due to the addition of many millions of bibliographic records from nonmember, and primarily non-US, libraries. The studies were also done on physical libraries, and a combination of physical and digital holdings today could yield different results.

The results in these libraries varied by the type of library: the theological library had numerous older books in its collection, and showed a high rate of “families” in its history area. OCLC, being a union catalog, had the greatest variety of work types. The university libraries each had their specialties, which affected the results of the study. In the end, however, Smiraglia concludes that the “only strong predictor of derivation was the age of the progenitor work” (Smiraglia 2001).

In other words, families develop over time. They also tend to develop more for some genres, like fiction and drama, than for scientific works.

Both Wilson and Smiraglia emphasize that what begins as a new work can give birth to a large family of works through a variety of changes such as revisions, augmentations, performances, and adaptations. Where one draws the line and declares that a new work has been created, however, is not clear.

Coyle's Cognitive View

This is a previously unpublished theory, so I must describe it here at some length. In the section on "Work, the Word," above, I presented a brief explanation of how cognitive science approaches "meaning" and the concepts that are conveyed when we use words to communicate. Cognitive science has studied numerous models of conceptual thinking as part of the human understanding of the world. Concepts have an element of generality/specificity whose exact function in understanding and communication is not yet clear. Regardless of our inability to define how thinking works, every moment provides proof that we do share enough of our conceptual matter to function together in the world. All of this has a strong social component. One of those commonalities is something referred to as the *basic level of categorization,* which means that within a social group we have understood common levels of specificity for things and concepts (Murphy 2004). A simple illustration is this:

Jane and John are walking down the street when they see their neighbor's calico cat. John says: "Hey, there's Fred's cat." Later, at the zoo, Jane says to John: "Take a look at that tiger." Both are felines, yet the words *cat* and *tiger* demonstrate different levels of categorization within our culture, probably based on how common these things are in our shared experience. Each is an understood shorthand for what is obviously a much more complex concept. There is no need to say: "Look, there's a vertebrate mammal of the feline species, sub-species house cat, variety calico, whose owner is Fred," even though that is indeed the case. Instead, "cat" is the level of categorization that allows us to efficiently express a concept that others in our environment will most likely understand. When you type "cat" into the English language Wikipedia, the article that is retrieved represents this same concept of cat as "house cat," while "tiger" gets its own page. This reflects a shared level of categorization in the English-speaking (and Wikipedia-editing) world.

The basic level of categorization is not an absolute, however, but depends on a social context. Experts in a field will have a different basic level than the general public (e.g., "*Passer domesticus*" and "sparrow") and aficionados amongst themselves will make distinctions that a less interested person will not ("Mercedes-Benz C 215 V6" and"car"). Analogously, librarians will have a shared professional understanding of bibliographic distinctions that is at a more detailed level of categorization than members of the general public.

Lubetzky and others frequently state that a library patron may state that he is looking for a book, when in fact he is interested primarily in the work rather than a specific physical item. The question, though, is what does the patron mean by "book" and what does the librarian mean by "work"? Smiraglia's study of the nature of the work shows that no one single definition of *work* exists among librarians.

If we look at the user view with basic level of categorization in mind, as well as the user's goals, we can then compare that with existing definitions. I'll take as a very simple case a person going to the library to find and check out a book. This person goes to the library and says that he is looking for "the book, *Moby Dick*." Lubetzky and others would say that the user is interested in the work, not a specific physical item. Shoichi Taniguchi (2003) would instead say that the user is interested in the actual text, not the abstraction that is the work. Cognitive science would say that "the book, *Moby Dick*" is a contextual shorthand, most commonly used to refer to a physical (or, today, electronic) book with the text of *Moby Dick*. The user doesn't distinguish between, in Wilson's terms, the work and the text and the exemplar, unless necessary to convey a specific query. The user may not include in her conceptual level that there are variations like translations, annotated editions or works about *Moby Dick* if those are not of interest to her, or not relevant to her immediate context.

The expert user view, for example, that of a professor of American literature who is doing a particular study of technical language in Melville's text, could be very different. Although the level of categorization will be different from that of the casual reader, the focus is still likely to be on the text in a concrete form (on the page or in a digital format). This user may qualify his request as being for "an authoritative version of *Moby Dick*" and may want to check the *bona fides* of the publisher or digitizer. This person is interested at the level of the manifestation, but is still hoping to exit with a real-world object that he can study.

If I say that I have read *Moby Dick,* I am speaking of an experience with a physical book or device that contained the words of Melville and the story those words express. As semioticians might claim, the ideas left in my head from that experience were developed through my experience with the physical book, the text on those pages, what was going on around me during the time that I was reading, and how I interpreted the meaning of that text in my personal context. Nevertheless, a real-world object was encountered.

In both speaking and thinking, we use single and simple terms to represent complex topics, otherwise we could not communicate efficiently. The shorthand used can be fairly imprecise and still support communication. "Have you read *Harry Potter*?" can mean any or all of the books in that "arc" or series. I could answer simply "Yes," meaning that I have read at least one of the works or perhaps all of them. In daily conversation, these shorthands do not cause us problems, in part because we can clarify in the conversation, "All of them?" "Which ones?" But we can also go straight to "What did you think? Good?"

In the cognitive sense, these are not abstractions, but are shared concepts for concrete things that we express with a commonly understood level of categorization that is not too broad to communicate to the other person, but not more specific than it needs to be. The work is often defined as an abstraction, an idea, yet when I ask "Have you read *Harry Potter*?" my question implies inclusion: that the shorthand "*Harry Potter*" represents the whole, and that I am asking my listener about one or more books that the person may have held and read.

In this cognitive model, there is no one definition for "work." It will have meaning within a context and that meaning will often be shared, but not always. The basic level of categorization within that context will vary depending on who is participating in the communication. Librarians are free to develop an expert meaning for the term, but cannot expect that meaning to be shared perfectly with the others. Interaction between libraries and library users of all levels of expertise and knowledge has to mimic the flexibility that humans use unconsciously when communicating, and cannot be so fragile that it is defeated by some degree of ambiguity. For this reason, we should focus on needs and functions, and not on a particular term.

Taniguchi's Expression-Dominant Model

Shoichi Taniguchi is a professor of library science in Japan. He began looking at the models for descriptive cataloging in the mid-1990s at the same time that work was being done by IFLA on the Functional Requirements for Bibliographic Records. Where Lubetzky's general feeling was that most users entering a library looking for a "book" want the "work"; Taniguchi's proposed model placed emphasis on the expressed text, rather than the more abstract work. In fact, Taniguchi's model is probably a better description of the basic level of categorization for texts. He proposes a model of bibliographic description that does not place the work nor the manifestation in the dominant position. He originally called his view "text-dominant," but that was before FRBR's *expression* was defined. His current work is a direct response to FRBR.

In Taniguchi's view, each bibliographic item must be described with a dominant expression entity. Titles, statements of responsibility (including added entries), and edition statements describe the expression; the manifestation bibliographic level contains only those attributes related to publication, physical format, and publication date. His assumption is that the majority of user tasks should be satisfied by the expression in most cases, but could include the work in those situations where work information is normally provided. Using Taniguchi's approach, it is not necessary for a library to create separate entries in its catalog that describe individual manifestations if those are not required by the particular user community. Therefore, time is saved by entering into the catalog the information about the expressions held by the library, and allowing most users to select between manifestations (if more than one exists) at the shelf. The catalog record therefore informs users which expressions the library owns, which is the minimum information needed to fulfill the user tasks. Those few users interested in the details of the manifestation can go on to that level of detail in display.

Taniguchi concludes that very little information about works is included in bibliographic records, although data derived from the manifestation, such as creators, titles, and subjects, is about the work, not the manifestation. The emphasis in cataloging rules is on describing the physical item "in hand," and therefore the dominant entity is the manifestation. He de-emphasizes the physical aspects and organizes his model around the content that the user encounters in the expression of the work.

Although Taniguchi's approach seems at odds with the cataloging of books in libraries, it is easier to appreciate when looking at federated search systems that combine both physical and digital versions of the same materials. This is especially true for journal databases where the particulars of the manifestation have little weight and the types of augmented editions (with added prefaces, illustrations, or commentary) that exist in monograph publishing are virtually unknown. The definition of work may not hold true over all possible bibliographic materials, and may evolve over time as new means of communication develop.

WORKS AND RELATIONSHIPS

Inherent in, but not necessarily explicit in, the definition of works is that bibliographic resources have relationships between them. One of these relationships is "this is a copy or version of the same work," but beyond the question of an exact copy the range and complexity of relationships grows. Most such relationships were not formalized in library catalogs. Instead, for some key relationships, like translations, supplements, and editions, the coincidence of collocation of entries under the same or similar headings (author, title) was enough to create a logical proximity between related resources. Where headings do not provide collocation, notes are sometimes added.

The new emphasis on works and work relationships spurred discussion of the types of relationships to be found between bibliographic resources. In her 1987 doctoral dissertation, Barbara Tillett undertook a comprehensive study of bibliographic relationships by studying a large set (over two million records) of MARC records from the Library of Congress database. Within these she studied the notes fields that represented statements of relationship, and categorized them. Tillett derived seven types of relationships: equivalence (the same content), derivative (adaptations), descriptive (reviews), whole-part, accompanying, sequential (series), and the more general shared characteristics relationships.

There is no single relationship in Tillett's categorization that translates to "same work" by any of the above definitions. The "equivalence" relationship is limited to copies, reprints, and other republications of precisely the same content. Derivative works include subsequent editions, translations, and adaptations, such as a rewriting of a book for a new audience. Because Tillett studied individual cataloging records produced by the Library of Congress, the bibliographic units in the relationship would be the cataloged publication. As per Wilson's caveat above about the limitation of the library catalog to separately published works, this study covered only some of the bibliographic items held in the library, because it did not include those literary units that were included in larger publications. Tillett did include whole-part relations in her study, but these had to be extrapolated from the existence of contents notes. Clearly the definition of relationships is related to the definition of the unit of bibliographic description, which will become clearer when we look at FRBR and the relationships defined in that model.

In summarizing the seven types of relationships (with their sub-relationships) Tillett wrote in her dissertation: "The primary categories of the above taxonomy meet the criteria of being mutually exclusive and totally exhaustive" (Tillett 1988). Wilson, not only with a more philosophical bent but in his position as a tenured professor, is much less inclined to make such a bold statement. In contemplating relationships, Wilson notes that he does not believe that there is a finite set of relationships, and thus discourages attempts to define such a set. In a practical application of relationships to bibliographic units, the truth is probably somewhere between these two views, with some set of relationships covering the majority of useful relationships, but always allowing for expansion as more is learned or as the nature of catalogs changes.

