METDATA FOR DIGITAL LIBRARIES

Sonoe Nakasone

Metadata for Digital Libraries

Sonoe Nakasone

Cover: “Elegant patterned notebook book cover,” courtesy of Canva.com

Design and composition by Sonoe Nakasone in Arial typefaces using Google Docs.

Cataloging data:
Metadata for digital libraries / Sonoe Nakasone. Includes bibliographical references. 1. Metadata. 2. Information organization.

Note on this edition and future editions

This edition is missing four forthcoming chapters and was “published” as an incomplete work so that completed chapters were made available to my students.

The next edition submitted will include the missing chapters. Subsequent editions will be posted as text files or in mark down to allow contributions. Contributors of future editions will be credited as contributors.

Images from pixabay.com are those not requiring attribution and are identified in caption or in the main text as being from pixabay.com. Other unidentified images are of my creation.

Contents

Introduction:

With all the books on metadata out there, why create another? This book was created to help me teach the course Metadata Applications for Digital Libraries at North Carolina Central University, which I've taught since Spring 2018. Although there are alternatives that are better in quality, none that I have seen adequately address an idea that is central to this course: that metadata is not neutral.

In *How to be An Anti-Racist* (2019), Ibram X. Kendi defines racism as "the marriage of racist policies and racist ideas that produces and normalizes racial inequities." Kendi further defines a racist policy as "any measure that produces or sustains racial inequity between racial groups," and a racist idea as "any idea that suggests one racial group is inferior or superior to another racial group in any way." This concept of policies and ideas that produce and sustain inequities or the notion of inherent inferiority or superiority based on group identities such as race, sex, gender, class, religion, ability, and other categories defines the landscape of power in which we currently live.

So what does that have to do with metadata? Metadata relies on several tools such as language and categorization to exercise control over information. These tools often perpetuate ideas that suggest inferiority or superiority of groups of people. Additionally, these tools become policies when they are standardized into metadata "rules" or systems that become widely used or required. Yet metadata is often taught and presented as the reporting of objective facts, as if the goal is neutrality, not acknowledging thenot neutral world in which we live.

In this textbook, I've struggled to integrate technical explanations and examples of metadata and how metadata used in digital libraries with explanations and examples of how metadata reflects historic and current inequities and power structures such as racism, sexism, classism, and additional isms. A lack of practice and my privellege as an English speaking, "able"-bodied, straight identifying, cis female, United States citizen with a

graduate degree no doubt limits my ability to accomplish this goal. My belief in social justice and my own experienced marginalization as a feminine presenting, Black and Asian person motivates me to continue to speak about the powers metadata exert.

My other reason for writing this book was to create a textbook that used minimal jargon, simple language, and relatable examples because learning about a new and complicated topic is hard enough. I hope I succeeded.

Each chapter focuses on one broad topic that is broken up into three to five sections. Each section starts with a list of important vocabulary and ends with quiz questions and a discussion question. My goal was to make each section no longer than seven pages, quiz questions included. In reading this textbook, it may help to read no more than one section per day. Read aloud, each section takes approximately 15 minutes to complete. Spending a half hour on each section by reading slowly to ensure understanding and information retention recommended. Ideally, each chapter represents no more than 2.5 hours of reading per week.

Chapter 1:

Metadata is a broad concept both inside and outside of libraries, archives, and museums. Originally a word used by computer scientists to describe structures within databases, “metadata” has become a sub-discipline in the study of libraries and other information centers.

This first chapter will define and explain what metadata is within the library context, discuss what it is used for, suggest reasons metadata is important, and finally, present ideas of what makes “good” metadata.

Section 1: What is Metadata?

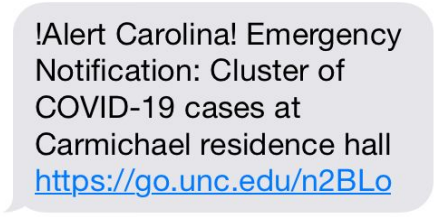
In this section, you will learn what metadata is and begin to recognize examples of metadata.

Important vocabulary in this section
Metadata: _____
Resource: _____

NISO (the National Institute for Standards Organization) author Gail Hodges defines *metadata* as:

> "Structured information that **describes, explains, locates**, or otherwise makes it easier to **retrieve, use, or manage** an information resource. Metadata are often called **data about data** or information about information" (2001).

I want to focus on the "data about data" part of this definition for now. What does it mean to have data about data? Consider your cell phone for a moment. Your cell phone creates and stores photographs, sound recordings, text messages, and more. Those are all data because they contain visual, sound, or textual information.



Three images left to right: 1) sound waves (pixabay.com), 2) a photograph of a tropical island (pixabay.com), 3) a text message from my phone. These images illustrate that information can be audio, visual, and textual, among other varieties.

View the photos on your phone, however, and you may see something other than just pictures. You might see metadata. Below is an example of photos on my phone. Two photographs are displayed under “Nelson Hall, Raleigh, NC, Feb. 3”; another eight are displayed under “Raleigh, NC, W Lane St., Feb. 4.”

If the photos on your phone are the data, and metadata is information about data, which information below is the metadata for the photos?

- ❏ Nelson Hall (location picture was taken)
- ❏ 81% (battery life of the phone)
- ❏ Feb 3 (date picture was taken)
- ❏ 11:31AM (current time)
- ❏ W Lane St (location picture was taken)
- ❏ AT&T (phone service provider)
- ❏ Raleigh, NC (location picture was taken)
- ❏ Feb 4 (date picture was taken

If you guessed that all the dates when the photos were taken and locations where the photos were taken are the metadata, then you are correct. In this case, the dates (February 3 and 4) and locations (Nelson Hall and W Lane St) are data about data (the photos). Can you think of other metadata for these photos that is not displayed in the example?

Now look let’s look at another part of the NISO metadata definition: metadata “describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage [a...] resource.” Hold up! What is a *resource*?

Before we go further, it will be helpful to define the word “resource” because I’ll use that word throughout this book. A resource is anything that contains or is the source of information or data. A book is a resource: books contain information. A person can be a resource: a person can provide information. A painting is a resource: paintings contain visual information. A mp3 file is a resource because it contains a song, which is also resources. Songs are resources because they consist of information in the form of musical notes, sound waves, and sometimes words. A resource can be a physical thing like a CD, a person, a painting; it can be digital, like a Microsoft Word document or Spreadsheet; it can be something intangible, something that you can’t put your hands on unless you put it into a container, like a song or a story.

In libraries, archives, and museums, metadata doesn’t only refer to data about data, but also data about resources. It’s possible you’re wondering, if resources contain data aren’t resources and data the same thing? My answer is “sort of.” For example, look at this photo of a mountain that I got from pixabay.com; pretend that I took it on my phone. It is digital (a computer file). As I noted earlier, the photos on my phone are data—they are visual information. Now pretend I printed the photo onto paper. Now we have a digital photo and a photo on paper. It is the same data—the same image of sky, mountain, and ocean—but they are different resources. The metadata for the data is the same whether it is digital or paper: the location of the photo is Honolulu; the date the photo was taken is October 5, 2018; depicted in the photo is Diamond Head (inactive volcano). The metadata for the resources, however, are different. Metadata for the printed photo resource might include the date I printed it, what type of paper I used, or how big I printed it. Metadata for the digital file might include that it is a jpg file and that it is 80MB large.

What is the definition of a resource that I shared?

a) Something you buy

b) A thing that contains or is the source of information or data
c) Data

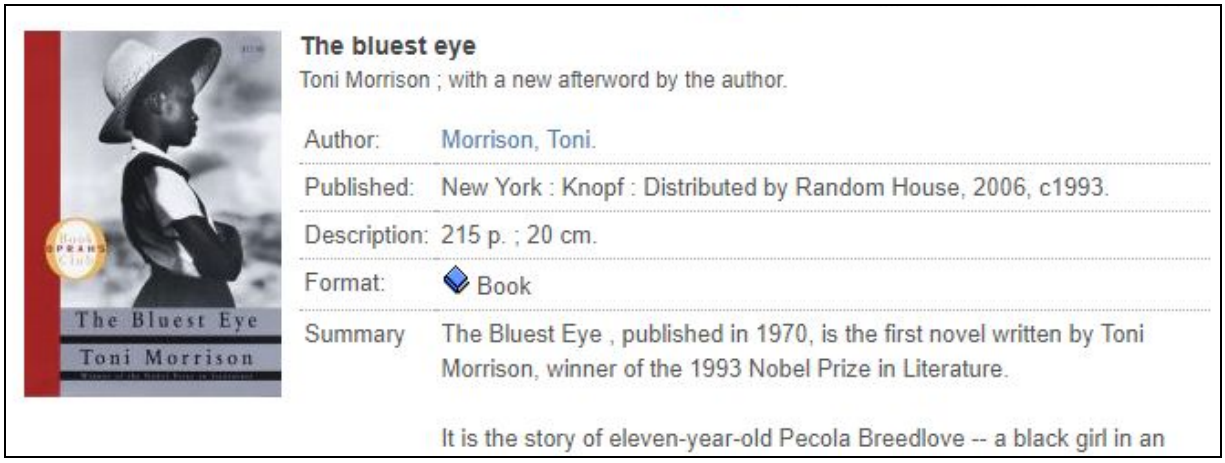
According to my example, can two resources contain the same data?
a) Yes
b) No

Let's return to the NISO definition of metadata now that we know that metadata is data about data and data about resources. NISO also says, metadata **"describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage"** a resource. This not only helps to define metadata, it tells us some of the things metadata is supposed to do.

When I look at the photos on my phone and see the dates and locations, do these pieces of information describe, explain, or locate my photos? Does the metadata for my phone's photos make it easier to retrieve, use, or manage my photos? I'd argue they do. The dates and locations provide context (an **explanation**) for the photos. Knowing when and where the photos were taken could potentially connect the photos to events happening at the same place and day or explain why people look or are dressed a certain way. It also explains the photo enough to help jog my memory of what was happening in the photo, why I took it, and why I decided to keep it (or why I might want to delete it). The locations make it easier for me to **retrieve** the photos by organizing them in chronological order, so I know how far to scroll to view and **use** them.

* * *

Libraries, archives, museums, and other cultural heritage institutions use metadata to describe, explain, locate resources or data and make it easier to retrieve, use, or manage them. Before we move on to the next section to talk more about what metadata does and why it is important, let's look at a typical example of library metadata. Below is a screenshot of metadata for a library resource. The resource is a physical book called *The Bluest Eye*.



Screenshot taken of the metadata for the novel The Bluest Eye from North Carolina State University Libraries catalog, accessed March 2018. https://catalog.lib.ncsu.edu/catalog/NCSU983879

Just as we examined the metadata for the photos on my phone, let's examine the metadata for the book called *The Bluest Eye*. If the book—the thing made up of bound together pieces of paper—is the resource, then which information is the metadata? As you think about this, ask yourself, do any of the following information listed describe, explain, or locate the resource? Do any of them help to retrieve, use or manage the resource?

- ❏ Title (The Bluest Eye)
- ❏ The length of the book (20cm)
- ❏ The author of the book (Toni Morrison)
- ❏ Format (Book)
- ❏ A summary of the story ("It is the story of eleven-year-old Pecola…")
- ❏ Number of pages in the book (215 p.)

If you said that all of these pieces of information are metadata, then you are correct. The length of the book and number of pages **describe** what the book looks like. The format **explains** what type of resource it is (a book as opposed to a CD or a magazine, for example), and the summary **explains** what the book is about. The author of the book **describes** and **explains** who created the story. The title *The Bluest Eye* makes it easier to **retrieve**, **use**,

and **manage** the resource, since calling it by a title is less confusing than calling it “the book.”

Take a moment to think of other metadata not shown or discussed that could also be used to describe, explain, or locate or make it easier to retrieve, use, or manage this book.

There are many more pieces of metadata for *The Bluest Eye*, but to include them all could be overwhelming to look at. Librarians have to make choices and judgements not only on which metadata to include to make the metadata easier to use but also how the metadata is represented. Metadata, then, is but a representation of a resource that will never be completely accurate because it will never exactly reflect the resource. If metadata will never be completely accurate, what is metadata good for? We’ll discuss that question further in the next section.

To summarize, one definition of metadata is data about data. In libraries and other cultural heritage institutions, metadata is not only data about data, but also data about resources. Resources are things that contain or are the source of data. Although it may sometimes seem that a resource and data are the same thing, it is important to note that two different resources can contain the same data. Metadata is important because it is not merely information about a resource, it is information that describes, explains, or locates a resource. Metadata also makes it easier to retrieve, use, or manage a resource.

Section 1 Quiz questions

Review the NISO definition of definition. According to this definition, metadata is...

a) Data about data
b) Information that describes, explains or locates a resource
c) Structured information
d) All of the above

Which is an example of metadata for a library resource discussed in the reading:

a) A friend's phone number
b) The title of a book
c) A library's hours of operation

Based on what you learned in this section, what might be an example of metadata for a CD?

a) The title of the CD
b) The songs on the CD
c) The disc itself

Section 1 Discussion Question

If two resources contain the same data, what are some reasons that / how would their metadata differ?

Section 2: What is the purpose of metadata?

In the last section, I wrote that metadata can never be entirely accurate because it is only a representation of a resource. If that's true, why have metadata at all? What's the point? By now, you may already have your own answers to these questions. In this section, however, I'll continue to discuss what metadata does and the role it plays for data and resources.