In the years since the important work that Tillett did to categorize relationships, the possibility that relationships should become incorporated more thoroughly into cataloging rules has gained traction. Her analysis influenced both the development of FRBR as well as that of the cataloging rules, Resource Description and Access (RDA), both of which Tillett was involved in creating.

WORKS IN CATALOGING PRACTICE

Without actually defining the difference between a "book" and a "work," both terms are used in the International Cataloguing Principles of 1961. The key to their use leads us back to Seymour Lubetzky, who, according to Richard Smiraglia and others, greatly influenced the creation of the 1961 principles. The use of *work* in the International Cataloguing Principles seems quite natural on the surface. The functions of the catalog include both "whether the library contains a particular book specified . . . by its author" as well as "which works by a particular author." The term *work* here presumably has the sense of "oeuvre," in the broad meaning of that concept. The Principles state that "The main entry for works entered under title may be either under the title as printed in the book, with an added entry under a uniform title, or under a uniform title." The uniform title is a contrived title that brings together some members of a bibliographic family. The instructions leave the definition of a work and when it should be represented in the catalog to the discretion of the cataloger, all along avoiding any need to tackle the very difficult task of defining what a work is.

By creating a special work title that would be assigned to all instances of the work in the descriptive cataloging, all editions of the same work would be collocated. In his 1989 article that primarily echoes the thinking in Ákos Domanovszky's 1975 book *The Functions and Objects of Author and Title Cataloguing,* Patrick Wilson suggests that one could go beyond recording merely the same edition of a work, but could form a family of works that could include strongly related texts, such as supplements, commentaries, and continuations. This view begins to approach Wilson's desire that a catalog make explicit the relationships between items in the library, with "same work" as only one possible relationship. The relationship "same work" (which may also extend to "same expression" or "same text") is implemented in library catalogs using the mechanism of the *uniform title*. First introduced in the A.L.A. Cataloging Rules of 1941, the uniform title gained additional prominence in the editions of the Anglo-American Cataloguing Rules (AACR). Unfortunately, the uniform title has been applied very unevenly in libraries, and this is at least in part due to the problem of scope.

The purpose of the uniform title is to collocate, that is, bring together in the same place, the versions of a single work. "Collocation" in library cataloging takes place through the relative position of the items in the alphabetically ordered list of the catalog. To overcome differences in how names of creators and titles of works are presented in actual publications, collocation within the ordered list is accomplished by using standardized "headings." These are controlled text strings for the bibliographic data that will be represented in the catalog, such as the names of authors, titles, and subjects. Collocation may sound simple, but in fact there are numerous adjustments that must be made in order to bring together items that the cataloging rules deem to be the same bibliographically. In particular, the collocation of works requires the cataloger not only to identify that different resources represent the same work, but also to provide a heading that will bring the works together in the catalog.

Collocation for works fails in some cases in spite of the normalization of author names because titles of publications of the same work can vary. In modern works this is most often true for translations:

> The magic mountain
> La montagne magique
> Der Zauberberg

Older and ancient works, such as the works of Shakespeare or early sagas that were written before their language or dialect was normalized, may also have titles that have varied over time, like:

> Hamlet
> Hamlet, Prince of Denmark
> The tragedie of Hamlet, Prince of Denmarke

To collocate these in the catalog as variations of a single work, an additional title is added between the author and the title of the printed book. This is called a "uniform title" and it serves as a normalized title that represents the bibliographic work. Where known, the uniform title represents the title of the original publication of the work. In other cases, the title is a selected title, such as "Hamlet," that contains the commonly known name of a work that was published under many different names, especially in its early period. The uniform title can also contain the language of the translation and/or the date of publication, to distinguish between different versions or editions.

```
Mann, Thomas
 Der Zauberberg
Mann, Thomas
 [Zauberberg. English]
 The magic mountain
Mann, Thomas
 [Zauberberg. French]
 La montagne magique
Shakespeare, William
 Hamlet
Shakespeare, William
 [Hamlet]
 Hamlet, Prince of Denmark
Shakespeare, William
 [Hamlet]
 The tragedy of Hamlet, Prince of Denmark
Shakespeare, William
 [Hamlet. Italian]
 Amleto
```

The uniform title, shown here between square brackets, represents the work with a "work title" combined as needed with something that distinguishes between different versions. In the above case that distinction is made with the language of translation, but for some works that appear in different versions in the same language, such as the works of Shakespeare, the expression may be represented by either a date or both a language and a date.

However, in the current cataloging rules, any publications whose title would be the same as the uniform title are not given a uniform title, and the majority of publications have only a single edition, and thus need no uniform title. AACR2 explains it this way:

> The need to use uniform titles varies from one catalogue to another and varies within one catalogue. Base the decision whether to use uniform titles in a particular instance on:
>
> a) how well the work is known b) how many manifestations of the work are involved c) whether the main entry is under title d) whether the work was originally in another language e) the extent to which the catalogue is used for research purposes.

> Although the rules in this chapter are stated as instructions, apply them according to the policy of the cataloguing agency. (Anglo-American Cataloguing Rules, 1978)

As you can see, the exceptions to the creation of a title for a work are both numerous and subjective. Bringing out the "workness" of a resource is the exception rather than the rule, and many libraries make little or no use of uniform titles for the work.

The first exception is that any item that has been published in only one edition or in only one language is not assigned a work title. Even the main proponent of identifying works, Seymour Lubetzky, stated that "wherever an author is identified in his works by one particular name and a work is represented under one title only" nothing more needs to be done to identify the author and the work.

In addition, the different editions of a work are not given a work title in cases where the titles of the editions do not interfere with collocation, as in reprintings or updated editions:

> Eysenck, Michael W., and Mark T. Keane. *Cognitive Psychology: A Student's Handbook*. Hove [u.a.]: Psychology Press, 2010. 6th edition
>
> Eysenck, Michael W., and Mark T. Keane. *Cognitive Psychology: A Student's Handbook*. Hove [u.a.]: Psychology Press, 2007. 5th edition
>
> Eysenck, Michael, and Mark T. Keane. *Cognitive Psychology: A Student's Handbook*. Hove: Psychology Press, 2003. 4th edition

Thus, different editions or versions of a work (or members of a work family) are only identified through a heading in those cases where the work title is needed to collocate the entries. If they already collocate by the coincidence of having the same titles, no work is identified.

Note also that, as shown above in the Thomas Mann example, the catalog entry for the item *Der Zauberberg* does not require a uniform title because the uniform title would be the same as the title of the book. This complicates the rules for sorting in catalogs because it requires a cascading sort of uneven membership, where the "real" title must sort before the uniform title that contains the exact same characters.

The uniform title is a great illustration of the tension between serving the individual library's users and the efficiency that can be gained through massive sharing of cataloging copy. Although allowing each library to make its own decisions as to when to bring out the "workness" of a resource is sensible both from a question of workflow and user service, it has a definite effect on data sharing. What makes a work useful or necessary in one library could be a distinct hindrance in another. A library may have copies of Tolstoy's *War and Peace* in English, Spanish, and Chinese, but not in Russian because the library does not serve a Russian-speaking population. Therefore, each translation can be found under the title in the translated language, which is logically where readers would look to find the book:

Guerra y paz   War and peace   戰爭與和平

but it may not be useful to also include an entry under the original Russian title, война и мир. Yet making that decision and adjusting the cataloging copy for the individual library has a cost in terms of cataloger time.

The concept of work in library catalog data is currently unevenly applied in practice. Individual libraries or library groups can and do opportunistically decide whether to make use of this feature based on the criteria in the cataloging rules, plus the perceived needs of their users and the capabilities of their catalog software. Key to the upcoming sections on FRBR is the fact that prior to FRBR, the work and the expression were considered bibliographically significant only under certain circumstances. In part this was because the creation of a specific heading for the work had effects on the catalog and the user experience that were either deemed unnecessary or even detrimental to the users of that library.

So far I've spoken of only one kind of work or "uniform" title. There are two others. The first is the collective title, like "Complete works" or "Selections." The second is the particular type of uniform title used for music materials. Both of these perform the same collocation function that is the basis of the work title, but they have significantly different meanings. The collective title identifies a particular type of publication, often not used on the title of the piece. My own observation is that this is unevenly used, even in large libraries. The music title, however, is a thing unto itself, and is probably the most successful application of work titles to a bibliographic group. Music is in many ways a special case because, unlike texts, musical compositions often do not have a single distinctive title. In addition, we experience music through performances, not through the original creation of the composer. And, last but not least, recorded music is most often packaged by publishers with two or more musical pieces per package, meaning that there is extensive use of the “added entry,” an author/title heading that essentially has a part/whole relationship to the main bibliographic entity.

Music uniform titles are crafted descriptions of the music piece, which sometimes ignore what most of us would consider a true title for the piece. For example, Beethoven’s symphony known as the “Eroica” (but also as Beethoven’s Third symphony) is given this uniform title:

Symphonies, no. 3, op. 55, E flat major.

No one would consider this artificial construction as the proper name of the symphony. Yet the method neatly orders music—at least classical music—and overcomes the lack of uniform practice in naming such works: “The Eroica,” “Beethoven’s Symphony #3,” “Beethoven’s Third Symphony,” “Sinfonie in Es-Dur,” “Symphonie no 3 en mi bémol majeur,” and many more.

Although important conceptually, as we’ve seen here, direct presentation of the work in cataloging is limited to a relatively small number of cases in libraries today. Taniguchi points out that in current cataloging the work does not “perform a key role in describing an item being cataloged, although its existence is supposed to be a prerequisite in making a bibliographic description.” Catalogers simultaneously describe the item in hand and extrapolate some degree of “workness” in assigning headings, but only when that seems called for. Moving to a bibliographic description that recognizes the work sufficiently to reveal the bibliographic families that Patrick Wilson describes means a significant change in cataloging practice. Recognizing those works in a way that the bibliographic families can be identified and offered to users as such is a much more difficult task plagued with some deep philosophical and practical questions. Among these is that of defining the boundaries within which bibliographic decisions take place. By elevating the bibliographic discourse from publications to works, the universe expands from the physical library and the item in hand to an essentially unbounded abstraction. Exactly where that abstraction should be addressed, whether within the inventory of a single library or in some aggregated bibliographic layer that is not limited to a library’s holdings, is a question that has not been answered, and often is not even asked.

Summary

The preceding definitions of the work are not to be taken as exhaustive nor conclusive. These definitions hopefully provide a bit of perspective for when we talk about the more functional approach of current bibliographic models.

There are many issues that are not addressed here but that pertain to how we define works. There is no in-depth discussion of whether all resources have some degree of workness. The studies cited here were either limited to text, or text and music. Recorded knowledge comes in other forms, including aerial photography, topographic maps, and scientific datasets. Whether each of these is also imbued with the work quality as defined by our thinkers is not clear. In his 1989 article "Second Objective of the Catalog," Patrick Wilson identifies some resources that are publications without being works, such as collections of shorter works between a single set of covers. This is not a universally accepted point of view, as we will see in the section on FRBR and aggregates. Although this opens up the possibility that there are "non-works" it does not provide criteria that we could use to divide the works from the non-works.

There is also little discussion of the domain of discourse in these definitions. Patrick Wilson's *Two Kinds of Power* addresses the need to define a domain, but rather oddly defines the domain of the library catalog as being the items in the library as well as items being considered for inclusion in the library. In other areas, he speaks of the "bibliographic universe," which is the broadest view one can take. How the library catalog intersects with the bibliographic universe is not stated, nor is what this means for the definition of the work. Lubetzky and Smiraglia's investigations generally use the context of the library catalog, and in Smiraglia's quantitative studies the boundaries for the work families are always inside a single catalog (even though that single catalog, WorldCat, can be an aggregation of many library catalogs). The question that isn't answered is whether there is a work family if the members of that family are not present in your catalog. Yet how we model our universe depends on having a clear answer to that question.

TWO

THE MODEL

There are various reasons to create models of the real world, mostly having to do with the difficulty of manipulating the real world directly. Architects create models of buildings they have designed, car-makers create clay models of new automobile designs, and chemists create physical models to represent molecules. Oftentimes our model of the world is not a physical model but a symbolic data model. These models are abstractions of the real world, and their resemblance to reality is conceptual rather than physical. Unlike an architect's or car-makers model, a data model doesn't physically resemble the thing we are modeling. This necessary abstraction from the real world makes the development of data models complex and prone to error. There are numerous competing techniques for the development of data models that help guide one in this difficult task. These techniques are used even by modeling experts.

Models generally begin with a macro view of the area of interest, such as growth plans for a city. They place the subject of the model in context and state general goals. The next step is often articulation of use cases. Use cases can be more or less specific, but they should state in clear terms what functionality the data in the model must support. The use case for a car is that a person can get into it and drive it from one place to another. Because one might drive a car after dark, it has to have lights that one can turn on that illuminate the road well enough for traveling. There must be a steering mechanism so the driver can turn the car in needed directions. Only when this type of functionality is articulated does the design team then get down to the details of implementation. In data models, the macro level is the enterprise. If the enterprise is large and complex, more than one system may be needed to serve all of its needs, and therefore sub-units with distinct boundaries become the area being modeled. The overall goals of the enterprise ("build cars and sell a lot of them") are the context for the model of a data system that serves all or some portion of the enterprise.

SHORT HISTORY OF DATA MODELS

We can credit libraries with developing some of the earliest data models with the development of the card catalog. Card catalogs were indeed "paper machines," as Markus Krajewski (2011) calls them, with interchangeable parts and a predictable retrieval method. The punched card had essentially the same functionality as a manual card file, only it could be run through a machine process that acted on the information encoded on the cards. Punched cards had limited capabilities because they only held eighty (actually seventy-two after eight were dedicated to sequencing) character positions.

The next advance was the ability to store the previously encoded punch-card data inside the computer itself. As computers became more powerful, the limitation of seventy-two characters per line was lifted, and we got an automated spreadsheet that looked not unlike the ledger book of olden days. If you are accustomed to working with spreadsheets, you may be familiar with data that has a form like this:

NAME	STREET	CITY	STATE	ZIP
John Smith	123 Main St.	Anytown	New York	10101
Mary Jones	33 High Road	Sometown	California	93003
Jane Doe	77 Lower Road	Anytown	New York	10101
James Roe	989 Norton Pl	Anytown	New York	10102

Spreadsheets are called "flat file" technology because they are simply a list of entries, one after the other, in a single file. You can search spreadsheets, sort them, and extract selected data from them. However, once the amount of data becomes very large—as would be needed for banking or to manage a large warehouse—the spreadsheet technology is not efficient enough to produce results in a short enough amount of time to make use of the data of the enterprise "in real time." If you don't want to have to wait overnight to get an answer to your query, you need better technology.

Flat files can become very bulky with repeated data. For example, if you have a list of customers and the products they have purchased, you quickly get a large file where some data is represented many times. If a customer buys more than one product, you need to list the customer again for each product purchased.

NAME	STREET	CITY	STATE	ZIP	PRODUCT	QTY.
John Smith	123 Main St.	Anytown	New York	10101	X12	2
John Smith	123 Main St.	Anytown	New York	10101	X13	1
Mary Jones	33 High Road	Sometown	California	93003	X12	1
Mary Jones	33 High Road	Sometown	California	93003	P38	6

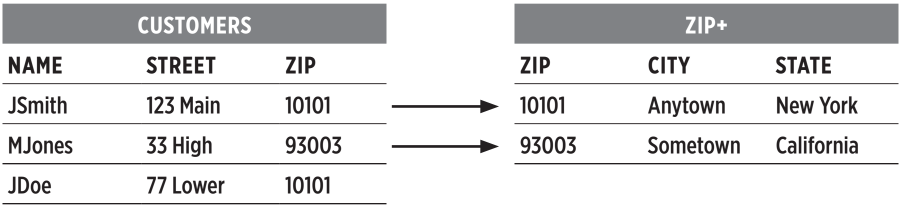
Every repeated element requires an entire new entry in the table. You can see how a file could grow quickly in size. The solution, at least the solution in the last decades of the twentieth century, would be to use a "database management system" rather than a spreadsheet. Early database management systems used a hierarchical model that could query particular paths in order to arrive at results. Like the classified library shelving system, these hierarchies forced designers to provide one and only one place for each information unit, which naturally cut off some possible data combinations at the same time that it facilitated others. In our example above, the model would need either to store customers in a hierarchy under products, or products under customers. Neither would be ideal, and there would still be repetition at the lower levels of the hierarchy. By the 1970s a new type of database management system was developed that was much more flexible than the hierarchical system: it was called a "relational database management system," or RDBMS.

The primary goals of a relational database are to eliminate duplication of the same information in the database, and to create relationships among bits of information such that it would be possible to approach the data from almost any starting point and still retrieve what you need. A relational analysis of the first spreadsheet shown above would begin by noting the duplication in the city, state, and zip code columns. That could then be designed as seen in Figure 2.1.

Each separate entry in a relational database is called a table, and figure 2.1 shows a mock-up of a database design based on the spread sheet, but now with two tables.

FIGURE 2.1

Data redesigned as two database tables



CUSTOMERS		
NAME	STREET	ZIP
JSmith	123 Main	10101
MJones	33 High	93003
JDoe	77 Lower	10101

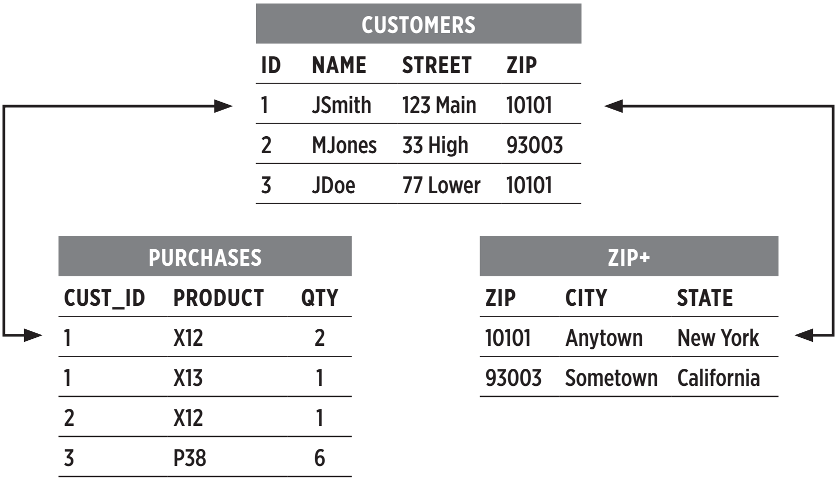
ZIP+		
ZIP	CITY	STATE
10101	Anytown	New York
93003	Sometown	California

There is still duplication here, within the city, state, zip-code table. The three columns for city, state, and zip code have a built-in relationship: the same zip code is always related to the same city and state, but the same city and state can have multiple zip codes. Therefore, the zip code can be considered a "key" for the city and state, and those can be placed in a separate table.

The purchase information related to customers becomes an additional set of tables that have relationships with the customer information. The logical database design therefore becomes something like in figure 2.2, although actual designs are generally much more complex.

FIGURE 2.2

Data redesigned as three database tables



CUSTOMERS			
ID	NAME	STREET	ZIP
1	JSmith	123 Main	10101
2	MJones	33 High	93003
3	JDoe	77 Lower	10101

PURCHASES		
CUST_ID	PRODUCT	QTY
1	X12	2
1	X13	1
2	X12	1
3	P38	6

ZIP+		
ZIP	CITY	STATE
10101	Anytown	New York
93003	Sometown	California

This process of analysis of the data to eliminate duplication is called "normalization." Normalization is generally the second or third step in a multistep analysis. This analysis might use a technique called "entity-relation modeling." Imagine that you work in a highly complex enterprise that is planning to computerize its operations. You have hundreds of employees in offices that each manage the data for a different function of the enterprise, such as manufacturing, purchasing, sales, and personnel. You wish to integrate all of these so that each office has access to the information it needs, and the data moves through the workflow without being duplicated (or lost). You ideally don't begin by tossing in all of your spreadsheets and paper files and beginning a normalization of your data. Instead, your model begins with a macro view that would make sense to management and nontechnical employees. From that you move into more detail, finally looking at individual data elements and the capacity of the actual database management system that you will employ.

Entity-relation (E-R) modeling is a technique developed in the 1970s and 80s to describe the elements of the data universe that you wish to organize and their relationships to each other. The technique was developed specifically to aid in the design of relational databases, although it has value in other data mapping situations as well. The first step in E-R modeling provides a conceptual view of your data. A *conceptual model* serves to define the data "things" (entities) that your business works with, and how they relate to each other in the bigger picture. Once the conceptual model is well understood, the process moves on to the creation of a *logical model*. This is where you complete the list of data elements, and define what type of data value will be stored for each data element (text, date, currency). This is the phase where you discover duplicate data coming in from different functions and perform normalization on the data. As you can imagine, the resulting picture can be very complex, and may vary considerably from the conceptual model. A *physical model* is the final step in database design, and may be combined with the logical model into a single step. The physical model should reflect the actual database structure and contents.