Important vocabulary in this section
User _____
User tasks (FRBR) _____
Known item search _____
Identifier _____
Browsing _____
Interoperability _____

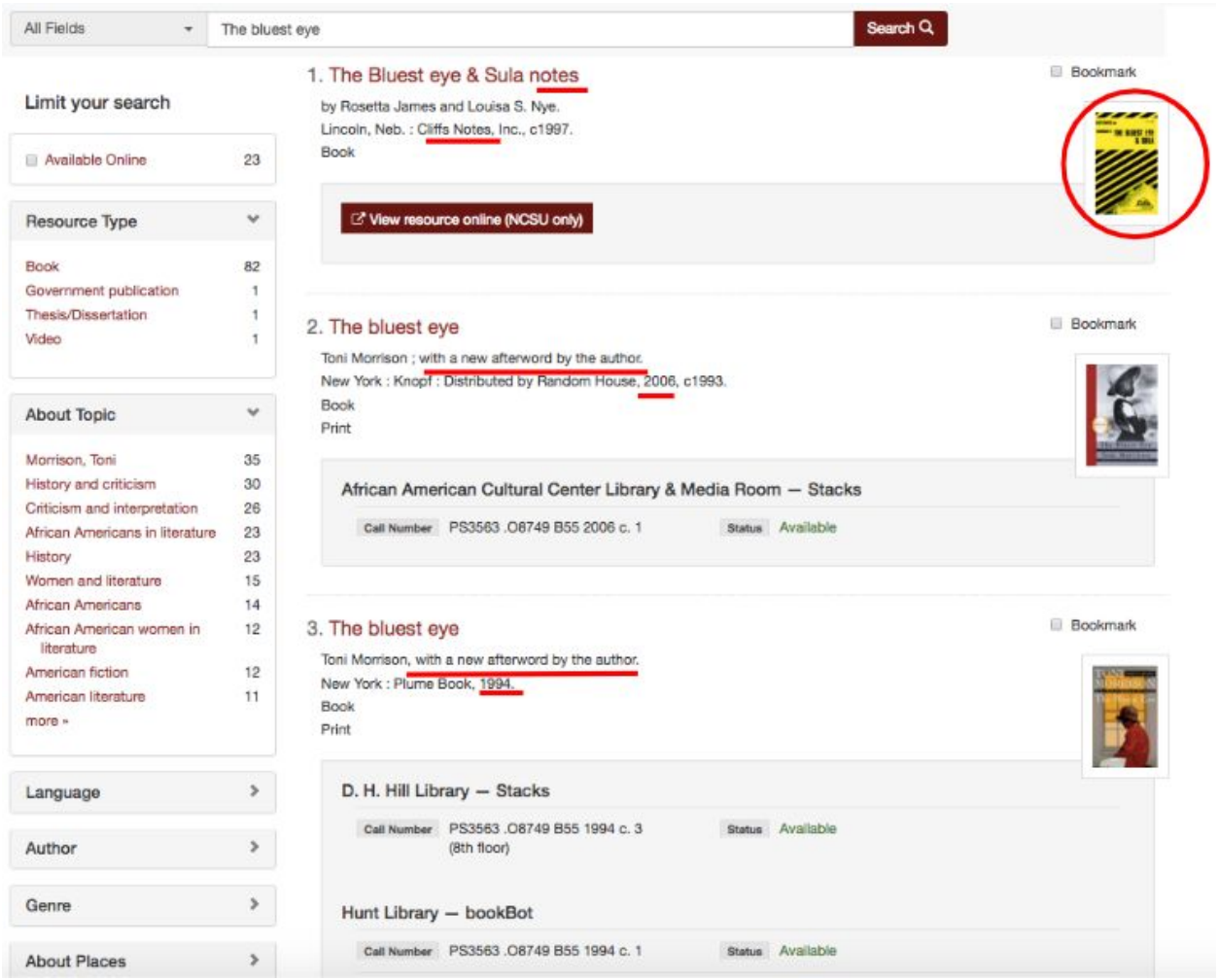
To talk about why metadata is useful, I'll use the "user tasks" outlined by the Functional Requirements of Bibliographic Records (FRBR)[1]. In the FRBR, a *user* is a person or group of people who seek information or resources. A *user task* is an action taken in order for a user to satisfy an information need. FRBR is more elaborate than user tasks, but I focus on FRBR user tasks here because they are one way of understanding how metadata helps people achieve information goals.

The FRBR user tasks are *find*, *identify*, *select*, and *obtain*. In the following examples, let's pretend I am a user and my information goal is to read *The Bluest Eye*. The metadata will help me perform the four user tasks to reach my goal.

Find

[1] FRBR is a conceptual model for understanding library resources and their metadata. Conceptual models allow people to define concepts and relationships between concepts for systems or other complex subjects.

Metadata helps people **find** resources. Search boxes are a popular way for people to find information or resources online. In this image of a library catalog, I entered “the bluest eye” in a search box. There are many ways to use a search box: you can search for resources related to keywords, subjects, genres, authors, and more. This example features a *known item search*, in which a person is searching for a specific, known resource.



Screenshot taken of the search results for the novel The Bluest Eye from North Carolina State University Libraries catalog, retrieved March 2018.

The search results are listed on the right; each result represents a different resource. From the list of resources, it looks like I’ve found several resources that match my search. But how do I know that? I know that because there is key descriptive information (metadata) listed: the title of the book.

Identify

The library catalog search results provide multiple options for *The Bluest Eye* in the image above. Sometimes metadata will include a unique *identifier*, a number or number and letter combination that is unique to a resource, like the ISBN on a book. An identifier could help if I knew the resource's identifier and could match it to one listed in the metadata. Without such information, I will rely on the other metadata from each search result to **identify** the resource needed.

In addition to helping me find the book in the catalog, the book's title also helps identify appropriate resources. For example, the title of the first result "The Bluest eye & Sula notes" isn't quite right. Additional metadata such as the thumbnail image of the book cover and the publisher illustrate that these are Cliff Notes, a popular series of summaries and study guides, for *The Bluest Eye* rather than the novel itself.

Select

The titles of the second and third options listed titles match completely. Found it! Now, which do I select? Some of the metadata listed can help explain if or how the two options are different enough to affect my choice. For example, although both include an afterword by the author, they were published in different years by different companies. These metadata indicate that although the main content of the novel may not have changed, there may be small changes in pagination (page numbering) or accompanying text like prefaces, introductions, forwards, annotations, afterwards, etc. Although some people may not care about these differences and would be happy with either resource, metadata can help users who do care about such differences to select the version of the book they want or need.

Obtain

I decide I want to obtain a copy of the third option for *The Bluest Eye* listed in the search result: what metadata would help me?

a) Place the book was published or printed (New York)
b) Current location on the library shelf (Call number: PS3563…)
c) Type of resource (Book)

If you said the current location of the book on the library shelf (the call number), you are correct. The call number will tell me where on the shelf I can locate the book so that I can grab it and borrow it.

* * *

The FRBR user tasks are not the only way to think about why metadata is useful. Jennifer Riley mentions several reasons metadata is used in her NISO article Understanding Metadata (2017).

Riley notes that metadata allows for *discoverability*, that is the ability to look for and find resources. Discoverability relates to the examples of the FRBR user tasks finding, identifying, and selecting above. It's also possible to "discover" resources, through *browsing*. Browsing relies on metadata that organizes resources according to categories or characteristics. In the earlier FRBR user task examples, using a search box was efficient because I already knew the title of the book I wanted. If I wanted to find a new, unknown novel to read, however, I might have used the facets to narrow results based on categories or characteristics like topic, author, genre, and more.

Riley mentions a few other uses of metadata including interoperability, preservation, digital asset management and navigation. Broadly speaking, these other uses of metadata relate to the following part of the NISO definition for metadata related to use and management of a resource. You'll learn a little more about these additional functions of metadata in Chapter 2 when you learn about different types of metadata. For now, I will note that *interoperability* is the ability to go from one system to another without losing information. For example, two libraries might want to borrow each others' library resources from time to time. If each has interoperable metadata, the metadata can easily go from one library's catalog to another library's catalog.

* * *

I'll conclude this section with two additional reasons for metadata: 1) context for resources and 2) saving time and energy. Metadata provides context to help you understand if you've found the right thing and how to use it to achieve your goals, for example, with a list of relevant subjects covered by the resource or a summary of the contents. Metadata also saves you time and energy like a map to a city. Instead of traversing an entire city to understand it, looking at a map summarizes the landscape of a city at a glance. Similarly, rather than spend time and energy reading through an entire book or stack of books to find something you want, metadata provides a window into a resource, giving you key pieces of information to help you get to the resources you need.

To summarize, there are many reasons we have metadata. Metadata provides description, context, organization, and saves time when we want to find, identify, select, and obtain resources. Although metadata helps library users with "discoverability," it can also help people to use and manage resources, for example, by helping with interoperability (the ability to move between systems without information loss).

Section 2 Quiz questions

Which of the following are reasons discussed in the reading to explain why metadata is useful?

a) Metadata assists users find, identify, select, and obtain resources
b) Metadata helps with context for and discoverability, use and management of resources.
c) A & B are true

The reading defines interoperability as

a) The ability to go from one system to another without losing information.
b) The ability for libraries to borrow each others resources
c) The ability to display metadata in different languages

Reflect on what you've read in this section and apply it to this new scenario. If I want "discover" which books by Toni Morrison a library has, which piece of metadata would be most helpful for me?

a) Publication year
b) Author name
c) Location of a specific book

Section 2 Discussion Question

Think of a resource you own and some metadata about the resource. Explain the ways that metadata is useful and why someone would want to know about that metadata.

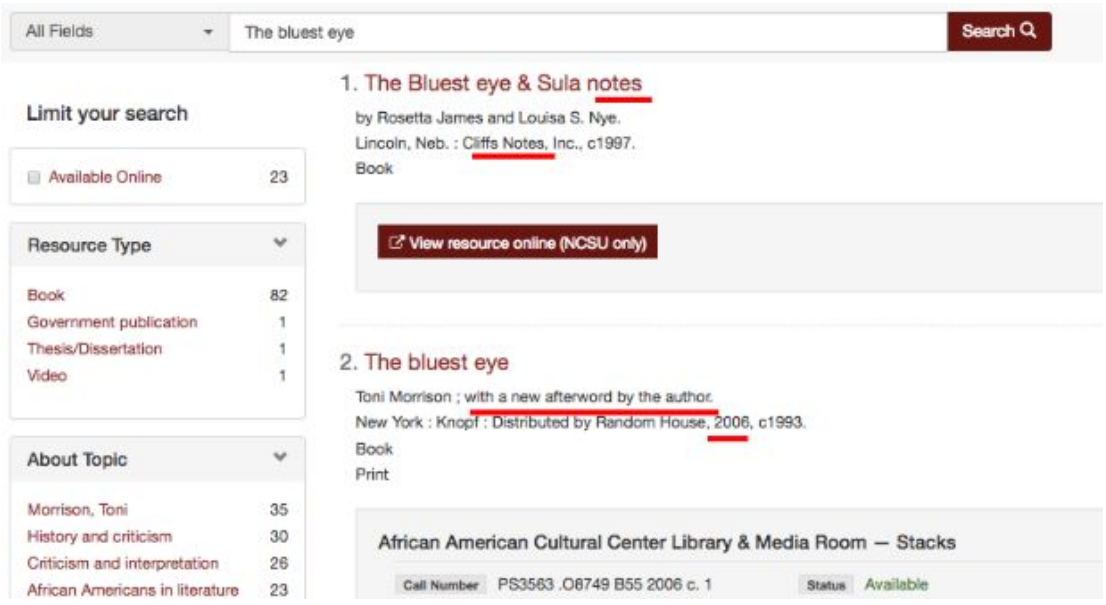
Section 3: What is the purpose of metadata (continued)?

In this section, I will revisit some of the ideas discussed in the previous section and further explore their consequences.

You may have noticed metadata helps people to control information and resources. In some cases, that control is exercised in order to help us to find a good read, a pair of shoes, the latest cancer research, and much more. In other cases, the control and power metadata exerts over information has numerous implications on what information is available and how it is made available.

To find / to hide

In this image of library catalog search results, the facets on the left side allow you to narrow your search. These facets help people to find resources through browsing, discussed in the last section. Perhaps you've used similar facets to limit your search by color, size, or some other category when shopping online.

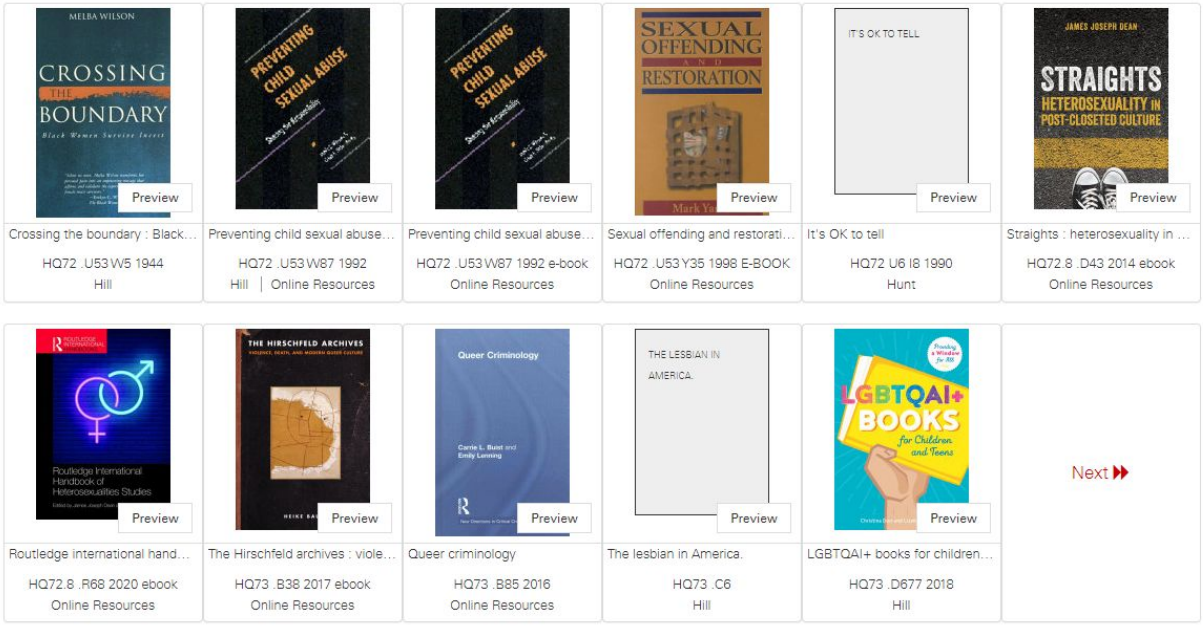


Screenshot taken of the search result for the novel The Bluest Eye from North Carolina State University Libraries catalog in March 2018.

Just as easily as metadata can populate such facets to help people find resources, it can also allow people to hide resources and information, for example, Facebook's use of metadata to enable discrimination. Facebook metadata enabled advertisers to post illegal discriminatory housing advertisements that excluded people who were thought to have disabilities or children or who were thought to be Black, Latinx or have other "cultural affinities." The housing ads discriminated based not on the specific content on a person's Facebook page (the data), which would also be illegal, but on metadata Facebook created based on how it described the content on a person's profile page.

Find / narrate

The ability to organize resources or information into categories not only allows people to find or hide information based on those categories, but also tells a story. In libraries, the way we order or organize resources also can affect the story told about them. We'll talk more about classification systems in a future chapter, but below is an example of books organized according to the Library of Congress Classification System or LCCS. This example is based on one that is discussed more fully in Melissa Adler's *Cruising the Catalog* (2017).

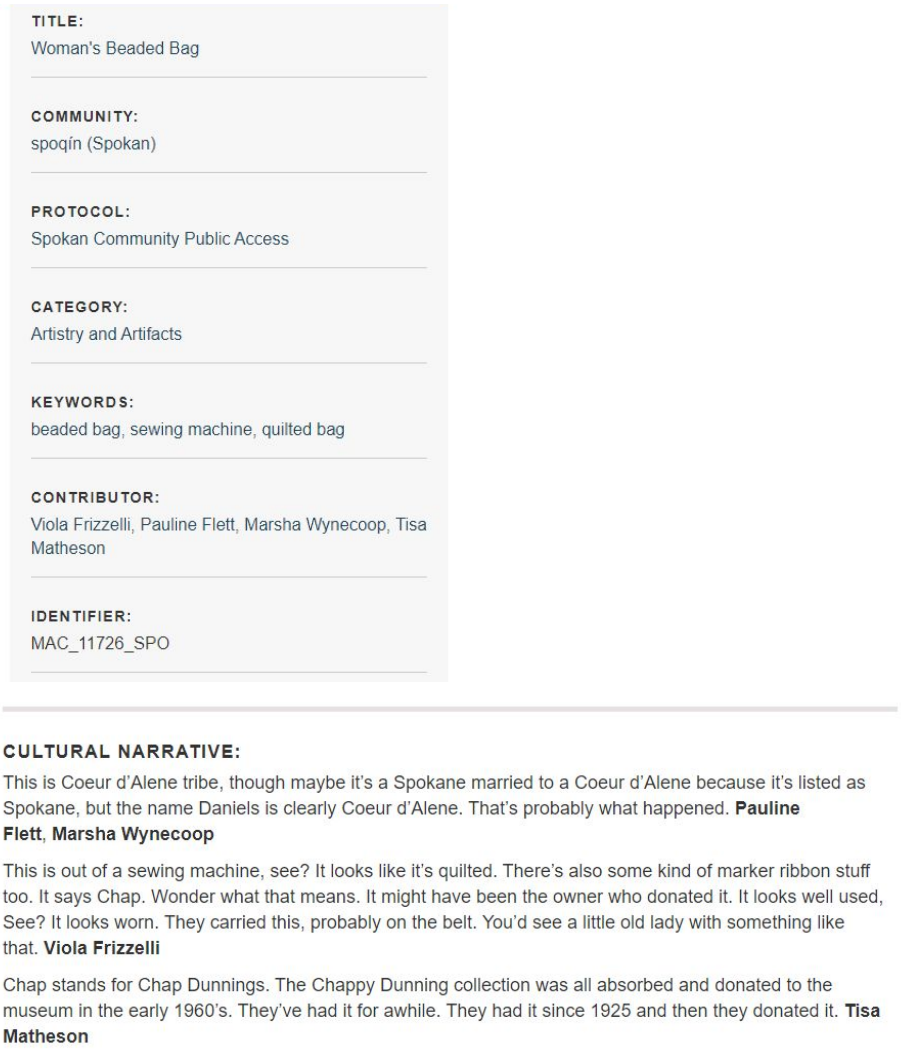


Screenshot of the virtual browse function from North Carolina State University Libraries catalog in March 2018. This function emulates the way books would appear on the physical shelf due to their Library of Congress Classification numbers.

In this image, books on LGBTQIA+ topics are placed near books on child sexual abuse because of their LCCS (Library of Congress Classification System) numbers. LCCS is a system of letters and numbers used to organize books into related categories. In the library, books located closely together are assumed to be related. By placing these books side by side, what is the library saying? At best, this is confusing. At worst, this suggests a false narrative about LGBTQIA+ topics and is hurtful to patrons who use the library to find books on LBGTQIA+ topics.

Interoperability / assimilation

Section two briefly discussed metadata's role in interoperability, especially if we want to share a resource from one system with another system. If two systems are not completely the same, we might lose information when trying to manage a resource taken from one system in another system. Interoperability would tell us that metadata can help to manage the transition from systems by providing information about a resource that is useful in both systems. In this way, interoperability assumes that data and metadata must assimilate into the system it enters just as a person might assimilate to the culture of a new country or that the system must change to accommodate the data and metadata.



TITLE:
Woman's Beaded Bag

COMMUNITY:
spoqín (Spokan)

PROTOCOL:
Spokan Community Public Access

CATEGORY:
Artistry and Artifacts

KEYWORDS:
beaded bag, sewing machine, quilted bag

CONTRIBUTOR:
Viola Frizzelli, Pauline Flett, Marsha Wynecoop, Tisa Matheson

IDENTIFIER:
MAC_11726_SPO

CULTURAL NARRATIVE:

This is Coeur d'Alene tribe, though maybe it's a Spokane married to a Coeur d'Alene because it's listed as Spokane, but the name Daniels is clearly Coeur d'Alene. That's probably what happened. **Pauline Flett**, **Marsha Wynecoop**

This is out of a sewing machine, see? It looks like it's quilted. There's also some kind of marker ribbon stuff too. It says Chap. Wonder what that means. It might have been the owner who donated it. It looks well used, See? It looks worn. They carried this, probably on the belt. You'd see a little old lady with something like that. **Viola Frizzelli**

Chap stands for Chap Dunnings. The Chappy Dunning collection was all absorbed and donated to the museum in the early 1960's. They've had it for awhile. They had it since 1925 and then they donated it. **Tisa Matheson**

Screenshot of metadata for the Woman's Beaded Bag from the Plateau Peoples' Portal. From the Spoqín (Spokan) community. Accessed February 2019 from https://plateauportal.libraries.wsu.edu/digital-heritage/womans-beaded-bag-8.

This metadata from the Plateau Peoples' Portal contains "cultural narrative" metadata for a cultural belonging or artifact from the Spoquin community. "Cultural narrative" provides space for someone in the community to talk about the meaning, use, or importance of a cultural belonging. Many libraries across the country do not currently include "cultural narrative" in their metadata. If there was a need to share this information about the beaded bag with another library, the other system would need to provide space for the "cultural narrative" metadata. By not creating space in the new system for "cultural narrative," the information, with its rich cultural context, could be lost.

To summarize, metadata has the ability to help hide resources, tell stories about resources, and assimilate resources. We must think carefully and critically about what metadata we create in order to understand the potential losses and harm it can enable.

Section 3 Quiz questions

Metadata can be used to find and hide resources. What are some examples of how this works?

a) Limiting by facets in a library catalog
b) Discriminating against people who are labeled as belonging to certain categories
c) Filtering out all blue shoes while online shopping
d) All of the above

Which statement best reflects what you just read about the way we organize information in libraries

a) Makes no difference to people looking for materials
b) Is the best way for people to find what they are looking for
c) Can tell a story about the materials

According to this section, what is one limitation of interoperability?

a) When things are interoperable, you lose the chance to be creative
b) Interoperability doesn't always help you organize things properly
c) Information and meaning might get lost when moving to systems built with different perspectives, norms, and values in mind

Section 3 Discussion Question

In what ways is the creation of metadata the exercise of power? What are ways not already discussed that this power can affect people?

Section 4: What makes good metadata

I hope you've noticed by now that metadata is not objective. It relies on librarians' choices and judgements and the way it is used or presented creates narratives, signal values, and has the ability to be inclusive and exclusive whether intentional or not. Similarly, the answer to the question of what makes good metadata, the focus of this section, is open to interpretation.

Important vocabulary for this section

Fields _____

The NISO (National Institute for Standards Organization) Framework for Digital Collections suggests six metadata principles that presumably make good metadata[2]. Although the NISO principles can be helpful, I'd argue "Good" metadata is relative to the reasons it's created and who is being centered when it's created. Good metadata isn't achieved by including a checklist of characteristics, but through a discussion. To help with this discussion, I want to use the following questions:

1) What resource is the metadata for?
2) Why is the metadata being created?
3) Who is being centered (i.e., who is the metadata (or resource) for)?
4) Where will the metadata live?

Let's practice answering some of these questions.

What resource is the metadata for?

Pretend that I wanted to create metadata for the item depicted below. I'd first want to understand this item: a ball gown, a piece of clothing that someone could wear. What categories of metadata could I use to describe the gown? What about the gown is describable? Could I describe the color? Length?

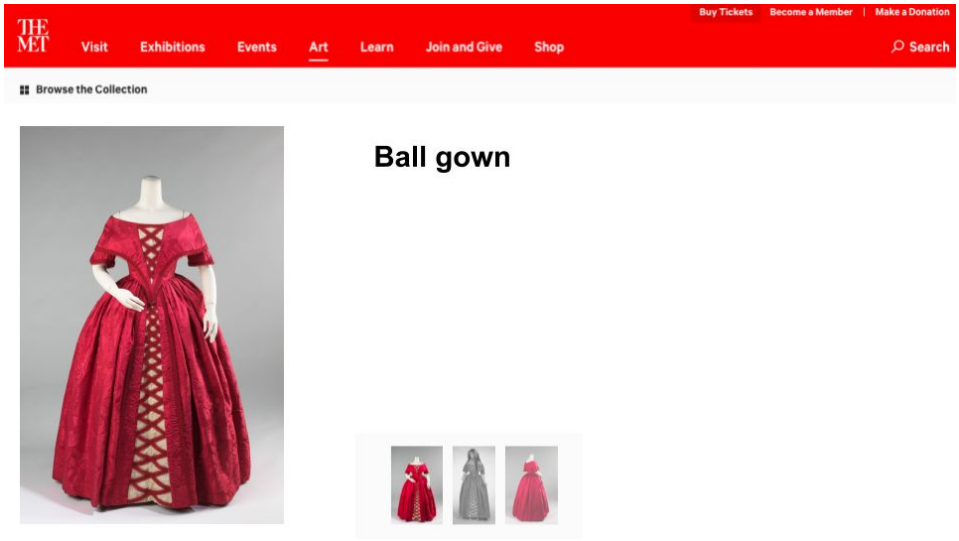
[2] The six principles can be viewed here: http://framework.niso.org/24.html.

Style? What else? What other categories could be used to describe a piece of clothing generally or a ball gown specifically? By the way these categories such as “color,” “length,” “style” are known as *fields*.

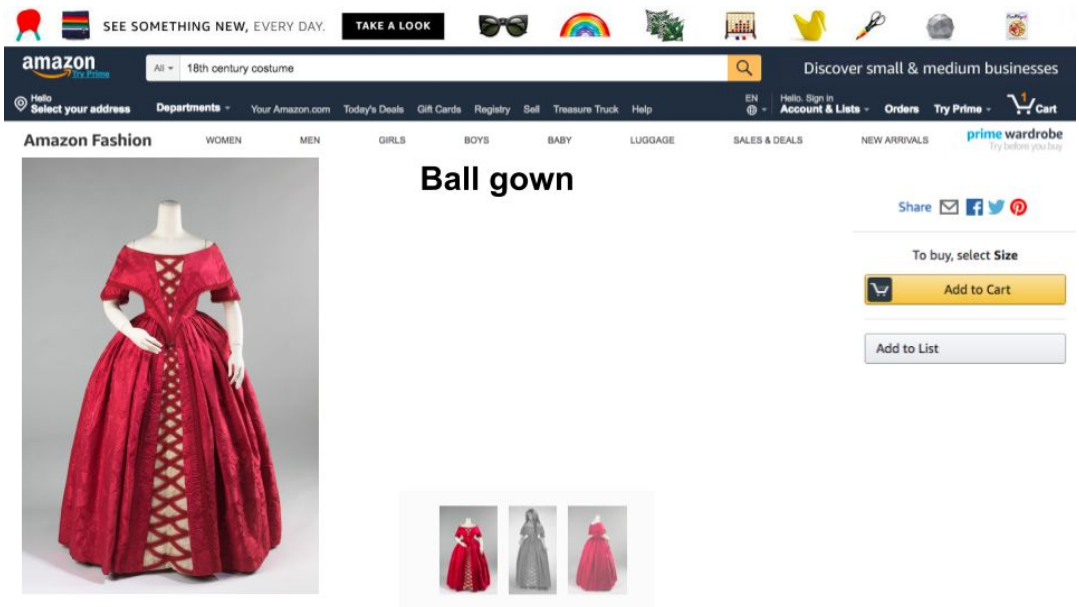
Ball Gown. Brooklyn Museum Costume Collection at The Metropolitan Museum of Art. Retrieved October 2018 from https://www.metmuseum.org/art/collection/search/155675

Why are you creating metadata?

Imagine that this Ball Gown is on display at the Metropolitan Museum of Art in New York (Met). One of the goals of the Met is to educate visitors on various types of art and artifacts and their place in history. Are there more fields (categories) to describe the ball gown now that it is in a museum? Are there any fields you used before that you wouldn’t use or that you would modify?



Now, pretend the gown is being sold on Amazon? Amazon is an ecommerce site, whose goal is to sell products online. How would your metadata fields change? What new fields might you add?



Did you notice that the fields you chose to describe the ball gown changed depending on the reason you were describing the ball gown–for history and educational purposes or for online shopping? Although some common pieces of information applied to both examples, some categories were unique to

each. What's more, if you decided to use the metadata from the Met example on Amazon, it would be missing key information to support the primary goal for Amazon to sell or for a customer to find something to buy. In this sense, you can already see how assessing the quality of metadata is relative to what is being described and why.

Who is centered?