The "conceptual model" of E-R modeling is not conceptual in the philosophical or cognitive science definition of "conceptual," but is a first step toward development of an actual data processing system. In philosophy or cognitive science, concepts can be imprecise, changeable over time and within different contexts, and probably could not be accurately developed into anything as mechanical as a database management system. In E-R modeling, the concepts define the main categories of things that must be described in the data in order to support the functional requirements of the system, and the relationships between them. Quite often the conceptual model is much simpler than the subsequent logical model.

E-R modeling is still used, as are relational databases, although in the 1990s a new model of data processing was developed, called "object-oriented" (OO). Object-oriented concepts are behind the programming languages C++ and Java, as well as being the basis for current languages like Python and Ruby. Object-oriented design makes extensive use of *classes* to gather data elements and processing routines that are shared by data types. OO classes can function as modular routines that encapsulate existing programming code, thus protecting that part of the code from changes made to the program elsewhere. A new design notation was developed to help developers who were working with OO models: the Universal Modeling Language, or UML. UML can be seen as an evolution of E-R modeling; it is possible to create E-R models using UML, but UML supports over a dozen types of modeling needs, including structure modeling, behavior or process modeling, and interaction modeling. Other than the extensive use of classes, one of the more significant differences between OO and E-R designs is that object-oriented programming and design often focuses on dynamic processes rather than static views of data. OO data is more like a factory than a finished product.

The next leap forward in data-planning and design is that brought on by the development of the Semantic Web. At this writing, the Semantic Web revolution is still in progress, and data designers are just beginning to gain experience with this new way of looking at the data we manage and share. The Semantic Web uses the concept of a web or graph of data, with the Internet as its underlying technology. The Semantic Web emphasizes growth and interconnections between data that can come from different environments. Although it is being used in business applications, the Semantic Web is oriented more toward discovery and knowledge enhancement than control. This will be covered more comprehensively in the chapter on technology.

LIBRARY DATA MODELS

Libraries have a number of functions that are served by their data systems: acquisitions and fund accounting, personnel administration, inventory control, user identification, and, of course, the library catalog. The actual function of the library catalog is where I will focus our discussion of modeling here, but before I do I want to talk about the bigger picture in libraries.

If you grab a book on data modeling, it will give you steps to take that lead from functions performed by employees all of the way to a database design that allows them to do their jobs with the help of automation. These books assume that the database that is being designed will be built. That seems like an obvious thing to bring up, after all why would you be designing a database unless you intended to build it? However, this is exactly the situation that libraries are in: libraries do not build systems, and they have only minor control over the systems that are built for them. For this reason, what few modeling exercises take place in libraries are quite different from those that we see coming from the enterprise information technology sector.

There is one aspect of library information management that overshadows all others, at least in library data theory, and that is the catalog of the library's holdings. To some extent, the catalog *is* the library, because it is itself a model, in metadata, of the essence of the library: the information it offers. The library catalog is to the library as the architect's miniature is to the real building. You would think, then, that there would be a large body of work around the model of the catalog and its implementation in technology. That is not the case, however. There is a body of work on the theory and practice of cataloging, but it is distinctly separate from any discussion of satisfying those goals in technology design. The library profession models its data, but not the system solution that uses that data. This leads to an awkward situation where the goals of cataloging may not be the same as the functions of the catalog as implemented.

Goals of the Catalog

In 1875, Charles Ammi Cutter stated the goals of the library catalog as:

1. **To enable a person to find a book of which either**
    - A. the author
    - B. the title is known
    - C. the subject

2. **To show what the library has**
    - D. by a given author
    - E. on a given subject
    - F. in a given kind of literature

3. **To assist in the choice of a book**
    - G. as to its edition (bibliographically)
    - H. as to its character (literary or topical)

Cutter defines a catalog as a "list of books which is arranged on some definite plan." He distinguishes the catalog from a bibliography in that a catalog is a "list of books in some library or collection," while a bibliography is a list of books around some other organizing principle, such as subject, place or period. To Cutter, the catalog's main goal is to be "an efficient instrument."

Cutter's list of goals could be considered a high-level set of use cases. What is not articulated here, but obviously was clear enough to him that he could develop his cataloging rules, was exactly how the catalog is to provide this functionality. There is nothing here to say how users will find an author, or what it means that the catalog will "show what the library has." Of course, Cutter was working nearly one hundred years before the concept of systems analysis was common among modelers, so to point out this shortcoming is not a criticism of the great man, but does show how modeling has changed as a concept.

In 1961, the International Conference on Cataloguing Principles (known as the "Paris Principles") gave these as the functions of the catalog:

> The catalogue should be an efficient instrument for ascertaining
>
> **2.1 whether the library contains a particular book specified by**
> (a) its author and title, *or*
> (b) if the author is not named in the book, its title alone, *or*
> (c) if the author and title are inappropriate or insufficient for identification, a suitable substitute for the title; and
>
> **2.2** (a) which works by a particular author and
> (b) which editions of a particular work are in the library.

The similarities between these functions and Cutter's goals are striking. The 1961 Paris Principles, written ninety years after Cutter, change his wording somewhat but have essentially the same meaning: the purpose of the catalog is to provide an identity for the resources in the library by a small set of known qualities, such as the author of the work, or the title, that a catalog user can employ to discover if the library has a copy of the item sought. There is no question that these principles adhere to the distinction between bibliography and a library catalog that was defined by Cutter. The library catalog is a sophisticated finding aid. Unspoken but implicit is that users can also discover what a library does not have because it will not appear in the catalog.

Significantly, the Paris Principles do not mention subject or genre access, both of which were included in Cutter's requirements for the catalog. Cutter's rules devoted fifteen pages to describing subject access, less than ten percent of the total, although Cutter conceded the exact subject description methodology to sources external to his cataloging rules. The scope of the Paris Principles was limited to entries by authors' names and titles (and the latter only when author entry was for some reason not available). In this sense, the Paris Principles can be seen as an updated version of Panizzi's rules, which preceded them by over a century. Both require author entry where the author name is available, define title entry for those works without authors, and deal with the form of the author's name and a set of exceptions. And no more. These principles comprise only a portion what one generally considers a complete catalog for users.

The most recent version of these principles was issued in 2009, nearly 50 years after the original Paris Principles and over 125 years since Cutter laid out his goals.

> 4. Objectives and Functions of the Catalogue
>
> The catalogue should be an effective and efficient instrument that enables a user:
>
> 4.1 to find bibliographic resource in a collection as the result of a search using attributes or relationships of the resources:
>
> 4.1.1. to **find** a single resource
>
> 4.1.2. to **find** sets of resources representing
> all resources belonging to the same work
> all resources embodying the same expression
> all resources exemplifying the same manifestation
> all resources associated with a given person, family, or corporate body
> all resources on a given subject
> all resources defined by other criteria (language, place of publication, publication date, content type carrier type, etc.), usually as a secondary limiting of a search result;
>
> 4.2. to **identify** a bibliographic resource or agent . . . ;
>
> 4.3. to **select** a bibliographic resource that is appropriate to the user's needs . . . ;
>
> 4.4. to **acquire** or **obtain** access to an item described . . . ;
>
> 4.5. to **navigate** within a catalog and beyond . . .

The change here is significant, and is entirely due to the fact that this version of the Paris Principles follows (temporally and philosophically) the entities described in the Functional Requirements for Bibliographic Records (FRBR). The "book" has been replaced with the FRBR bibliographic entities "work, expression, manifestation," even though those are not defined anywhere in this version of the document. Subjects return in this edition, although as we will see they are actually given short shrift in the FRBR model. The principles also include an interesting smattering of "additional access points" that don't appear to have any particular theoretical basis, such as "bibliographic record identifiers," "language of expression," and "content type." None of these are defined or explained, and the suggestion is that these may be used as a "limiting device for a search." Such devices are found in some online catalogs, but there doesn't appear to be a philosophical basis for their existence in the Principles.

Although user-seeking behavior was implied in previous versions (users "found" in Cutter, and "ascertained" in 1961), this 2009 version includes the user tasks defined in FRBR: find, identify, select, and obtain. It also adds the concept of sets, an acknowledgment of what the introduction to that document refers to as the "OPAC (Online Public Access Catalogues)" technology in wide use. The term *set* refers to the technology of retrieval that, based on a query, returns a selected group of entries that meet the criteria of the query. This may seem to be a small change, yet in fact the change from the linear, alphabetic (or "dictionary" catalog, as Cutter would have it) is a change of great import that is hardly acknowledged in the practice of bibliography.

This is undoubtedly not the first time that you will have seen Cutter's rules, because his rules for a dictionary catalog continue to be widely quoted as the basis for library cataloging today. To some this is proof that there are strong, underlying purposes to the catalog that have withstood the test of time. On the other hand, it seems unlikely that Cutter's objects of the catalog are sufficient for today's information seekers.

In 1875, when Cutter's rules were published, a very large library was one that held 500,000 volumes, and most libraries were much smaller. Information seeking in a collection of that size is clearly different from information seeking in a library holding millions of books and tens of thousands of motion pictures and pieces of recorded music, and also provides integrated access to tens or hundreds of millions of indexed articles. The library user of 1875 was of course also significantly different from the library user of the twenty-first century. Some of the arguments launched against Panizzi's plan to create a detailed catalog of books in the British Museum Catalog were that any reasonably educated gentleman came to the library knowing exactly what he sought, and therefore the additional information in the catalog was unnecessary.

In the midst of all of this orthodoxy around library catalog goals, some interesting ideas came from outside of the cataloging community. One particularly unorthodox thinker was Professor Patrick Wilson, and his exposition of a concept he called "two kinds of power."

Patrick Wilson's *Two Kinds of Power,* published in 1968, and introduced in chapter 1, is a book that is often mentioned in library literature but whose message does not seem to have disseminated through library and cataloging thinking. If it had, our catalogs today might have a very different character. A professor of Library Science at the University of California at Berkeley, Wilson's background was in philosophy, and his book took a distinctly philosophical approach to the question he posed, which most likely limited its effect on the practical world of librarianship. Because he approached his argument from all points of view, argued for and against, and did not derive any conclusions that could be implemented, there would need to be a rather long road from Wilson's philosophy to actual cataloging code.

Wilson takes up the question of the goals of what he calls "bibliography," albeit applied to the bibliographical function of the library catalog. The message in the book, as I read it, is fairly straightforward once all of Wilson's points and counterpoints are contemplated. He begins by stating something that seems obvious but is also generally missing from cataloging theory, which is that people read for a purpose, and that they come to the library looking for the best text (Wilson limits his argument to texts) for their purpose. This user need was not included in Cutter's description of the catalog as an "efficient instrument." By Wilson's definition, Cutter (and the international principles that followed) dealt only with one catalog function: "bibliographic control." Wilson suggests that in fact there are two such functions, which he calls "powers": the first is the evaluatively neutral description of books, which was first defined by Cutter and is the role of descriptive cataloging, called "bibliographic control"; the second is the appraisal of texts, which facilitates the exploitation of the texts by the reader. This has traditionally been limited to the realm of scholarly bibliography or of "recommender" services.

This definition pits the library catalog against the tradition of bibliography, the latter being an analysis of the resources on a topic, organized in terms of the potential exploitation of the text: general works, foundational works, or works organized by school of thought. These address what he sees as the user's goal, which is "the ability to make the best use of a body of writings." The second power is, in Wilson's view, the superior capability. He describes descriptive control somewhat sarcastically as "an ability to line up a population of writings in any arbitrary order, and make the population march to one's command" (Wilson 1968).

If one accepts Wilson's statement that users wish to find the text that best suits their need, it would be hard to argue that libraries should not be trying to present the best texts to users. This, however, goes counter to the stated goal of the library catalog as that of bibliographic control, and when the topic of "best" is broached, one finds an element of neutrality fundamentalism that pervades some library thinking. This is of course irreconcilable with the fact that some of these same institutions pride themselves on their "readers' services" that help readers find exactly the right book for them. The popularity of the readers' advisory books of Nancy Pearl and social networks like Goodreads, where users share their evaluations of texts, show that there is a great interest on the part of library users and other readers to be pointed to "good books." How users or reference librarians are supposed to identify the right books for them in a catalog that treats all resources neutrally is not addressed by cataloging theory.

Wilson's analysis presages the search and retrieval capabilities of Internet search engines like Google, Bing, and Yahoo. He also writes that power of bibliography is greatest if it extends over the entire bibliographic universe, not just a single selection (one universal library as opposed to the local collection); that the user is better served the fewer retrieved items must be reviewed before satisfying the user's request (as in targeted ranking); and that direct access to the text is a greater power than restrictive use (open access).

Due to the philosophical nature of the book, one has to tease out these brilliant ideas; they are not laid out as headlines or clear conclusions. Yet in the text Wilson may have laid out a new direction for libraries decades before those same principles were discovered by Internet entrepreneurs using new technologies. Imagine if Internet search engines had the same goals as library catalogs and designed their products to cater to only those users who came to the search box knowing either the title or the author of the document they were seeking. Not only is that not the goal of these systems, but they do not even assume that the search engine user is even aware that any documents satisfying their need exist. This is the difference between seeing information space as a finite set of items on a shelf, versus an ever-changing, nearly infinite set of unknowns. The setting of boundaries around the library collection is one of the tenets of library cataloging goals—to define exactly what the library does and does not have. Although such an inventory is clearly needed, it is a mistake to also assume that this inventory and its boundaries is what interests today's information seeker. Cutter's goals for the catalog were written at a time when the information world was still contained within a relatively small number of published texts, and even fewer of those were available to information seekers at any given time and place. Although users may have entered a library seeking information, the only possible way to pose the question at that time was "do you have a book on?" A person facing the nearly blank Google home page is free to ask "is there anything out there about my topic?" without having to predetermine the limitations that may exist in the information resources available on that topic. Failure in these systems is undoubtedly a common occurrence, but the failure in the library catalog comes about by limiting the questions the user can ask, and limiting, by design, the utility of the response.

The Larger Context

I began this section saying that a model begins at a macro level. A model that covers the library catalog and the user interaction with that catalog is clearly already focused on a small slice of both the library's functioning and on the activities of the user. You could argue that this is a self-contained unit that is well-defined, but it is easy to prove otherwise.

Many library management functions revolve around the resources owned or controlled by the library, such as acquisitions and collection development. This is the basis behind the idea of the "integrated library system," or ILS. There is a workflow not unlike that of a business where resources are selected for purchase, added to budgets, paid out as expenses, received as goods, processed, and stored. Prior to the integration of these workflows, separate systems had their own separate databases, and these often carried information duplicating that of other areas of the library's management. The integrated system brought at least some of these data stores together, resulting in less duplication and greater efficiency. Given this, it would seem only sensible that the catalog would be studied within the entire library workflow. If it were, there would be goals like:

- Show what the library has on order.
- Allow the input of minimum records for items under review.
- Keep a record of requested inter-library loans for future purchasing decisions.
- Manage statistics about use and co-use of materials.

The catalog that is described in the cataloging rules and in the models of catalog data does not acknowledge the existence of library management functions.

Not that the library cataloging rules would necessarily be the correct place for information like account management, circulation statistics, or serials receipts, but the failure to place the catalog in the larger context means that there isn't a place in the model for the interaction of these necessarily connected functions.

At the same time, look at any request for proposal for an integrated library system, and neither cataloging goals nor users receive much attention, just as the needs of library systems are not addressed in cataloging rules. This split between the goals of the user catalog and the goals of the library as a place of business is also visible in the standards environment. Technical standards are developed by the National Information Standards Organization (NISO). There are standards for circulation data, for statistics, for automated data retrieval, for recording licenses, for serials management, and a number of identifiers. The base format for recording the catalog data is also a NISO standard, but the specific format used is managed elsewhere, at the Library of Congress. Although NISO has a work area called "Discovery to Delivery" this area does not include any direct interaction with the cataloging rules, which are developed by a separate and independent organization. NISO also does not have standards that would overlap with the library cataloging rules, nor with the goals for the catalog.

The upshot is that libraries have moved into the twenty-first century with nineteenth century user service goals, at least as far as information seeking in the library catalog is involved. Although today's systems could provide a wide variety of user services, there is no interaction between technology standards development and cataloging standards. The addition of "all resources defined by other criteria (language, place of publication, publication date, content type, carrier type, etc.), usually as a secondary limiting of a search result"; to the 2009 International Catalog Principles is in its way proof of how distant cataloging is from technology design. It is ironic that almost none of the "other criteria" that are actually used in systems and that allow limiting by such come from the cataloging rules. In practice, these systems make use of the fields in the machine-readable record standard that the cataloging rules do not describe, much less mandate, as catalog information.. The information is usable in this way precisely because it is coded information designed for use by computers, not as visible information for human users.

The User in the Model

The catalog goals also provide a very narrow view of the user's interaction with the library. We will see this again when we look more closely at FRBR, even though its "find, identify, select, obtain" appears to be broader than Cutter's "find a book of which ______ is known."

First, what do the goals tell us about the user? The first thing is that some users come to the library looking for a known item. This is indisputable. Whether they really know what they are looking for is another question, and we have seen that online systems use technologies like query completion and "did you mean . . . ?" because this is a common problem.

Next we have the user finding sets that represent logical groupings, such as all of the works of a single author. Once again, it appears that users need to come to the library with this information, because nowhere is it stated that the system should offer these sets through some other mechanism. In fact, many systems do, by allowing users to click on a linked heading and retrieve everything associated with that heading, but because there has been no definition of the functions of the catalog, this isn't something we can assume.

What is key about these goals, however, is that they limit themselves to the user finding an entry in the catalog (albeit FRBR goes on to having the user obtain the item represented there). A study done by the University of Minnesota Libraries in 2006 (UMN 2006) took a much broader view of their users and user needs. They asked their faculty and graduate student users questions like "Where do you work when you are conducting research?" "How do you share source materials?" Just these two questions already reveal quite a lot: the librarians are not assuming that one conducts research in the library, and acknowledge that many people work in teams or groups that share resources among themselves. They also asked about library use: how often do these users visit the physical library, and how often do they visit the library web site, and what do they do there?

The authors of the report (who modestly remain anonymous) then developed a model to describe what they had learned. They borrowed the core of their model from a humanities researcher, John Unsworth, who described the primitives of humanities research as *discover, gather, create,* and *share.* Of these, only discover is usually seen as directly related to the library, and many, perhaps even most, discoveries take place outside of the library catalog. Yet if your view is that libraries support the research function, then all of these primitives could possibly have some interaction with the library. The *share* primitive includes teaching, and the library may be directly connected to the course management system such that course materials are shared through library functions. The *gather* function includes acquiring and organizing, which might mean library support of bibliographic tools. And the *create* function could be supported through shared annotation tools, which could be especially important in those disciplines where research is done through collaborative work.

Libraries have recently begun to take a role in the storage and sharing of research data. Oftentimes institutional repositories for the storage and delivery of research papers written by the faculty of an institution are also managed by the library. In many of these, the library users are not using the library to find materials, but are instead providing resources that the library will manage. Even if those materials do not go through the same cataloging process as more traditional library holdings, it would be hard to argue that they should not be equally available for searching.

Although libraries have taken on many of these functions, and some of them do interact directly with the library catalog, they are not included in the objectives and functions of the catalog listed in the International Cataloguing Principles. Those principles expound an unfortunately narrow view of the catalog, isolated from the user services that modern libraries are endeavoring to provide.

The objectives of the catalog say little about the users themselves and why they would come to the library seeking resources. Wilson addresses this in *Two Kinds of Power* when he states that it is obvious that people are looking for the best book for their needs or desires. I characterize the traditional library catalog goals as beginning with "a man walks up to a catalog. . . ." Nothing before or after the interaction with the catalog is under consideration. What those objectives do is put a tight fence around the freedom of a person to then ask the question that would satisfy their need. Because of how the catalog is designed, the question "Do you have a good book on dogs?" is not going to result in an answer, although it is, in Wilson's view, simply illogical to think that someone would ask the question "Do you have a book on dogs that I will find insufficient for my needs?" It also seems unlikely that someone would come looking for "a list of books on dogs where there isn't enough information for me to determine which meet my needs."

From this view it becomes clear that the objectives of the catalog are not stated in terms of satisfying the user's query, but to delineate what queries can be made, and to manage the expectations for what responses will be experienced. Library instruction in universities teaches users what they can—and cannot—ask of various resources available through the library, precisely because none of them can answer the question: "Do you have what I need?" Bibliographic research is often a tedious and unsatisfying task. Course syllabi and best-seller lists exist precisely because this is so.

The question comes down to the moral role of the library. As historian Dee Garrison pointed out in her book *Apostles of Culture* (1979), in the early twentieth century libraries saw their role as uplifting the ignorant masses by providing them with “good books.” The library as neutral keeper of the “stuff” came about later, but arguments for moral education still come forward around allowing comic books into the library and providing unfiltered access to the Internet. Thus the debate over whether the library provides what the user does want, or provides what the user should want, continues. In the area of the catalog, however, the solution appears to be to provide only discernible facts about resources.