Which audiences are you centering or prioritizing when creating metadata? Libraries and other information centers like archives and museums often claim to serve broad audiences, but the language and technology used to provide access to resources can tell a different story. Consider the following metadata from the Met:

> Object Details
>
> Title: Ball gown
> Date: ca. 1842
> Culture: British
> Medium: silk, cotton

There's nothing wrong with the metadata displayed, but take a moment to think about what kind of person might find this most accessible? One thing I noticed is the use of "ca.," the abbreviation for "circa," in the date "ca. 1842." Although "circa" or "ca." is not an uncommon word or abbreviation, it's often found in more formal and scholarly settings. The average museum visitor or even young person may not be familiar with this term and might prefer more common phrases like "approximately" or "around." I also noticed the use of "medium" to refer to the material the gown is made of. Although "medium" is very commonly used in art museums or galleries, it's a less common way to say "what something is made of" or in this case "material" than the average person expects. Based on just these two pieces of metadata alone, I'd guess the metadata was created for a more scholarly audience or at least one that frequently visits art museums. This might be fine if that is the intended

audience of the Met, but the Met's goal is to be a museum for all. So is this good metadata for anyone who is not a scholar or frequent museum visitor?

Where will your metadata live?

I'll talk about this more in a future chapter, but the quality of metadata can be relative to where it will live or be stored. Are you keeping metadata on paper or on a computer? If keeping metadata on paper, for example, it makes sense both for the person creating the metadata and the person using the metadata to keep it short and simple. Doing so reduces the numbers of errors on the part of the person creating it and the amount of reading and details to sort through as a person using it to find a needed resource.

* * *

In concluding this section, let me say that "good" metadata is always aware of its nature as metadata, which is an imperfect representation. Good metadata acknowledges these limitations and does not claim to be objective or neutral, allowing those who create it to continually reevaluate and revise.

To summarize, there are many ways to answer the question of what is good metadata. Good metadata can depend on the situation, and it is important to ask questions that help you to understand that situation such as "what is the metadata for", "why is it being created," "who is being centered," and "how is it being stored"? Additionally, actively working against potentially harmful effects of one's implicit bias can also add to the quality of metadata.

Section 4 Quiz Questions

What are two questions that I suggest in this section that can help us think about how to create "good" metadata?

a) What is metadata and why is it important?
b) Who are you creating metadata for and what will they do with it?
c) What are you creating metadata for and why are you creating the metadata?

Which statement best explains what those two questions mean?

a) Context changes the metadata you use
b) There is a way to make metadata universal to all situations
c) Be considerate when you are creating metadata

Based on what you read, which statement best explains how implicit bias is related to “good” metadata?

a) Implicit bias affects how metadata are created, so being aware and actively questioning biases can lead to better metadata
b) Implicit bias helps us function more efficiently in the world so it makes the creation of metadata more efficient
c) Implicit bias is bad and we need to eliminate it completely in order to create the best metadata

Section 4 Discussion Question

Although many businesses and public services want to serve a wide demographic of people, “you can’t be everything to everybody,” as the saying goes. Go to a frequently used website that serves a broad audience and spend a few moments assessing it. Based on the data or the metadata there, describe the type of person the website is centering and why you think so.

Chapter 1 Recommended Reading

Riley, J. Understanding Metadata: What is Metadata, and What is it For?: A Primer (2017). *National Information Standards Organization*. http://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf.

Chapter 1 References

Adler, M. (2017). *Cruising the Library Perversities in the Organization of Knowledge*. Fordham University Press.

Hodges, G. (2001). Metadata Made Simpler. *National Information Standards Organization*. Retrieved November 2020 from https://earthref.org/ERDA/download:328/.

Jan, T., & Dwoskin, E. (2019, March 28). Hud is reviewing Twitter's and Google's ad practices as part of housing discrimination probe. *The Washington Post*. https://www.washingtonpost.com/business/2019/03/28/hud-charges-facebook-with-housing-discrimination/.

Riley, J. Understanding Metadata: What is Metadata, and What is it For?: A Primer (2017). *National Information Standards Organization*. http://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf.

Chapter 2: How do you create metadata?

Now that you know more about what metadata is and why and how it is used, you may be wondering how one creates metadata. Although later chapters in this book will share more detail about the technical aspects of metadata creation, this chapter will introduce you to broad approaches for creating metadata and the use of metadata standards. This chapter will focus on how descriptive metadata is created, but the basic concepts may apply to other types of metadata as well.

Many of the metadata examples explored in this book so far have been examples of descriptive metadata. Descriptive metadata is just one type of metadata; Chapter 5 (Types of Metadata) will explore other types of metadata. “Descriptive metadata describes a resource for purposes such as discovery and identification” (Hodges, 2001). Examples of descriptive metadata shared in Chapter 1 include but are not limited to the title of a resource, dates a resource was created, and the color, size, or shape of a resource.

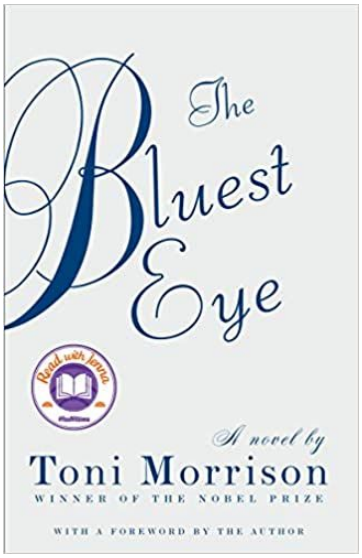
Section 1: How do you create metadata?

There are different approaches to creating metadata, but the two I have used most are recording (capturing or copying information that is observable) and analysis (examining, synthesizing, and interpreting information). Although analysis is more prone to variation, neither method of metadata creation is completely accurate. This section will discuss these two approaches to creating metadata as it relates to descriptive metadata.

Important Vocabulary in this Section
Transcribing _____
Subject analysis _____
Is-ness _____
About-ness _____
Access points _____

Recording

Transcribing information on a resource is one method of "recording" metadata. Transcribing involves copying text written in or displayed on a resource or spoken words or sounds from a resource. For example, look at the image below of the cover for *The Bluest Eye* (image copyright: 2007, Knopf Doubleday Publication Group). Imagine you want to create descriptive metadata for the book this image represents.



One way to create metadata for the book is to transcribe textual information you find on and in the book that can help someone to find it and understand what it is. Some helpful pieces of information might include "The Bluest Eye," "Toni Morrison," and "novel." You could transcribe them exactly as you see them: "The Bluest Eye Read with Jenna #readWithJenna A novel by Toni Morrison WINNER OF THE NOBEL PRIZE. WITH A FOREWORD BY THE AUTHOR."

Alternatively, you could categorize and format and select the metadata e.g., “Title: The Bluest Eye; Author: Toni Morrison; Genre: Novel”

You can create metadata also from transcribing spoken or displayed words in audio or visual resources. For example, listen to the first 13 seconds of the following audiobook: https://www.youtube.com/watch?v=eD_P84f5OOM. Just as you transcribed information about *The Bluest Eye* based on textual information read from the cover of the book, you could transcribe the words of the speaker, as they contain metadata such as the title, author, and publisher of this audiobook, *The Adventures of Pinocchio*.

Other methods of recording include but are not limited to noting quantifiable characteristics (e.g., number of pages, physical dimensions, file size), other characteristics (e.g., color, texture, shape), or creating an image of a resource (e.g., thumbnail image of a cover).

Analysis

Transcription and other forms of recording require some level of judgement, selection, and interpretation: recording everything (i.e., completely reproducing the resource) is not metadata and is not useful as metadata. Analysis requires even more interpretation than recording. Analysis requires examination of a resource in order to provide additional context for finding or understanding the resource.

Is-ness, about-ness, and subject analysis

One common type of analysis for descriptive metadata is *subject analysis*: distilling the essence of a resource into main concepts referred to as *subjects*. Subjects are typically expressed as a single word or short phrase. Two popular ways to perform subject analysis is to consider a resource’s *“is-ness”* and *“about-ness.”*

Were you to describe *The Bluest Eye*’s “is-ness,” you would use metadata that describes what *The Bluest Eye* **is**: *The Bluest Eye* **is** a novel; *The Bluest Eye* **is** a book; *The Bluest Eye* **is** fiction. In these examples, genres like

“novel” and “fiction,” as well as formats like “book” are just three ways to describe what a resource is. Were you to describe *The Bluest Eye*’s “about-ness,” you would describe topics, themes, or concepts discussed (explicitly or implicitly) within the book: *The Bluest Eye* is **about** women; *The Bluest Eye* is **about** Black people; The Bluest Eye is **about** a town in Ohio.

These subjects that describe a resource’s is-ness and aboutness can be used as *access points*, key pieces of information that aid searching or browsing for resources. Names, titles, subjects, dates, formats, are just some examples of access points.

> Which word or sentence best describes is-ness for the movie *Coming to America*? (Never saw *Coming to America*? Look it up!)
>
> a) Eddie Murphy
> b) Comedy
> c) Marriage
>
> Which option best describes about-ness for *Coming to America*?
>
> a) Prince Akeem
> b) VHS
> c) Arsenio Hall

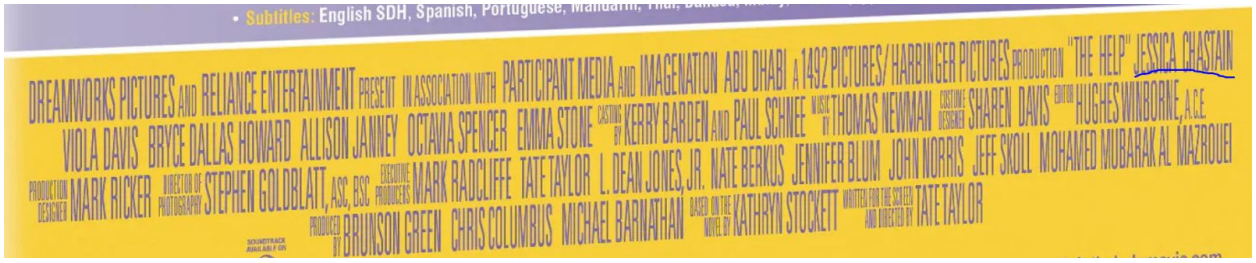
Coming up with words to describe is-ness and about-ness can be challenging even if you can pick those words from a list or copy them from the resource. Particularly challenging is subject analysis for a complex or abstract resource. What is Beyonnce’s *Lemonade*? A series of music videos? A “visual album”? A musical storybook? What is Georgia O'Keeffe’s *Music Pink And Blue II* about? Music? Synesthesia? Movement? ???

* * *

Both approaches to creating metadata—recording and analysis—require interpretation. Although it could be said that recording relies on “facts,” the selection of which “facts” and how they are presented is interpretive. There is

no such thing as objective metadata. For example, look at the images below for two editions of the DVD for the film *The Help*. One could argue that the information printed on the cases is metadata, albeit unstructured, because it is data about a resource (the movie). The DVD cases present similar information, but you may notice some differences.

Back and front (L-R) of DVD case for the movie The Help. *Back shows film credits and images of scenes from the movie. This edition is distributed by Poh Kim Corporation PTE LTD. Image retrieved December 2020 from https://pohkimvideo.com/shop/english-movies/the-help-dvd/*



Enlargement of the back of the DVD case shown in the previous image. Enlargement shows film credits with actors' names listed in alphabetical order, starting with Jessica Chastain.

Back and front (L-R) of DVD case for Special Edition of The Help. Back shows film credits and images of scenes from the movie. Image retrieved December 2020 from http://covers-layers.blogspot.com/2012/01/help-dvd-cover.html

Enlargement of the back of the Special Edition DVD case shown in the previous image. Enlargement shows film credits with actors' names listed in billing order, starting with Emma Stone.

One difference is the order in which actors' names are listed. On the first DVD, actors' names are presented alphabetically, whereas on the second DVD, actors are listed by billing, the order in which actors are credited based on their level of notoriety (perceived or actual) or screen time or both. The names of the actors in the movie are facts; what is interpretive is the way the names are presented in particular their order. Many factors go into the order in which actors' names are listed in movies and accompanying materials: contract negotiations, an actor's notoriety, screen time, importance of the role. Still, the order in which the names appear may reflect a slightly different interpretation of the facts. For example, does the second DVD, the Special

Edition of *The Help*, include more scenes with Emma Stone? Is that why she goes from being listed alphabetically (and therefore last) to listed first?

The selection of information to include, too, can be an interpretation of the facts. For example, the images from scenes in *The Help* on each DVD case are facts about the resource—these images in fact appear in the movies. Each DVD, however, uses different images. Based on the first DVD case, I would expect *The Help* to focus primarily on Viola Davis's character; based on the second DVD case, I would expect the movie to focus on Emma Stone's character and her friendships. The facts of the images remain facts, but which images are selected and how they are presented can create different focal points and thus interpretations.

To summarize, although there are many approaches to creating metadata, the two focused on in this section are recording and analysis. One type of analysis used to create metadata is subject analysis, which considers the isness and aboutness of a resource in order to identify subjects. Although recording "facts" about a resource seems straightforward and objective compared to analysis, no metadata are objective. The selection of information to record, as well as how it is presented, is a level of interpretation.

Section 1 Quiz questions

Which pair below reflects an example each of 1) recording and 2) analysis

a) 1) subjects; 2) reproduction
b) 1) transcription; 2) subject analysis
c) 1) abstract; 2) summary

According to this section, can metadata be objective?

a) No, it is always colored by interpretation of some kind
b) Yes, you just have to be very careful

Which of these examples of isness or aboutness for Toni Morrison's novel The Bluest Eye reflects the explanation of isness and aboutness?

a) Isness. Genre: novel
b) Aboutness. Length: 300 pages
c) Isness. Topic: Beauty

Section 1 Discussion Question

In what ways might the example of the DVD cases for the film *The Help* reflect inequities that could be found in metadata?

Section 2: Introduction to Standards

Recording and analysis both require a lot of decision making: how much information do you record; which information gets recorded; how is information formatted; how do you come up with words to describe isness and aboutness? Also challenging is making the same decisions each time you create metadata. Furthermore, if each person at each library is making such decisions, how different will their decisions be? If different decisions are made by different people and institutions, can their metadata be efficiently or effectively shared? To account for these concerns, librarians use *metadata standards*. An introduction to standards will be the focus of this section.

Important vocabulary in this section
Metadata standards _____
Content standard _____

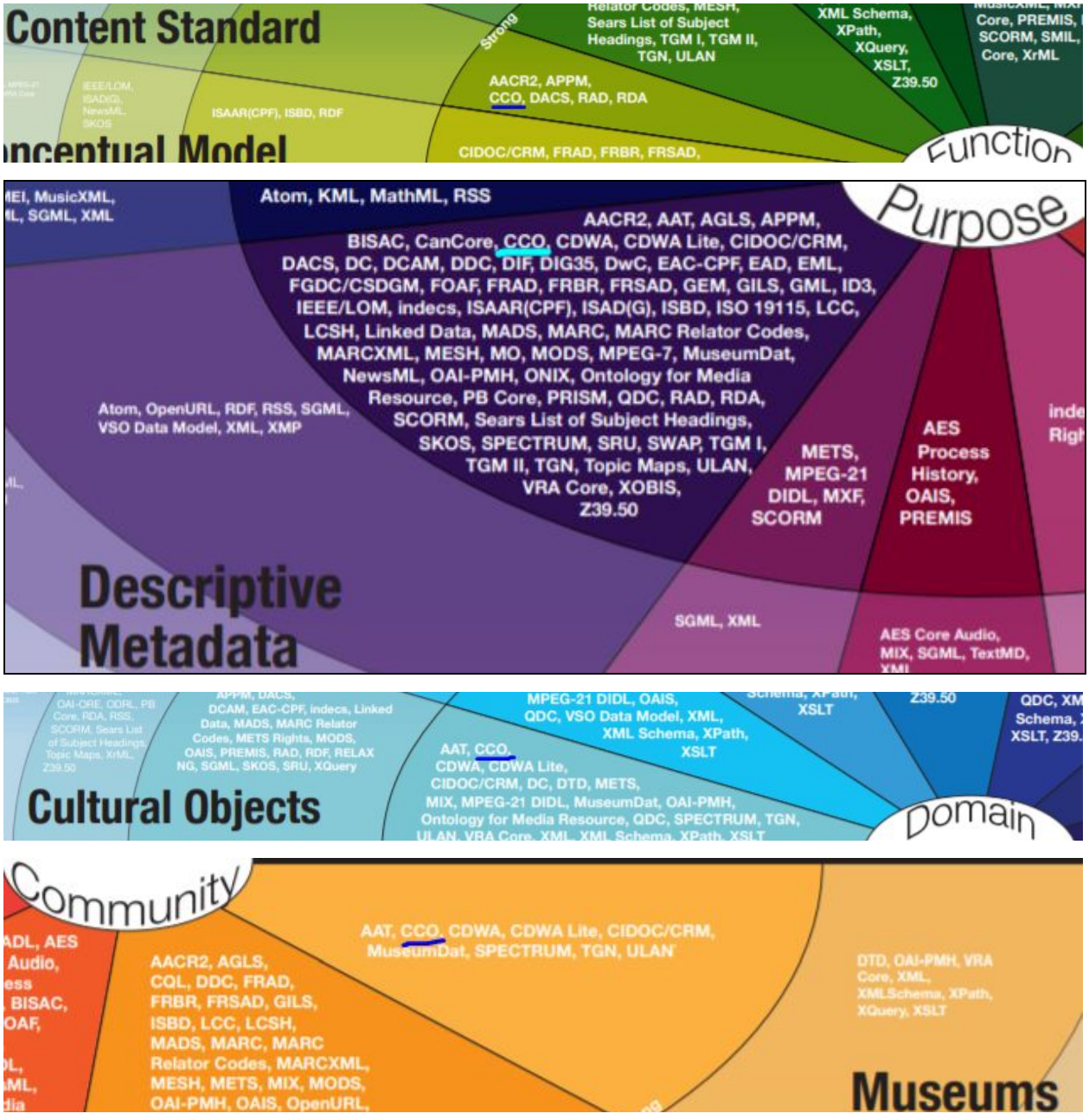
Metadata standards are guidelines, benchmarks, rules, instructions, or required structures for recording metadata . The purpose of metadata standards is to save time, aid interoperability (loss-less information exchanging between systems), and contribute to quality through consistency when creating and using metadata. You don't have to use metadata standards to create metadata, but standards can be a useful tool for the reasons mentioned above.

Many types of standards

Different types of standards support the creation of metadata in different ways.

Jennifer Riley divides standards into four broad categories in her poster *Seeing Standards: A Visualization of the Metadata Universe*: domain, community, function, and purpose (2009-2010). Riley's full poster can be viewed here: http://jennriley.com/metadatamap/seeingstandards.pdf. Riley's four broad categories are divided further into 25 subcategories. Names of standards related to each category are listed, but some standards are listed

multiple times. For example, the standard CCO (Cataloging Cultural Objects) is listed in each of the four broad categories and in several of the subcategories because it is a *content standard* (function) for descriptive metadata (purpose) of cultural heritage objects (domain) collected by museums (communities).



Screenshots from Riley's Seeing Standards *poster showing standard CCO listed in each of the four broad categories of types of standards (function, purpose, domain, and community) in order to*

illustrate that membership in categories are not mutually exclusive for any standard. http://jennriley.com/metadatamap/seeingstandards.pdf

As you can imagine, it would be difficult to discuss all 25 subcategories of standards or the hundreds of standards on Riley's portster. For that reason, the remainder of this chapter will discuss just three types of standards: content standards, controlled vocabularies, and classification systems. Future chapters will touch on a few other types of standards such as standards for different types of metadata in Chapter 3 (purpose), standards for storing and encoding metadata in Chapter 4 (function), schemas in Chapter 5 (function), and Archives metadata (community) in Chapter 8.

Content standards

The rest of this section will focus on content standards. Content standards are guidelines or "rules" that instruct how to identify, select, and structure information from a resource to create metadata. Content standards are often specific to a domain or a community in order to provide guidelines that will be helpful for the types of resources found in that situation.

In libraries in the United States, one of the most widely used content standards is called RDA (Resource Description and Access). Because libraries traditionally collect text or paper based published resources, RDA is especially equipped to help describe those kinds of resources, although it can also be used to describe non-text, non-paper, or unpublished resources, too. Similarly, DACS (Describing Archives: A Content Standard) is the default content standard for archives in the US and is especially equipped to instruct in the description of unpublished resources found in archives. CCO and CDWA (Categories of Description for Works of Art) both were particularly designed to describe museum and art objects.

Content standards are often robust and provide guidance in a number of ways. Content standards can lay out an overall philosophy. For example, RDA's "purpose and scope" section states that the purpose of the RDA is to guide the creation of metadata that supports the Functional Requirements for

Bibliographic Records (FRBR) user tasks: find, identify, select, and obtain. Content standards can also provide guidelines on what aspects of a resource should be described. For example, RDA's Section 3 provides guidelines on describing carriers whereas section 7 provides guidelines on describing content (e.g., a CD is a carrier for music, which is the content). Similarly in CCO, Section 3 covers physical description (e.g., how to record measurements and materials), whereas section 4 provides guidelines on how to describe stylistic / cultural / chronological information.

Content standards may also provide specific instructions on where to find information to record and how to structure information. RDA, for example, instructs you to use the title page as the main source of information for metadata for a book. RDA also specifies that names of people who are from English speaking countries should be recorded using the following pattern: family name, first name.

Content standards typically do not tell you which "record format" or "structure standard," as Riley calls them, to use. You will learn more about record formats and structure formats in Chapters 4 and 5. For now, think of them as the containers into which you put metadata. For example, the content standards CCO and CDWA provide instructions for what kind of information should be recorded about a cultural object, such as requiring that an object's measurements be recorded, but neither specify how the data should be contained. The information could be contained using labels such as "measurements" or "dimension" (e.g., Measurements: 19 x 23 x 8 cm; Dimensions: 19 x 23 x 8 cm) or any number of other options. Furthermore, the information could be contained on paper or in a variety of types of computer files or systems in a variety of computer languages.

Often, different types of standards will work together to provide robust guidance for creating metadata. Content standards can be combined with other types of standards you will learn about in this chapter and in later chapters.

To summarize, standards are used by libraries and other information organizations to create metadata efficiently and consistently, and also to increase interoperability. There are many categories of standards, and an individual standard may fall into more than one category. Different types of standards support the creation of metadata in different ways. Towards the end of this section, we focused on content standards, which are standards providing guidelines or instructions for identifying, selecting, and structuring metadata.

Section 2 Quiz Questions

How many broad categories and how many subcategories of metadata are listed in Jennifer Riley's *Seeing Standards* poster?

a) 2 and 8
b) 4 and 25
c) 3 and 15

What category of standard is a content standard according to Riley's poster?

a) Function
b) Purpose
c) Domain

Which of the following DOES NOT describe what a content standards does?

a) Provides a list of vocabulary to use in your metadata.
b) Tells you what type of information to record about a resource.
c) Tells you from where to select the information you record.

Section 2 Discussion Question

Look at the definition of "Function" and "Content Standard." Why do you think Content Standards are considered a subcategory of the broader category?

Section 3: Controlled vocabulary

In this section you will learn about a type of standard called a controlled vocabulary and how controlled vocabularies are used.

Important vocabulary in this section
Controlled vocabulary _____
Thesaurus / Thesauri _____
Authority record _____
Subject heading _____
Name heading _____

A *controlled vocabulary* is a predetermined list of words or phrases (referred to as terms) (Riley 2017). Controlled vocabularies try to account for the variation in people's word choices and interpretations when referring to the same thing by establishing a single word or phrase for that thing or concept. For example, some people might describe the object depicted in this image as a "soccer ball," whereas many others would label it as a "football." If I search an image site like Pixabay.com, a site on which anyone can upload and label their images, I would get slightly different results depending on which word I used. Search "futbol," another popular term for the same ball, and yet a different set of results would display. Without controlled vocabulary, a single word or phrase that can be applied to all similar images, finding all relevant images is complicated.

Although a controlled vocabulary can be as simple as a list of words, the most widely used controlled vocabularies in libraries are more complex vocabularies called *thesauri*. A *thesaurus* (singular of thesauri) is a controlled vocabulary in which each term contains definitions and references. References in a thesaurus point to other terms or words that are related in some way. References may be hierarchical: e.g., **broader term**: "footwear";

narrower term: “high heels,” “boots,” “sneakers”. References may reflect *preferred* vocabulary: e.g., **Use** “Myanmar” instead of the older term “Burma.” References may point out other relationships: e.g., “Obama, Barack” **see also** “United States. President (2009-2017: Obama).”

Different thesauri cover different types of vocabulary. The Library of Congress Name Authority File (LCNAF) and the Union List of Artist Names (ULAN), for example, are thesauri that focus on names. The Library of Congress Subject Headings (LCSH) and the Art and Architecture Thesaurus (AAT) focus on vocabulary that is more descriptive such as topics, formats, and genres. Medical Subject Headings (MeSH), as the name suggests, is a thesaurus specifically for medical and health related vocabulary.

Words such as *authority* / *authorized* or *heading* are used in controlled vocabularies to indicate the word or phrase that the controlled vocabulary wants you to use in your metadata. You may have noticed these words in the names of thesauri listed above. *Authority records* are metadata that establish and define headings. Below are two simplified examples of authority records: one for the *subject heading* “Soccer” from the Library of Congress Subject Headings (LCSH) and one for the *name heading* “Clement, Jemaine” from the Library of Congress Name Authority File (LCNAF).

LCSH *Subject heading*

Subject Heading (topic): **Soccer**
Use the heading instead of: Association football
Use the heading instead of: English football
Use the heading instead of: European football
Use the heading instead of: Football (Soccer)
You can see this related subject heading: Football
Source of information: Seasons in the fall by P. W. Dumas, copyright 2007

This above authority record tells you that the Library of Congress wants you to use “Soccer” if you want to include a subject or topic about the sport in your metadata. Below the heading are words to avoid, as well as subject headings that are related, and the source of information for the metadata.

LCNAF *Name heading*

Person name Heading: Clement, Jemaine
Associated place: Masterton (N.Z.)
Occupation: Actors; Singers; Musicians; Comedians
Gender: Males
Fuller name: Jemaine Atea Mahana
Source of information: Tongan ninja, 2004; credits (Jemaine Clement)

The name heading for the actor Jemaine Clement is “Clement, Jemaine.” Names of people in the LCNAF and many other controlled vocabularies are often represented with the family name first because of the history of alphabetical filing. Beneath Jemaine’s name is his place of birth, occupation, gender, fuller name, and sources of information. LCNAF maintains vocabulary for the names of people and organizations.

* * *

Controlled vocabulary can help metadata creators to save time and be consistent, which in turn, makes it easier for people to search for things through access points such as names, subjects, genres, etc. The establishment of controlled vocabulary, however, is an exercise of power.

How does a word or phrase get selected to be in a controlled vocabulary? Controlled vocabularies assume that a concept, person, or thing is popular enough that it would be helpful to have standardized language that can be

reused. The Library of Congress, for example, uses a concept called “literary warrant,” to establish a heading. This concept bases the establishment of headings on if and in what form these concepts or names frequent the books and other resources the library collects. Historically, however, publishing was a privilege mostly extended to white men; those from marginalized groups who did get published, were often ignored by the mostly white male librarians at the Library of Congress.

Recall that “soccer ball” or “football” are words for the same object. If we had to choose a single term to use for this object, our choice may reflect our cultural biases or that of our intended audience. Most of the world calls this sport “football” or some derivative of that; using “soccer ball” in a controlled vocabulary may demonstrate a bias in favor of nations like the United States. This example may seem harmless, but consider how cultural biases in controlled vocabulary can contribute to “othering” practices such as racism, sexism, and homophobia. Othering is the marginalization of people based on characteristics opposite of what is dominant. Othering assumes a standard way of being and sees anyone who deviates from that standard as unacceptable, dangerous, inferior, or not human.

One prominent example of othering in controlled vocabulary is well chronicled in the documentary *Change the Subject* (2019), which tells of Dartmouth students’ efforts to change the subject heading “Illegal aliens” in the library catalog to “Undocumented immigrants.” The word “alien” has an hurtful connotative meaning and is dehumanizing. Further, dehumanizing is the idea that a person could be illegal. We don’t refer to people who have been arrested as “illegal people;” why refer to this group as “illegal”? The students also argued that the term is outdated and no longer widely used. They lobbied the Library of Congress and won! Unfortunately, the Library was overruled by Congress. It should be noted that Congress has rarely if ever interceded in such library matters.

Even the language of controlled vocabulary—preferred, heading, authority—raises questions of power. Whose authority? Who has the

authority to decide which name is “preferred” and which name is a variant or subordinate?