Patrick Wilson later addressed a topic of more specific interest to catalog theory, and that is the identification of the library resources that represent that same “literary unit.” Lubetzky referred to this as cataloging’s “second objective.” Whereas it would be a notable expansion of bibliographic description for libraries to attempt to fulfill Wilson’s second kind of power, library catalogs already include some bibliographic relationships between the items in the library and beyond. Both Cutter and the original Paris Principles include the identification of the edition of a book as a basic function of the catalog. This goes beyond the mere description of individual items to adding certain bibliographic relationships between items where appropriate. Unlike Wilson’s second kind of power, this idea has actually gained some traction.

In any functional model it is necessary to define a clear scope of operation: what are the boundaries within which this model will operate? Cutter was clear in his objectives that his rules applied to the catalog of a library, and served to show what books the library did hold, and, by deduction, what books it did not. He had a clear universe for his rules, and it was the single library. The challenge to the neat, finite boundaries of single library’s walls came about twenty-five years later when the Library of Congress began distributing sets of catalog cards to libraries across the United States. With this seemingly small gesture, the closed walls of the individual library catalog were breached.

Since then libraries have had to seek a balance between the efficiency of bibliographic data sharing and the desire to serve their unique population of users. The development of combined catalogs of the holdings of multiple libraries, including the massive WorldCat database containing the holdings of tens of thousands of libraries, makes the creation of a boundary for a bibliographic data model all the more elusive. Creating a viable model when such a key question is unresolved is difficult if not impossible.

THREE

THE TECHNOLOGY

Today when we say "technology" it is often shorthand for "computer technology." The Technology section of a newspaper reports on Silicon Valley news and reviews the latest consumer gadgets that are powered by bits and bytes. Of course this is not the only technology in our lives, but it is the one that defines our modern age. A century and a half ago, the defining technology was electricity and all things electric. The light bulb was literally the bright idea of the day. Today we have LED light bulbs that we can control with a smartphone app, turning on the lights when we are still on our way home, or creating a romantic atmosphere by changing the color and intensity of the light at the touch of a screen.

If we move back in time we see ages defined by their technological innovations: steam power, water power, or the precision use of metals that made it possible to create accurate timepieces and to automate the production of fine cloth. We can go back to the printing press, clearly a defining technology for all that came after it. Printing technology depended both on innovations with metals and also on the development of paper-making techniques that greatly improved on previous writing surfaces, like sheepskin, papyrus, wax, clay, and stone.

Basically, it's technology all of the way back—back to fire and the first stone axes. We naturally take for granted the technologies that precede our own age, and we marvel at the ones that are new.

Libraries of course have been technology-based from the beginning of their history. The earliest libraries that we know of were furnished with writings in the form of scrolls. Medieval libraries held bound manuscripts. The big leap forward was the Gutenberg revolution and the concomitant increase in the production of copies of texts. The number of books not only increased but they also become more affordable as a result of their abundance. Other technologies also had effects on libraries, such as the aforementioned development of electric lighting, which reduced the threat of fire and allowed readers to make use of the library outside of daylight hours.

In the eighteenth and nineteenth centuries, not only were more copies of books produced than ever before, but the numbers of new writings and new editions also grew. Library holdings thus increased as well, which led to difficulties in keeping up with an inventory of the items held by the library. Today we assume that every library has a catalog, but even in the 1800s some libraries had no actual record of their holdings or relied on a brief author list. Much "finding" done in libraries at the time relied on the memory of the librarian. Charles Ammi Cutter, writing about the catalog of the Harvard College Library in 1869, took pity on the librarian overseeing a collection of 20,000 books without a proper catalog, who had to attempt to answer subject-based queries using only his own knowledge of the content of the collection.

The library catalog technology of Cutter's day was a printed book. Printed book catalogs had the same advantages as books themselves: they could be produced in multiple copies and were highly portable. A library could give a copy of its catalog to another library, thus making it possible for users to discover, at a distance, that a library had the item sought. The disadvantages of the printed book catalog, however, became more serious as library collections grew and the rate of growth increased. A library catalog needed near-constant updating. Yet the time required to produce a printed book catalog in an era in which printing required that each page be typeset meant that the printed catalog could be seriously out of date as it came off the printing press. Updating such a catalog meant reprinting it in its entirely, or staving off an expensive new edition by producing supplementary volumes of newly acquired works, which then made searching quite tedious.

In the mid-1800s the library card catalog was already winning hearts and minds. Cutter attributed the development of the card catalog to Ezra Abbot, head of the Harvard College Library, in 1861 (Cutter 1869). Although neither the book catalog nor the card catalog meets all needs as efficiently as one would desire, the card catalog had already proven itself as an up-to-date instrument for library users and librarians alike. German professor Markus Krajewsky, in his book on the history of card files, *Paper Machines* (2011), shows that cards on paper slips had been used in earlier times, in particular by the early bibliographers and encyclopedists who needed to create an ordered presentation of a large number of individual entries. It was libraries, however, that demonstrated how useful and flexible the card catalog could be.

Cards were lauded by Melvil Dewey in his introduction to early editions of his Decimal Classification, although his classification and "relativ index" in no way required the use of a card system. However, the "Co-Operation Committee" of the newly formed American Library Association announced its decision on the standardization of the catalog card in *Library Journal* in 1877; not coincidentally, Dewey's library service company, The Library Bureau, founded in 1876, was poised to provide the cards to libraries at a cost lower than custom-produced card stock. The Library Bureau soon branched out into the provision of catalog furniture and a variety of card-based products for a growing business records market. In fact, before long providing cards to libraries was only a small portion of The Library Bureau's revenue as businesses and other enterprises in the United States and Europe turned to card systems for record-keeping. Krajewski considers these card systems the early precursors of the computerized database because of the way that they atomized data into manipulatable units, and also allowed the reordering of the data for different purposes.

It should be obvious that both the book catalog and the card catalog were themselves technologies, each with different affordances. They also were affected by related technological developments, such as changes in printing technologies. The typewriter brought greater uniformity to the card catalog than even the neatest "library hand" could, and undoubtedly increased the amount of information that one could squeeze into the approximate 3" x 5" surface. When the Library of Congress developed printed card sets using the ALA standard size and offered them for sale starting in 1902, the use of the card catalog in US libraries was solidified.

After Melvil Dewey, the person who had the greatest effect on library technology was Henriette Avram, creator of the Machine Readable Cataloging (MARC) format. This was not only an innovation in terms of library technology, it was generally innovative in terms of the computing capability of the time. In the mid-1960s, when MARC was under development, computer capabilities for handling textual data were very crude. To get an idea of what I mean, look at the mailing label on any of your magazines. You will see upper-case characters only, limited field sizes, and often a lack of punctuation beyond perhaps a hash mark for apartment numbers. This is what all data looked like in 1965. However, libraries needed to represent actual document titles and author names, and languages other than English. This meant that the library data record needed to have variable length fields, full punctuation, and diacritical marks. Avram delivered a standard that was definitely ahead of its time.

Although the primary focus of the standard was to automate the printing of cards for the Library of Congress's card service, Avram worked with staff at Library of Congress and other libraries involved in the project to leverage the MARC record for other uses, such as the local printing of "new books" lists. To make these possible the standard included non-text fields (in MARC known as "fixed fields") that could be easily used by simple sort routines. The idea that the catalog could be created as a computerized, online access system from such records was still a decade away, but Librarian of Congress L. Quincy Mumford announced in his foreword to Avram's 1968 document *The MARC Pilot Project* that MARC records would be distributed beginning in that year, and that this "should facilitate the development of automation throughout the entire library community." And it did.

Melvil Dewey did not anticipate the availability of the Library of Congress printed card service when he proposed the standardization of the library catalog card, yet it was precisely that standardization that made it possible for libraries across America to add LC printed cards to their catalogs. Likewise, Henriette Avram did not anticipate the creation of the computerized online catalog during her early work on the MARC format, but it was the existence of years of library cataloging in a machine-readable form that made the OPAC a possibility.

The next development in library catalog technology was the creation of that computerized catalog. It would be great to be able to say that the move from the card catalog to the online catalog was done mainly with the library user's needs in mind. That wasn't my experience working on the University of California's online catalog in the early 1980s. The primary motivators for that catalog were the need to share information about library holdings across the entire state university system (and the associated cost savings), and to move away from the expense and inefficiency of card production and the maintenance of very large card catalogs. At the time that the library developed the first union catalog, which was generated from less than a half dozen years of MARC records created on the systems provided by the Ohio College Library Center (later known solely as OCLC) and the Research Libraries' Group's RLIN system, the larger libraries in the University of California systems were running from 100,000 to 150,000 cards behind filing into their massive card catalogs. This meant that cards entered the catalog about three months after the book was cataloged and shelved. For a major research library, having a catalog that was three months out of date, and only promising to get worse as library staffing decreased due to budget cuts, made the online catalog solution a necessity.

We, and by "we" I mean all of us in library technology during this time, created those first systems using the data we had, not the data we would have liked to have. The MARC records that we worked with were in essence the by-product of card production. And now, some thirty-five years later, we are still using much the same data even though information technology has changed greatly during that time, potentially affording us many opportunities for innovation. Quite possibly the greatest mistake made in the last two to three decades was failing to create a new data standard that would be more suited to modern technology and less an imitation of the library card in machine-readable form. The MARC record, designed as a format to carry bibliographic data to the printer, was hardly suited to database storage and manipulation. That doesn't mean that databases couldn't be created, and to be sure all online catalogs have made use of database technology of some type to provide search and display capabilities, but it is far from ideal from an information technology standpoint.

The real problem is the mismatch of the models between the carefully groomed text of the catalog entry and the inherent functionality of the database management system. The catalog data was designed to be encountered in an alphabetical sequence of full headings, read as strings from left to right; strings such as "Tolkien, J. R. R. (John Ronald Reuel), 1892–1973" or "Tonkin, Gulf of, Region—Commerce—History—Congresses." Following the catalog model of which Charles Cutter was a primary proponent, the headings for authors, titles, and subjects are designed to be filed together in alphabetical order in a "dictionary catalog."

Database management systems, which are essential to permit efficient searching of large amounts of data, work on an entirely different principle from the sequential file. A database management system is able to perform what is called “random access,” which is the ability to go seemingly directly to the entry or entries that match the query. (The actual internal mechanism of this access is quite operationally complex.) These entries are then “retrieved,” which means that they are pulled from the database as a set. A set of retrieved entries may be from radically different areas of the alphabetical sequence, and once retrieved are no longer in the context intended by the alphabetical catalog.