To summarize, controlled vocabulary is a list of words or phrases used to enforce consistency in language choices in metadata. Controlled vocabulary can be a simple list or a complex thesaurus with definitions and references. Although controlled vocabulary is helpful in creating and managing metadata, implicit biases can result in codifying language that is harmful.

Section 3 Quiz Questions

All controlled vocabulary are thesauri

a) True
b) False

According to this section, which statement best explains the function of an authorized term or heading?

a) A popular word or phrase that shows up repeatedly in resources.
b) The correct term for a concept.
c) The preferred term to use as endorsed by a controlled vocabulary.

Names are found in controlled vocabularies for names, such as LCNAF.

a) True
b) False

Section 3 Discussion Question

In what ways can subject headings contribute to power over marginalized people?

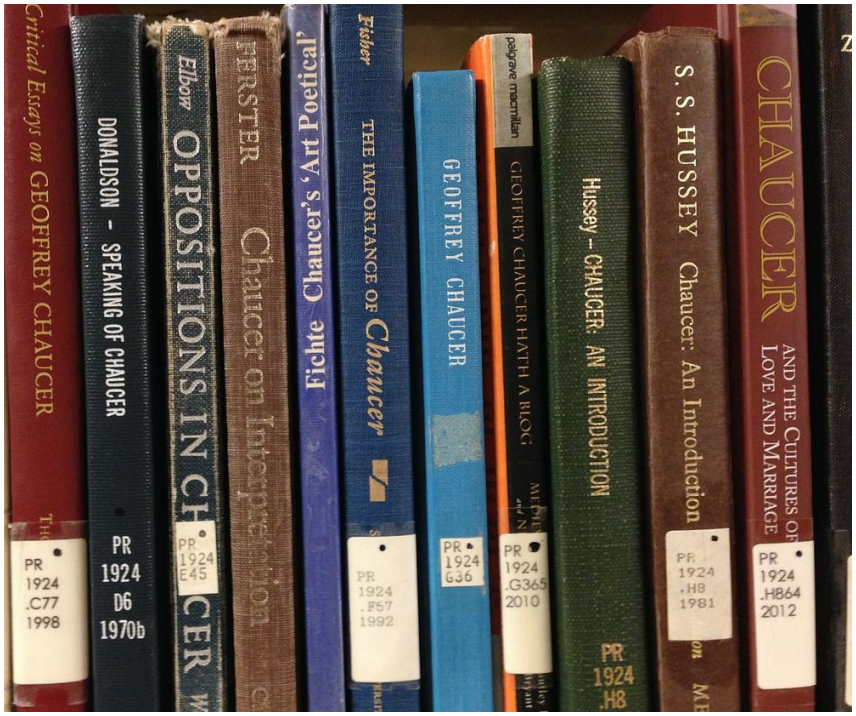
Section 4: Classification

This section will provide a brief overview of library classification systems, highlighting the two most widely used library classification systems: Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC).

Important vocabulary for this section
DDC _____
LCC _____
Classification system _____
Class _____

A *Classification system* is a type of standard. Each classification system organizes and categorizes knowledge into classes according to the goals of the system. A *class* is "a grouping of objects or concepts based on one or more [shared] characteristics, attributes, properties, qualities, etc. [...]" (Reitz, 2004). In libraries, classification systems bring order to and increase findability of resources. Resources are given a classification number that represents how the resource has been grouped. Classification numbers also correspond to shelf locations because resources are listed in the order of their classification numbers.

The image below is an example of books on a shelf in LCC number order. All the books in the photograph have classification numbers that start with "PR 1924." According to LCC, "P" is the class for literature; "PR" is the subclass for English literature. English literature is further divided into sub-subclasses indicated by numbers. The number "1924" in this case is the number used for literary criticism about Geoffery Chaucer's writings. The classification system enables all resources containing literary criticism about Chaucer's writings to be organized together for efficient locating and collocated to help people stumble across new, related resources.



The image shows books listed in order of their classification numbers. The book on the far left has a number PR 1924 .C77 1998. The next book is PR 1924 D6 1970 and is listed second because the "D" in "D6" would alphabetically come after the "C" in "C77."

There are different types of classification systems in libraries, for example, those specific to subject areas like the National Medical Libraries (NML) classification, or those specific to a nation or language like the Nippon Decimal Classification for resources in Japanese. One of the most popular classification systems, DCC, is a universal classification system. Universal classification systems attempt to categorize all of the world's knowledge.

DCC and LCC

First published in 1876, Dewey Decimal Classification (DDC) became the first widely used library classification system in the United States. The influence of neoclassical thought, which emphasized reason and order, and scientific classification can be seen in the DDC's supposition that the world's knowledge can be categorized into mutually exclusive hierarchies. In 1904, the Library of Congress published their own Library of Congress Classification (LCC). Unlike DDC, LCC was not meant to be a "universal" system, as it was developed based on the Library of Congress collection. Due to the large size

and breadth of the collection, however, LCC covered a wide range of topics and gained popularity in other libraries nationwide.

DDC and LCC are the most widely used classification systems in this country today, yet both classification systems were created well over 100 years ago by white, middle-class men (although the second and third editors of DDC were women). Though there have been many revisions to both systems, their main classes still carry the legacy of limited perspective. DDC and LCC make it possible for thousands of libraries to organize their resources and support millions of people in finding information, yet both systems reinforce structures of privilege and power enjoyed by their creators and maintainers. These structures create inequities by providing inadequate or misleading access to materials or access in harmful ways.

One example of inequity is DDC's bias in favor of Christianity: 90% of its religion categories are Christian-based. Another example is the LCC HQ subclass, "The family. Marriage. Women." The title of HQ is problematic on its own: it tethers women to family and marriage as if women don't exist outside of these contexts. Exploring further, one sees that under the sub-subclass "Sexual life," LGBTQ topics are sandwiched between "sexual deviations" and "Sadism. Masochism. Fetishm, etc." for no good reason (e.g., these are not alphabetical listings) (Adler 2017). Further, there is no mention of heterosexuality. Rather heterosexuality is assumed in all other categories in the HQ subclass.

Subclass HQ

HQ1-2044	The Family. Marriage. Women
HQ12-449	Sexual life
HQ19-30.7	Sexual behavior and attitudes. Sexuality
HQ31-64	Sex instruction and sexual ethics
HQ71-72	Sexual deviations
HQ74-74.2	Bisexuality
HQ75-76.8	Homosexuality. Lesbianism
HQ77-77.2	Transvestism
HQ77.7-77.95	Transexualism
HQ79	Sadism. Masochism. Fetishism, etc.
HQ101-440.7	Prostitution
HQ447	Masturbation
HQ449	Emasculation. Eunuchs, etc.

LCC schedule for the HQ subclass, “The Family. Marriage. Women.” Screenshot retrieved December 2020 from https://www.loc.gov/aba/cataloging/classification/lcco/lcco_h.pdf

It is important to examine not only individual classification systems, but also how we understand classification. For example, Hur-li Lee’s examination of the *Seven Epitomes* (~26 BCE) suggests that unlike the analytic and taxonomic approach to classification seen in many western classification systems, traditional Chinese classification uses correlative and holistic thinking (Lee, 2012). The *Seven Epitomes* was the catalog for the Former Han Dynasty's Chinese Imperial Library. Lee describes the *Seven Epitomes*’s classification system as ranked dichotomies that serve as arms of a whole body of knowledge rather than separate knowledge areas, as in western classification (Lee, 2012). If, as Lee concludes, the western mode of classification isn’t universal, how does this affect the way non-western materials are characterized in systems like DDC and LCC?

* * *

Classification systems were built in a world without computers in which resources were physical objects that could be placed only in one location, reflecting their classification. Resources, however, are rarely about one thing, nor are they useful to just one area of study. This begs an important question:

in a digital, interdisciplinary world, are library classification systems still needed?

To summarize, library classification systems attempt to organize and categorize library resources. This section focused on two of the most widely used classification systems in this country: Dewey Decimal Classification (DCC) and Library of Congress Classification (LCC). These systems can often reflect and reinforce racist, sexist, and additional power structures, which can lead to inequitable access to information.

Section 4 Quiz Questions

According to this reading, what is a class?

a) A place you go to study and learn about a topic.
b) A system in which you categorize resources into groupings.
c) A grouping of ideas or things based on shared characteristics.

Based on what you've learned, which conclusion about classification numbers is most accurate?

a) Classification numbers tell you what category a resource belongs to and where to find the resource on a shelf.
b) Since a resource can be about many things, libraries will give resources a classification number for each category.
c) Classification numbers are best suited to the virtual environment.

Classification systems are based on science and reflect the truth.

a) True
b) False

Section 4 Discussion Questions

Why or why aren't classification systems needed in a digital environment?

Chapter 2 Recommended Readings

International Society of Knowledge Organization article “Subject (of documents)” by Birger Hjørland. https://www.isko.org/cyclo/subject

Billey, A., Drabinski, E., & Roberto, K. R. (2014). “What’s Gender Got to Do with It? A Critique of RDA 9.7.” Cataloging & Classification Quarterly, 52(4), 412–421.

Change the Subject (2019) by Sawyer Broadley, et al. https://collections.dartmouth.edu/archive/object/change-subject/change-subject-film

Duarte, Marisa Elena, and Miranda Belarde-Lewis. 2015. “Imagining: Creating Spaces for Indigenous Ontologies.” Cataloging & Classification Quarterly 53 (5–6): 677–702.

Chapter 2 References

Adler, M. (2017). *Cruising the Library Perversities in the Organization of Knowledge*. Fordham University Press.

Baca, M., et al. (2006). Part II Elements, 3.2.1 Rules for Measurements. In *Cataloging cultural objects: CCO; a guide to describing cultural works and their images*. essay, American Library Association.

Broadley, S., Jill Baron, Cásares, Ó. R. C., & Padilla, M. (2019). Change the Subject. https://collections.dartmouth.edu/archive/object/change-subject/change-subject-film (accessed October 2020).

Harpring, P. & Baca, M. (2016). 6. Measurements. In *Categories for the description of works of art* (Revised). essay, J. Paul Getty Trust.

https://www.getty.edu/research/publications/electronic_publications/cdwa/ (accessed December 2020).

Lee, Hur-Li. (2012). Epistemic foundation of bibliographic classification in early China: A Ru classicist perspective. Journal of Documentation. 68. 10.1108/00220411211225593.

Library of Congress. Subject Cataloging Division. (1988). *Classification. Class P. Subclasses PN, PR, PS, PZ. Literature (general), English and American literature, fiction in English, juvenile belles lettres.* 3rd ed. Washington: The Library .

Library of Congress. Subject Cataloging Division. Library of Congress Classification Outline. Class HQ, https://www.loc.gov/aba/cataloging/classification/lcco/lcco_h.pdf (accessed December 2020).

Morrison, T. (2007). Book cover from The Bluest Eye. United States: Knopf Doubleday Publishing Group. https://www.google.com/books/edition/The_Bluest_Eye/12_KUGLXigMC?hl=en&gbpv=0 (accessed November 2020).

Reitz, J. M. (2004). *Dictionary for library and information science*. Westport, Conn: Libraries Unlimited.

Resource Description and Access. RDA Toolkit, http://rdatoolkit.org/ (accessed December 2020).

Riley, J., & Becker, D. (2009-2010). *Seeing Standards: A Visualization of the Metadata Universe*. Seeing Standards. Accessed December 2020 from http://jennriley.com/metadatamap/.

Riley, J. (2017). Understanding Metadata: What is Metadata, and What is it For?: A Primer. *NISO*. http://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf (accessed October 2020).

Satija, M. P.. *The Theory and Practice of the Dewey Decimal Classification System*, Elsevier Science & Technology, 2013. *ProQuest Ebook Central*, https://ebookcentral-proquest-com.libproxy.lib.unc.edu/lib/unc/detail.action?docID=1575007 (accessed December 2020).

Chapter 3: Storing and Encoding Data

This chapter discusses various ways metadata has been stored and encoded historically, as well as related current practices and trends. As with most of this textbook, the topics covered in this chapter could have books written about them, so here is presented brief overviews highlighting a few commonly used methods.

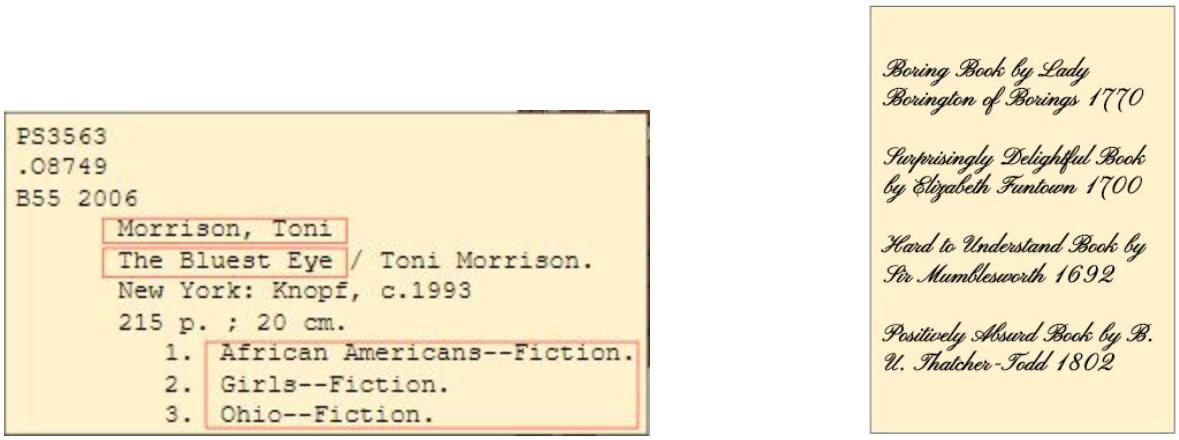
Section 1: Brief history of storing metadata

In this section, I will give an overview of ways metadata has been stored in libraries throughout Western history. As I learn more about metadata traditions not centered around European or US traditions, I hope to revise this section to include multiple histories.

Important Vocabulary for this Section
Catalog _____
Record _____
MARC _____
Field _____
OPAC _____

Over the last two millennia in the western world, metadata went from being stored on scrolls to being stored on computers. Jeffery Pomerantz outlines this history of western metadata in the introduction to his book *Metadata* (2015). The *Pinakes* (~240BCE) is the Greek bibliography of the Library of Alexandra largely accepted as the oldest example of library metadata. The *Pinakes* was written on scrolls, but over the next few centuries, the codex, what we now call a book, replaced scrolls as the primary way to store data—and metadata.

These collections of metadata, whether on scroll or in a book came to be referred to as a *catalog*, a listing of resources held by a library or any other place that holds a collection of resources (e.g., an archive; a museum).



Left: Card catalog example of metadata for The Bluest Eye. Right: Codex (book) catalog example with metadata for several fictitious titles on one page of the codex.

According to Pomerantz, the card catalog was invented in France as early as around the time of the French Revolution. Card catalogs contained similar if not more metadata to the codex or book catalog, but instead of metadata for multiple resources listed on the same page in the order of their location, separate cards each contained metadata for a different resource. Cards catalogs were easier to revise than book catalogs: a library could add and remove cards for new or discarded resources in the appropriate alphabetical order. Like the book catalog, however, searching across records was limited to a few access points.

Both the printed book catalog and card catalog relied on consistent data formatting based on local, and later national standards such as *Rules for Descriptive Cataloging in the Library of Congress* (1949), *AACR* (1967), *AACR2* (1978), and those long before them. The formatting provided a predictable experience for people using a metadata *record*, a group of metadata about a single resource, to find resources. These standards in structuring metadata would influence how metadata is stored in computers.

```
LEADER 00000cam a2200313 a 4500
001    29356775
005    19970407091548.0
008    931029r19931970nyu           000 1 eng
010       93043124
020    0679433732 :|c$22.00 ($29.00 Can.)
040    DLC|beng|cDLC|dNOC
100 1  Morrison, Toni.|0http://id.loc.gov/authorities/names/
       n80131379
245 14 The bluest eye /|cToni Morrison ; with a new afterword by
       the author.
250    1st Knopf ed.
260    New York :|bKnopf :|bDistributed by Random House,|c1993.
300    215 pages ;|c20 cm
336    text|btxt|2rdacontent
337    unmediated|bn|2rdamedia
338    volume|bnc|2rdacarrier
650  0 African Americans|0http://id.loc.gov/authorities/subjects/
       sh85001932|zOhio|0http://id.loc.gov/authorities/names/
       n79049197|vFiction.|0http://id.loc.gov/authorities/
       subjects/sh99001562
650  0 Girls|0http://id.loc.gov/authorities/subjects/sh85055012
       |zOhio|0http://id.loc.gov/authorities/names/n79049197
```

Screenshot of a record for Toni Morrison's The Bluest Eye*, encoded in MARC21. MARC's numeric fields are highlighted in the picture on the left; the specific 100 field for author and 245 field for title information are also highlighted.*

Around 1968, the Library of Congress developed *MARC* (MAchine Readable Cataloging), a computer language for encoding library metadata. MARC took formatting metadata further than the book and card catalog by coding metadata into *fields*, categories of metadata, which separated parts of a catalog record. For example, authors would go into an author field; titles would go into a title field; and so on. This allowed each field in a record to be independent, theoretically allowing someone to search that field across many records in the catalog. The computer could also store more metadata for resources, making descriptions more robust.

Despite MARC, print and card catalogs were the predominant public access catalog for decades, waning during the 1990s with increased affordability of computers and the development of Online Public Access Catalogs (*OPAC*). The *OPAC* is often referred to as a "discovery layer" because these user friendly search tools are a layer of technology between the library user and the database where MARC records are stored (the Library Management System (LMS) or the Integrated Library System (ILS)).

Paper metadata has not completely disappeared; they are still used for personal collections and in some small and specialized libraries today. Some libraries also keep microfilm or microfiche (super tiny photographs) versions of these print resources for historical uses.

MARC is still possibly the most prolific form of computerized library metadata in the world, used primarily to create library *bibliographic metadata*, metadata for published materials, but also for digital resources as well as for some archival and museum resources.

To summarize, metadata has historically been stored on paper in various forms and on computers. The one of the most notable computerized metadata formats in libraries today is MARC, an encoding language that formats catalog records into metadata fields, which are then stored in a relational database.

Section 1 Quiz Questions

What is a record in the context of this Section?

a) A collection of music
b) A written account of an event
c) A group of metadata related to a resource

Which of these best defines what a catalog is based on the reading?

a) Metadata that describes each resource held by anyone with a collection of resources
b) Bibliographic metadata for a library stored in MARC
c) An OPAC, Online Public Access Catalog, which is also referred to as a discovery layer

Which of the following is a broad definition of a “field” in metadata, according to this reading?

a) A category of metadata
b) Metadata about a wide open space with grass
c) A numeric code used in MARC

Section 1 Discussion Question

What are the advantages and disadvantages of the various examples in this section of how metadata has been stored (i.e., paper vs electronic; book vs card vs MARC “record”)?

Section 2: Tabular and Relational Data

In this section, you'll learn about tabular and relational data and how metadata is stored using these structures.

Important Vocabulary for this Section
Tabular data _____
Relational database _____
Delimited text files _____
CSV _____

Structure makes data more usable, and "structured data" is part of the definition of metadata. For example, consider the difference between the following sentence and the formatted metadata below it:

The Bluest Eye (1970) is a novel by Toni Morrison taking place in an African American community in 1940s Ohio.

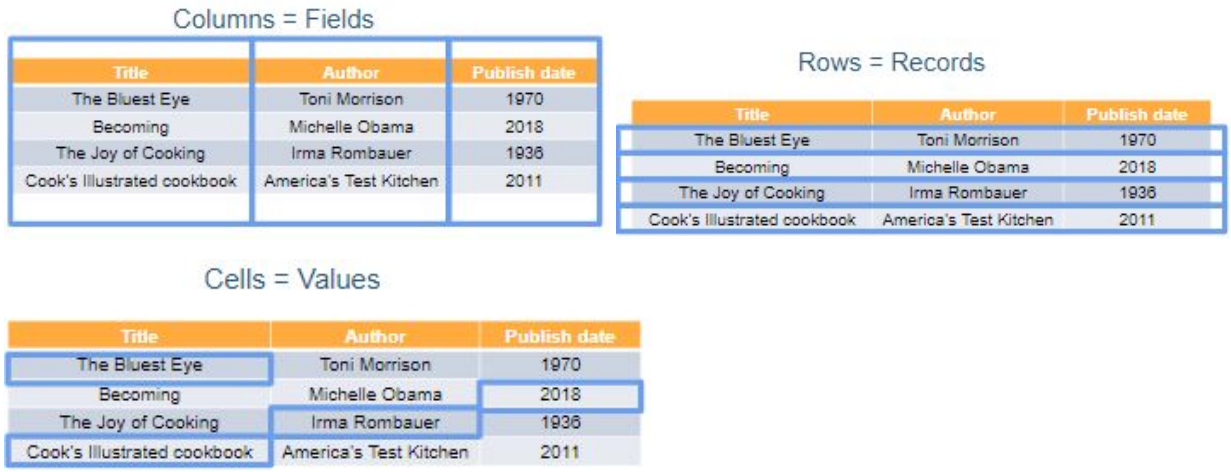
vs

Title: The Bluest Eye
Genre: Fiction
Author: Toni Morrison
Original publication date: 1970
Subject matter: 1940s Ohio
Subject matter: African Americans

In the first example, while we could use keyboard shortcuts like ctrl + f to find the specific title and author, it would be difficult to identify every title and author if we had 700 similar descriptions. In the second example, we'd be able to more easily find title, author, and other information across 700 similar descriptions because the information is separated and formatted into data

fields. Instead of searching for individual titles or authors, we might be able to get a list or report of all 700 titles or authors.

If you are familiar with spreadsheets, the idea of separating metadata into fields might be familiar to you. In a spreadsheet, data is stored in columns and rows. Columns represent the *fields* (categories) of data stored; names of those fields are at the top of each column. Rows represent records, which are groups of data about the same resource. Cells in a row are *values*, individual pieces of data that correspond to a field. This format is often referred to as tabular data (tabular for table).



Three images illustrating columns, rows, and values. Top left, Columns = Fields: columns "Title," "Author," "Publish date" are highlighted using vertical lines. Right, Rows = Records: rows are highlighted using horizontal lines; each row contains the record for a different book. Bottom left, Cells = Values: various cells are highlighted using boxes; each cell contains a piece of information within a record that relates to a specific field (category). Original design concept by Thu-Mai Christiansen.

Spreadsheets like Microsoft Excel or Google Sheets are computer programs that not only store tabular data, but also provide tools to sort, filter, and otherwise manipulate data. Popular alternatives to spreadsheets for storing tabular data are *delimited text files*. Unlike spreadsheet programs, delimited text files merely store data, although they are easily imported into spreadsheet programs. Common delimited text files include Comma Separated Values (*CSV*) and Tab Separated Values (TSV). These files use commas (CSV),

tabs (TSV) or other delimiters (e.g., pipe, double dagger, dollar sign, etc.) to separate columns or fields of information.

myCatalog.csv

```
title,author,publish date
The Bluest Eye,Toni Morrison,1970
Becoming,Michelle Obama,2018
The Joy of Cooking, Irma Rombauer,1936
Cook's Illustrated Cookbook,America's Test
Kitchen,2011
```

CSV (Comma Separated Values) example. Field names (columns) are established in the top row. Each field is separated by a comma. Subsequent rows contain the records for each resource. Commas separate the values within a record into their corresponding field. Imported into a spreadsheet application such as Excel, this tabular data will resemble the example above. Plain text computer applications such as Windows Notepad can be used to create and save files such as this.

The development of personal computing technology since the 1980s provided easier access to these tabular data formats for the general public. As such, these data formats are helpful alternatives for individuals, communities, and information centers that don't have the resources to create complicated MARC metadata in expensive Library Management Systems (LMS).

We won't discuss databases in detail in this book, but it's important to know that *relational databases* can be viewed as complex tabular data. Relational databases contain multiple tables that are linked together through relationships. Relational databases make it possible to create metadata that is multidimensional. For example, rather than storing information about resources and the people who created them in each record, the two can be described separately and linked together, creating two dimensions of description. Just as with Microsoft Excel, the availability of programs like Microsoft Access and File Maker Pro to the general population has also made it possible to create and store metadata in databases for those who do not have the resources or inclination to do so in MARC based systems.

To summarize, data formats provide structure to store data and make them more useful. Tabular data formats such as spreadsheets, delimited files, and databases can be used as alternatives for storing metadata for those who can't or don't want to work with MARC.

Section 2 Quiz Questions

What do columns represent in tabular data?

a) Resources
b) Fields or categories of data
c) Structures that keep a building up

Based on what you learned in this section and section one, which of the following statements is most accurate?

a) Numeric MARC fields are similar to tabular data columns in the role they play in a metadata record
b) The information recorded on a card catalog could never be recorded using tabular data
c) Each row within a MARC record is like each row of tabular data within a spreadsheet.

Section 1 Discussion Question

Discuss the reasons people might choose to use tabular data to create and store metadata rather than a more established standard like MARC.

Section 3: XML and JSON for metadata

In this section, I will focus on the computer languages XML and JSON as formats for storing, encoding, and representing metadata. Although XML and JSON are popular languages for digital library metadata, they require more time and training to learn than the tabular data formats discussed in the previous section.

Important Vocabulary for this Section
XML _____
Tags _____
Value (metadata value) _____
Root element _____
Tree _____
JSON _____
Key:value pairs _____

XML

MARC is still the dominant way to structure and store metadata in libraries today, but since the 1990s, other options have developed and flourished especially in digital libraries. In 1995, the Online Computer Library Center (OCLC) developed a metadata standard called Dublin Core, which we'll discuss in a future chapter. Dublin Core is arguably the most widely used standard for digital libraries today. Although Dublin Core can be implemented using any number of languages and formats, it was first designed to be implemented in a language called *XML*.

In the library context, *eXtensible Markup Language* (XML) is a computer language used to encode and store metadata and documents. The "markup" in XML refers to the process of tagging or marking information in a document in order to flag it for processing in a specific way or to separate metadata into fields. This section will focus on XML's use as a means to create metadata records, but it is important to note the use of XML to encode textual

documents for digital use through such projects as the Text Encoding Initiative (TEI).[3]

An XML record does not separate data into fields using columns as with tabular data. Rather, XML uses sets of *tags* (also referred to as fields or elements) to establish the name of a data field. Tags are recognizable because they are enclosed in angle brackets. A field or element consists of an opening tag and a closing tag. Data are stored between the opening and closing tags like filling in a sandwich. XML has no predefined tags, so tags are made up just as you would make up names for columns in a spreadsheet.

Compare the "title" field from this table of metadata for *The Bluest Eye* with the XML below the table:

Title
The Bluest Eye

<title>The Bluest Eye</title>

In this example, the table contains a single record (row) for the novel *The Bluest Eye.* The record for the novel contains metadata related to one field (column) called "title." The cell containing the text "The Bluest Eye" is the metadata value corresponding to the "title" field. The same metadata is stored between two sets of XML tags in the XML example. The opening tag is the name of the metadata field "title," and the closing tag is almost identical except that the name of the field is preceded by a forward slash. The *value* of the "title" field, which is the text "The Bluest Eye," is placed between the opening and closing tags.

In tabular data, each row is a new record, but in XML, records are stored as separate files or separate blocks of code within one file. For example, the following table of metadata containing two records, would be two blocks of

[3] For more information on the TEI, visit https://tei-c.org/

code within the same file or single code blocks in two separate files (depicted below) in XML.

Title	Author	Published
The Bluest Eye	Toni Morrison	1970
Becoming	Michelle Obama	2018

Record 1

```
<book>
        <title>The Bluest Eye</title>
        <author>Toni Morrison</author>
        <published>1970</published>
</book>
```

Record 2

```
<book>
        <title>Becoming</title>
        <author>Michelle Obama</author>
        <published>2018</published>
</book>
```

You may notice a couple things about the XML examples. First, there is an XML tag called <book> not present in the spreadsheet example. Second, the <title>, <author>, and <published> fields are indented and sandwiched between the opening and closing <book> tags. The <book> tag in these examples serves as the *root element.* The root element is the foundation of an XML document or record. Imagine that the root element in XML acts like the lines that form the borders of a data table. The metadata that makes up the record is indented to visually indicate that the <title>, <author>, and <published> information is contained inside of the <book> element.