Database management systems include the ability to treat each word in a sentence or string as a separate searchable unit. This has been accepted as a positive development by searchers, and is now such a common feature of searching that today most do not realize that it was a novelty to their elders. No longer does a search have to begin at the same left-anchored entry determined by the library cataloging rules; no longer does the user need to know to search “Tonkin, Gulf of . . .” and not “Gulf of Tonkin.” Oddly enough, in spite of the overwhelming use of keyword searching in library catalogs, which has been shown to be preferred by users even when a left-anchored string search was also available, library cataloging has continued its focus on headings designed for discovery via an alphabetical sequence. The entire basis of the discovery mechanism addressed by the cataloging rules has been rendered moot in the design of online catalogs, and the basic functioning of the online catalog does not implement the intended model of the card catalog. Parallel to the oft-voiced complaint that systems developers simply did not understand the intention of the catalog, the misunderstanding actually goes both ways: significant difference in retrieval methods, that is, sequential discovery on headings versus set retrieval on keywords, did not lead to any adaptation of cataloging output to facilitate the goals of the catalog in the new computerized environment. Library systems remain at this impasse, some three-and-a-half decades into the history of the online catalog. The reasons for this are complex and have both social and economic components.

It is not easy to explain why change was not made at this point in our technology history, but at least one of the factors was the failure to understand that cataloging is a response to technical possibilities. Whether the catalog is a book, a card file, or an online system, it can only be implemented as an available technology. Unlike most other communities, the library community continues to develop some key data standards that it claims are “technology neutral.” It is, however, obvious that any data created today will be processed by computers, will be managed by database software, will be searched using database search capabilities, and will be accessed by users over a computer network. One ignores this technology at great peril.

THE PRESENT AND FUTURE

We have made the error in the past of moving to new technologies without examining the fit between our data and the new technology. A perfect example of this is the development of an XML version of the MARC record. There are indeed similarities between MARC and XML, primarily that both can be used to mark up or encode machine-readable documents. Both can also encode structured data, although the MARC use of fixed fields is less flexible than XML, which allows variable-length data throughout. MARCXML was developed as a pure serialization of the MARC format. "Serialization" means that the data encoding of MARC was translated directly to XML without any related transformation of the data itself. Although this produced a record that could be managed with XML-aware software, it did nothing to improve the kind of data that could be conveyed in library bibliographic records. It also did nothing to address some of the limitations of the MARC record. The MARCXML standard is kept one-to-one with the original MARC record, with the single exception that field and record sizes are not enforced. (MARC fields are limited to a four-character length, thus to 9,999 bytes; the record itself cannot exceed 99,999 bytes.) But the limitation on the number of subfields to a field remains, even though there are fields that have no open subfields available for expansion. Other inconveniences also remain, such as the non-repeatability of the MARC fixed field information, which then forces some repeatable elements like languages and dates to be coded in more than one field to accommodate repeatability. MARCXML was never allowed to develop as its own technology, and therefore did not present a change. Library data in XML, rather than in MARCXML, could have represented a real change in capabilities. It might also have provided a better transition to new technology than we now have, because we could have resolved some of the more awkward elements of MARC over a decade or more, with a gradual update to the library systems that use this data. Today we either have to carry those practices on to our future data, or we need to make a great leap forward and break with our past.

We missed the XML boat, but now some are hoping to get on board the latest ship sailing by: the Semantic Web and its base technology, the Resource Description Framework (RDF). It should be noted that there is one other data technology development that could have been considered between XML and

RDF, that of object-oriented design (OO). By the early 1990s, when the FRBR Study Group was being formed, relational technology was no longer new and object-oriented technology was taking its place in many implementations. Programming languages like Java and Python are object-oriented, and data and databases can also be "OO." Library data is leap-frogging over this technology, or it will if it adopts RDF for its data, as it appears it might.

Unlike most of the data models that preceded it, from entity-relation to object-oriented, RDF does not arise from the world of business that prompted our previous technology upgrades. The Semantic Web, as the name implies, comes out of web technology. This is a significant difference from, for example, database technologies, because the web is an open platform and is the place where we put publicly accessible data, whereas databases are private and closed, housed within enterprises and often highly controlled in terms of access. This means that many of the design assumptions that drive the Semantic Web standards are quite different from those encountered in business data processing.

First, let's look at where the Semantic Web comes from and what is meant by "semantic." The Semantic Web comes out of a combination of web technology, with linking and identifying as primary requisites, and the artificial intelligence (AI) community, with smart "bots" as its goal. Where most of us read the term "semantic" as meaning "meaning," in the AI world "semantic" refers to a computable axiom, such as:

If A = B, and B = C, then A = C

Obviously, machine intelligence and human intelligence are significantly different. AI attempts to model human thinking by defining the world as information about things and rules that can be used to "understand" those things. As we know from the overly confident promises that have come out of the AI community since the dawn of computing, the world and how we humans understand it is more complex than it seems. Human intelligence is still a marvel that is unchallenged by machines, in spite of gains in such algorithm-rich areas like the game of chess.

Artificial intelligence on the World Wide Web is a more tractable problem than creating a robot that can navigate stairs, recognize human faces, and pass a Turing test, because the web is already a data abstraction with some distance from the sense-experienced world, and therefore more amenable to computation. The Semantic Web was introduced in an article in *Scientific American* in 2001 by Tim Berners-Lee, founder of the World Wide Web and director of the World Wide Web Consortium, and his associates James Hendler and Ora Lassila. The article told the story of a helpful bot that could find an available doctor, check your calendar, and make an appointment that fit into your schedule. Creating such technology over the web would require much less effort than creating this technology as a stand-alone system; the web already had solved the problem of a large distributed system capable of handling heterogeneous data and billions of users. The trick was to include in the web the kind of coding that would allow data to be used alongside the current web of documents and media files.

The technology to achieve this is all based on the Resource Description Format (RDF), which itself is a deceptively simple model of things and relationships that can be used to express very complex data. There are some particular aspects of RDF that are both essential and notably different from the technology that most of us have worked with during our careers. There isn't space here to fully elucidate the technology that is RDF, but some points are key to the analysis in the second part of this book. Let's begin with identifiers.

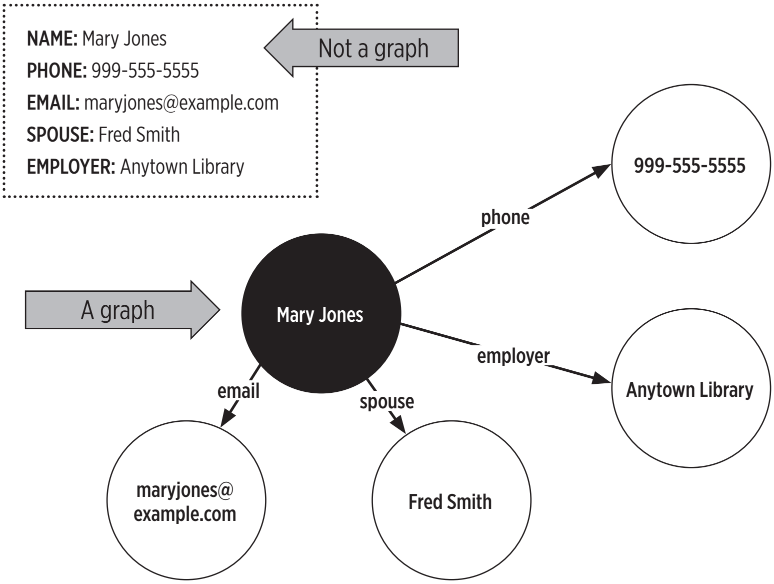
Everything being described in RDF must have a standard identifier that begins with "http://" followed by a domain name (e.g., "ala.org/") and a precise path (conference2015). That might seem confusing, because that is the same prefix that is used for a uniform resource locator (URL), which is the address of something on the web. RDF is using the same standard for its identifiers for a couple of reasons: first, the mechanism to create and manage domain names on the web already exists, which means that it will be easy to create these identifiers; second, the combination of identification with location means that information about the thing identified can be stored on the web at that location without any change in technology.

RDF identifiers are intended for machines, not humans. No one wants to read, much less type, "http://id.loc.gov/authorities/subjects/sh85038796" for the Library of Congress subject heading "Dogs." All identifiers can have human-readable labels, and the assumption is that in every situation where a human is interacting with the data, the human-readable label will be the one displayed. This includes input, which in many data creation scenarios in business applications already makes use of textual pull-down lists for easy and accurate input. Thus a cataloger will choose a subject heading, such as "Dogs in literature," from a list and the data stored will be "http://id.loc.gov/authorities/subjects/sh85038823."

Identifiers are in a sense merely a substitution of the normalized text we use today, often in the form of a formatted heading, with a particular string in the URL format. Other changes required in the shift to RDF are more radical. One of the ones that is most difficult to understand is that RDF data about resources is not stored as separate records; instead, information about a thing is in the form of a graph of statements. Graphs have no boundaries; they can grow and they can interconnect with other graphs where their data intersects (figure 3.1). To give a simple example, the identified author in a library catalog description can interlink with the author information page on Amazon or with the encyclopedic entry about the author on Wikipedia. This assumes that these systems have knowledge of each other's identifiers, but that is increasingly the case: library authority identifiers are already found in Wikipedia entries, so this connection can be made. Data in RDF resembles synapses, with multiple connections that allow new information paths to be created as more information is added (figure 3.1).

FIGURE 3.1

A graph



The next key piece of information about RDF is actually about the nature of the World Wide Web itself. The web is an open space where millions of people and corporations and governments can put information that they wish to make public. Most contributors to the web also have other information stored in private data repositories. Although these private repositories may be in some way connected to the Internet, they are protected by user accounts and passwords, and some are protected through layers of digitally locked doors. The Semantic Web has an emphasis on the public information space, although its technology can also be used for privately held data.

There are three main principles that govern the Semantic Web that are important for understanding the rules that are applied to Semantic Web data:

- the Open World Assumption
- the Non-Unique Name Assumption
- "anyone can say anything about anything"

The Open World Assumption describes the nature of the web, which is that the web is never complete, never done, and it may not be possible to have access to all of it at any one given time. What this means is that web applications must not rely on completeness. If your bibliographic description on the open web has no title, it doesn't mean that there will never be a title, or that there hasn't ever been one. You can assume that a title exists, just not in your current view. Contrast this to a database application that has strict control over input and output, and where rules governing the data are enforced: that title must be there. In a database, a bibliographic description with no author means that the resource has no author attribution. In the web environment, that negative cannot be assumed from the absence of the element.

The Non-Unique Name Assumption (NUNA) states that any identified thing can be identified with more than one identifier. This is like real web life, where I am identified by more than one e-mail address (one at kcoyle.net and another at gmail.com), an IRC handle, and a Twitter name, in addition to my social security number, passport number, driver's license number, and so on, in "real life." On the web you cannot assume that each identifier represents a unique entity. To avoid chaos, there are ways to code identifiers as identifying to the "same" or "different" resources, but the Non-Unique Name Assumption rules any identifier pairs without explicit relationships, such that you cannot draw conclusions from identifiers alone.

The statement that "anyone can say anything about anything" is as true for today's World Wide Web as it is for the Semantic Web: there is no technical restriction on who can put information on the web. There is also no restriction on who can link to resources on the web. You may exercise content control over a web site that you create, but you cannot stop anyone else from linking to it. The same is true on the Semantic Web, where anyone can create links to your data. There is, however, a difference in the effect of linking on the Semantic Web as compared to the web of web pages, because RDF links are more meaningful than links between web pages. Links between web pages have a single meaning, which is simply "this links to that." Semantic Web links carry a meaning to the link, such as “this is a sub-class of that,” or “this is the same as/different from that.” These are conditions that you should keep in mind when designing your data. To the extent that you can predict how your data might interact with other data in that vast data space, you need to design your data to “play well with others.”

The basic technology of the Semantic Web is RDF. Other technologies build on that. One of these is the Web Ontology Language, OWL, which is the language developed for the creation of Semantic Web vocabularies. First, yes, OWL should be WOL, but it is OWL. Second, the RDF documentation uses the terms *vocabulary* and *ontology* interchangeably. The term *ontology* comes out of the artificial intelligence community and it implies a level of rigor in the definition of terms and their relationships. OWL is to the Semantic Web what a metadata schema has been for us in the past: OWL is how you define the terms of your domain and how you will use those terms to create your data.

OWL is a difficult standard to understand if you are not familiar with certain aspects of artificial intelligence decision-making. Many of the features that are defined in OWL sound familiar but in fact mean something different from what most of us are accustomed to. OWL is designed for a particular Semantic Web function called “inferencing.” Inferencing allows you to draw conclusions from data that is present. Thus if:

Every man is a mammal
Fred is a man
Therefore, Fred is a mammal

OWL is quite a bit more sophisticated than this example implies, and includes concepts such as “inverse functional object property” and “negative data property assertion,” among many others. The purpose of OWL is to define a vocabulary that can be used in complex artificial intelligence work. It also includes the ability to define some common features of metadata languages, such as cardinality (mandatory, repeatable) and equivalence (same as or different from). Unfortunately, what these features mean in OWL can be quite different from what they mean in metadata standards with which we are familiar.

The meaning of the OWL terms is governed by the RDF concept of classes, in which things being described acquire their membership in a class from the terms that define them. In our simple example above, Fred acquires “mammal-ness” because he is described by the term “man,” which itself has been defined as being of class “mammal.” In artificial intelligence this mimics the human brain’s ability to draw conclusions from information in the environment, generalizing from knowledge gained in one experience to apply in other situations. The Semantic Web builds up knowledge from atoms of learning, which is the opposite of the top-down approach that is common in classifications of knowledge.

There have been controversies about OWL since its inception, because it is so very complex and also so easily misunderstood. Depending on your application, you can ignore much of that complexity, but for any OWL assertion that you do use you must make sure that you understand the consequences of its use. In particular, many of the OWL declarations about terms and classes seem identical to functions in familiar programming languages. A simple example mentioned above is cardinality. Cardinality in programming languages declares the minimum and maximum allowed occurrences of a data element. If the minimum cardinality of the element is “1,” that element is required—it must occur at least one time. If it is “0,” then the element is optional. If the maximum cardinality is “1,” the element is not repeatable, but any other number defines the number of times it can repeat in your data. In most programming situations, data that violates these rules is considered to be in error.

OWL has minimum and maximum cardinality, but their meaning has a different interpretation due to the application of the Open World Assumption and the Non-Unique Name Assumption. You can define your data as having, for example, a single creator for each given resource; the maximum cardinality of your creator element is therefore “1.” If you create or encounter data that has more than one creator for a single resource, this is neither an error nor even an inconsistency in the data. Instead, applying the rules of the Semantic Web, applications that interpret OWL data will conclude that all of the creator identifiers identify a single entity because your rule says that there is only one such entity, and that entity can have any number of identifiers. At times this OWL rule may come in handy because you want to find equivalent identities, but that presumes that the data has all been coded correctly, something that most of us have learned is rarely the case. This is the big “gotcha” of OWL. OWL-based software can examine data that exists and can return a response that the data either does or does not conform to the OWL rules that have been defined for those data elements. But OWL cannot control the creation of data that meets its rules; it examines but that it does not enforce, in large part because ”anyone can say anything about anything” and because OWL is intended to function in an open world that is always in flux.

This aspect of OWL generally confuses people because the OWL rules so closely resemble the rules that other programming languages use for a very different purpose: data quality control. In fact, because people often want to use OWL rules in the same way that they use programming rules in closed and controlled environments, there is now software that applies the OWL rules in closed environments, treating identifiers as uniquely identifying a single entity. This reverses two of the main truths of the Semantic Web, which are the Open World Assumption and the Non-Unique Name Assumption. It also operates on data stores where "anyone can say anything about anything" is definitely not allowed. In other words, a mirror copy of the OWL language is being used in the same way that we have always used programming languages, but not in the way intended for the Semantic Web.

Within your own closed environment, such as a local database, you clearly can do whatever you want with your data and you can impose any kinds of rules and controls that serve you and your organization. But if you open that same data to the web, the meaning of those rules will be interpreted using the Semantic Web standard meanings, which means that the Open World Assumption and the Non-Unique Name Assumption will be applied. The actual meaning of your data will be radically different in those two different environments, and operations like searches could yield very different results. The upshot of this is that the same OWL-defined vocabulary should not be used in both the closed and the open worlds.

This conflict between the controlled data stored in one's personal or corporate database and the open environment of the web is one of the hardest for data designers coming from other technology environments to overcome. There obviously is a real need to perform quality control on data, but the basis of the Semantic Web is one of discovery, not control. This is a conflict that, as of this writing, is unresolved, both in code and in terms of best practices. One possible solution, proposed by the Dublin Core Metadata Initiative (DCMI), the same people who develop the Dublin Core metadata terms, is to separate the controlling aspect of the vocabulary from its basic semantics. This isn't different from many existing metadata implementations: terms to be used are defined for their meaning, and a separate structure and rules are developed that turn those terms into a metadata record.

Dublin Core (DC) is a good example of this. Dublin Core terms are defined apart from their use in metadata. Dublin Core's element "title" is defined simply as "the name of the resource." Whether it is mandatory or optional, and whether or not it is repeatable, is not part of the definition of the term itself. Those rules would be defined in a metadata schema or in what the DCMI calls an "application profile," which is a definition of the metadata structure and rules for a particular application. The term can be used in different ways in different metadata implementations, and the DC terms are indeed used in a wide variety of situations. However, in all uses the term retains the same meaning. This separation of meaning from rules results in maximum flexibility that allows the same terms to be used in many different applications, as Dublin Core terms are today. That flexibility is the positive outcome of this method. The negative outcome is that the separation of meaning from rules results in maximum flexibility, so that data sharing requires some adjustment between communities. The application profile, if provided in a machine-readable form, can be the basis for data sharing because communities can easily understand the structure of data created by others. Through all of that, however, a Dublin Core title remains "the name of the resource" even if some communities allow only one, some more than one, and for some the element may be optional.

We can contrast this to the primary metadata standard used in libraries today, MARC 21. This standard defines the meaning of terms and also the rules for data quality in a single standard. This is not uncommon as a data creation and management approach, however, it is undeniably a definition of a closed data world. Anyone who would use the base MARC record structure and data elements with a different set of rules governing term meanings and cardinality would simply not be creating MARC 21 data, and there would be no expectation that one could successfully combine data created under such different sets of rules.

The final aspect of the Semantic Web that I'll cover here is classes. We're all familiar with the concept of a class from scientific taxonomy and classification systems. In those systems we assign things to classes to give them the meaning of the class, putting ourselves and cats in the class "mammals," and books on mammals in one of the sub-classes of biology. Classes have a different meaning and work differently in the Semantic Web; they are not categories or boxes to put things into, but are meaningful information about things that can be used in various contexts. Classes are not exclusive in their nature, and anyone or anything can have the qualities of more than one class. This is much like the real world, where a person can be an employee in one context, a parent in another, and a volunteer firefighter in yet another. Rather than assigning a thing to a class, the class is deduced based on how something is described. Our rules may say that persons with paychecks are employees, those with children are parents, and those who are members of Volunteer Brigade 7 are volunteer firefighters, and anyone can be all three. By attributing characteristics to the thing we are describing, we build up our world by describing it. This, too, fits into the methods of artificial intelligence where their creations must be able to make deductions about newly encountered things in the world based on information, as we do in real life. We recognize chairs as chairs even if we haven't seen a particular chair before. We understand that a person is a police officer because anyone wearing that uniform is a police officer, even if we haven't seen that person before. We are moved to open the door for a person carrying packages because we know that it's hard to open a door when your hands are full, in spite of not having been in this exact situation (same person, same door, same packages) before. All of this computation happens quickly and naturally in the human brain, and some of it can be imitated through code if the right information is given about things we describe on the web.

The preceding describes some of the fundamentals of the Semantic Web. The Semantic Web is implemented as *linked data,* a set of common practices for data on the web. One of these practices, the use of http-based identifiers, has been discussed above. Other practices have to do with making sure that your data can be used in the open environment of the World Wide Web. There are standard ways to define your metadata so that others can understand it and potentially use it. Linked data is a mix-and-match technology, and people are encouraged to make use of metadata definitions that exist rather than inventing their own. Any description can be made up of metadata from a number of different sources, and can use descriptive elements found anywhere on the open web.

From this description you can undoubtedly conclude that a future library data standard using linked data would look considerably different from the data we have today. The purpose of linked data is both discovery, through hyperlinks, and new knowledge creation, by linking between previously separate communities and their data stores. Those looking at linked data for libraries are focused on the library catalog and its discovery function. Our current catalog data is very different in its goals and content from data that would play well in a linked data environment. The challenge for us is to make this transition intelligently, and in a way that serves library users. The remainder of this book looks at current efforts with that challenge in mind.