XML differs from tabular data in that a XML record can support hierarchy. Hierarchy in computing is often referred to as a *tree*. We can talk about hierarchy in XML like we do a family tree. XML elements can have children (subelements), parents, and siblings. For example, if we wanted to say more about the author in the records we created, we could add child elements of <author> to capture details such as <name> and <dateOfBirth>. Adding these subelements, our first record would look like this:

```
<book>
        <title>The Bluest Eye</title>
        <author>
                <name>Toni Morrison</name>
                <dateOfBirth>1931</dateOfBirth>
        </author>
        <published>1970</published>
</book>
```

XML files are plain text files with a .xml extension. You can create these files using any plain text editor such as Notepad (Windows) or Text Edit (Mac) and save the file as .xml. On its own, XML doesn't do anything except store and encode data; it is not a computer programming language and cannot perform actions. Rather, actions can be performed upon XML. Computer programs or processes can be run on an XML file in order to display it in certain ways, extract information from it, sort, filter, or search it, and other possibilities.

There are many criticisms of XML.[4] One interesting criticism I've heard in conversation with colleagues (I'm sorry I forget who!) is about how the hierarchical structure of XML can codify one's value judgements metadata. Which metadata fields get to be "parent" fields and which are established as "child" fields; what do those choices say about the kinds of information we value? This criticism would apply to any format that supports hierarchy.

[4] See *XML Sucks* at https://wiki.c2.com/?XmlSucks for a listing of various critiques. Arguments in support are found on the counter-part site *Benefits of XML* at https://wiki.c2.com/?BenefitsOfXml.

JSON

JSON (JavaScript Object Notation) is a data format that emerged more recently than XML, around the early 2000s. Like XML, JSON was not developed for libraries; it was created by web developers working with the Javascript programming language. As with XML, however, libraries have realized JSON is useful for storing data. Because of its readability and relationship to Javascript, JSON is increasingly used for web based digital library metadata.

XML and JSON have similar metadata uses. Both store data in fields; both support hierarchy; both can't perform actions on their own. Structurally, the two are very different. Whereas XML stores data in between a set of tags to form a data field, JSON uses *key:value pairs*. To illustrate this difference, look at the following snippets of XML and JSON:

```
<author>Toni Morrison</author>
```

```
{
        "author" : "Toni Morrison"
}
```

In the JSON example, the "key" is the data field "author" and the "value" is the information captured, "Toni Morrison." You may have also noticed the curly braces in the JSON above. Along with key:value pairs, another feature of JSON are objects. Anything inside a set of curly braces in JSON is an object. Hierarchy in JSON is established by putting objects inside of other objects. Again, comparing JSON to a familiar example of XML from earlier:

```
<author>
        <name>Toni Morrisonz</name>
        <dateOfBirth>1931</dateOfBirth>
</author>
```

```
{
        “author”:{
                        “name”:“Toni Morrison”,
                        “dateOfBirth”:“1931”
                }
}
```

In the JSON example, the “name” and “dateOfBirth” elements are enclosed in curly braces, indicating they are an object. That entire object is the “value” to the “author” key.

To summarize, XML (eXtensible Markup Language) is a computer language that encodes or marks up text documents and stores metadata. Like tabular data, it can store records about resources with data separated into fields. Unlike tabular data, XML uses pairs of tags in angle brackets to denote fields and supports hierarchy in records. JSON is used in libraries in similar ways but features the use of curly braces, objects, and key:value pairs.

Section 3 Quiz Questions

Which of the following options appropriately captures the data from this table in XML

Photographer
Sonoe

a) <photographer>sonoe
b) <photographer>sonoe</photographer>
c) <sonoe>photographer</sonoe>

Which of the following options best illustrates the use of the root element?

a) <photo>

```
        <title>Landscape</title>
        <photographer>Sonoe</photographer>
    </photo>
```

b) `<title>Landscape</title>`
`<photographer>Sonoe</photographer>`

c) `<photo></photo>`

```
        <title>landscape</title>
    <photo></photo>
```

Section 3 Discussion Question

What criticisms or advantages do you see for XML compared to other data forms, languages, and formats discussed so far in this chapter?

Section 4: RDF

This section introduces Resource Description Framework (RDF) and discusses RDF as a data structure for metadata.

Important Vocabulary for this Section
RDF _____
Triples _____
Graph _____
Serialization _____

The World Wide Web Consortium describes the Resource Description Framework (*RDF*) as “a framework for representing information in the Web” (2004). One goal for RDF is to have a simple data model that enables the automated processing of information on the Web (W3C, 2004). This goal has led libraries, archives, and museums to become interested in RDF as a framework for representing metadata on the Web.

RDF allows information and relationships between them to be explicitly defined through triple statements or *triples*. According to RDF, we can describe resources using this syntax (grammar) that has a three part structure:

Subject - predicate - object

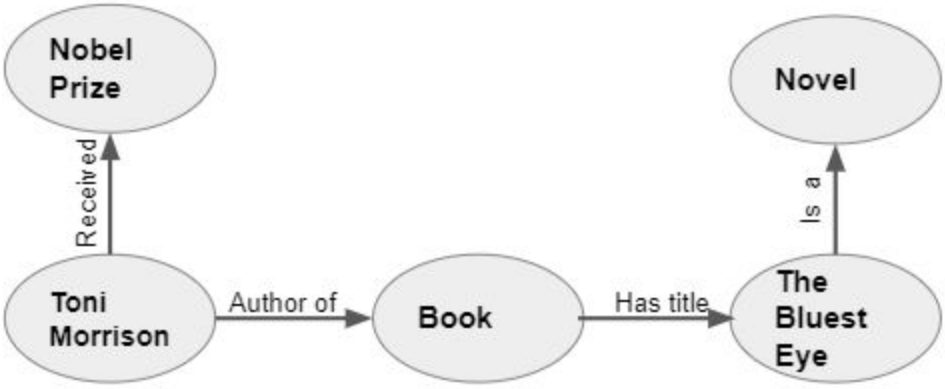
This is a triple. In the context of metadata, within this model, the subject is the resource being described; the object is the description or the metadata value; the predicate is the relationship between the two, or the metadata field. Look at the statement below and see if you can tell which parts are the subject, predicate, or object.

Toni Morrison - is author of - Book.

“Is author of” describes the relationship between the two pieces of data “Toni Morrison” and “Book.” This makes “is author of” the predicate of the statement. “Toni Morrison” is the subject of the statement, and “Book” is the object. Objects can become the subject of other triple statements. In this triple statement, “Book” is now the subject of a new triple:

Book - has the title - *The Bluest Eye*

“Book” is the subject; “has the title” is the predicate; “*The Bluest Eye*” is the object. If we join the two triple statements together, they form what is called a *graph*, a collection of triple statements. Below is a visual representation of a graph consisting of the two triples above and two additional triples (“Toni Morrison received the Nobel Prize” and “*The Bluest Eye* is a Novel”).



RDF is not a language like XML or JSON. As a framework, RDF establishes the subject - predicate - object grammar for how to structure data. RDF has many *serializations*, however, which are languages that can code these triple statements. JSON-LD, a form of JSON specifically for coding RDF triples, is one such serialization. Others include but are not limited to RDF-XML, N-triples, and Turtle.

Example of RDF triples as encoded in different languages or serializations including N-Triples and Turtle.

To summarize, XML is not the only metadata format used in digital libraries; JSON and formats supporting Linked Data are quickly gaining in popularity and use. Like XML, JSON stores fielded data and supports hierarchy. Unlike XML, JSON uses key:value pairs and the concept of objects. As Linked Data emerges as the next generation of metadata, XML, JSON, and other languages are increasingly being used to support the subject-predicate-object triples model for structuring data.

Section 4 Quiz Questions

Predicates describe the relationship between a subject and an object. Which comparison below is most appropriate based on the examples provided in this reading?

a) Predicates are like subjects of a resource
b) Predicates are like rows in a spreadsheet
c) Predicates are like metadata fields

RDF is a language

a) True

b) False

A an RDF serialization is RDF in a specific language

a) True
b) False

What is a graph in RDF?

a) A way to chart RDF success
b) A collection of triples
c) An image

Section 4 Discussion Question

In what ways does the RDF structure of triples change the way metadata is represented that is different from the other formats we've learned about?

Section 5: Data types

In this section, you will learn about data types and why it is important to understand these data types when creating and managing metadata.

Important Vocabulary for this Section
Data types _____
Character _____
String _____
Date _____

Data types categorize data so that they can be grouped based on common characteristics. Although different domains will recognize or define different data types, the following are commonly found and understood in library metadata: *character*, *string*, and *date*. *Integer*, and *geospatial* are also common, but are less frequently seen in library metadata.

Character

The *character* data type is a single unit of textual information. It can be a letter (a), digit (5), space ( ), or symbol (e.g., punctuation marks and symbols from Wingding fonts). The following is a table that shows a few examples of the *character* data type.

Character examples
a
5
“
&

Based on the definition above, which of the following is **NOT** an example of a character?

a) 2
b) The
c) □

Metadata fields containing a single character are most common in fields that use codes for filtering or indexing purposes and are often hidden from public view. For example, if a shoe seller's website uses "b" for all blue and purple shoes, "r" for all red and orange shoes, and "g" for all green and yellow shoes, they will be able to sort shoes into those categories simply using those letters.

String

A *string* is a series of characters. A string could be two characters or multiple paragraphs of text. This is the **most common** type of data found within a metadata field. Below are just some examples of what strings can look like.

String examples
22 cm.
Nakasone, Sonoe
This is a paragraph. This is another paragraph. This is yet another paragraph & sentence.
<this>could be a string too</this>

Date

A *date* is a string that uses specific formatting and vocabulary so that it can be recognized as a date (a day, month, or year, or any combination). Dates can be formatted in many ways. The following are some examples of the date data type. Different ways of formatting the date are displayed in each row.

Date examples
1990
05-05-1990
5/05/1990
May 5, 1990

Integer and Geospatial

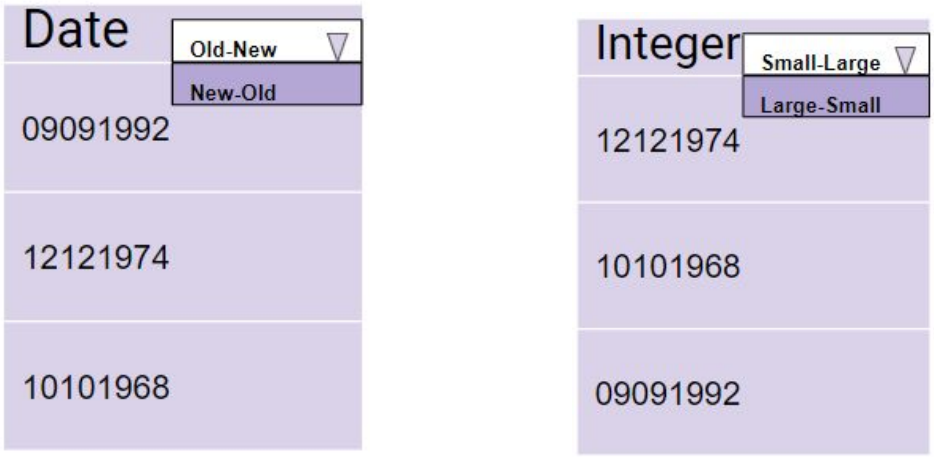
An *integer* is a string that is treated as a number. Sometimes digits are treated as numbers. For example, as a number or *integer*, the string "530591210" would be 530,591,210 (five hundred thirty million five hundred ninety-one thousand two hundred ten). As an integer, this number could be used for mathematical functions and has a value larger or smaller than other integers. Digits can also be treated as plain strings. For example "530591210" could be a social security number (530-59-1210) or a telephone number (530-591-210) or a randomly generated ID for a book. These do not have meaningful value as integers and are used for other purposes.

Geospatial is a type of data that relates to geographic and spatial information typically referring to coordinate information. Coordinates often come in pairs of latitude and longitude.

* * *

Why do different data types matter? Data types allow for searching / browsing, sorting, and filtering resources based on metadata. A metadata field with a *date* data type, for example, can be used to sort and filter resources based on whichever dates are recorded in the metadata. Imagine there are three books published on 9 September 1992, 12 December 1974, and 10 October 1968, respectively. In the image below on the left, each of these dates is represented as an eight digit string and has a *date* data type. The dates are sorted from the most recent to oldest date (new-old): the last four digits of each string are the years 1992, 1974, and 1968. When browsing

resources by date or limiting to specific date ranges, the strings must be recognized as dates to provide meaningful order to the resources. In the image below on the right, the same three dates are treated instead as integers. Integers can only be sorted by numeric value, i.e., largest to smallest / smallest to largest, so treating these dates as integers does not allow them to be sorted in a meaningful order.



Data types should also not be mixed because that will affect the ability to manage or manipulate metadata. For example, if the string “before the 20th century” is entered into a metadata field with a *date* data type, how does this metadata get sorted or filtered? How can this data be recognized as a date? Sometimes mixing data types is unavoidable because of constraints on time or other resources, but it’s important to understand the consequences.

In summary, the most common data types you will encounter in metadata are *character*, *string*, and *date*, although there are other data types. It is important to know how to differentiate between data types because each type of data allows us to search for, sort, and filter resources in different ways.

Section 5 Quiz Questions

Based on the definitions and examples provided for string, could a string include all the words and sentences in a book?

a) Yes
b) No

Based on what you learned from this section, why might you try to avoid storing a phone number in the same field as a date?

a) It will look ugly.
b) The way the numbers are formatted look similar, so that might confuse people.
c) They are different data types so it could negatively affect the ability to sort and filter the data.

In the example earlier showing the consequences of treating a date like an integer, what happened to the dates when treated as integers?

a) Sorted as integers, the dates were out of order because the oldest date was treated as the second largest number.
b) Sorted as integers, the dates were in exactly the same order they would have been were they sorted as dates.
c) The dates got reformatted when they were treated as integers.

Section 5 Discussion Question

Go back through this chapter and even a previous chapter to look at the data types you see illustrated. Which data types are most frequent, and why do you think that is?

Chapter 3 Recommended Readings

W3schools.com tutorials for XML:
https://www.w3schools.com/xml/default.asp

And JSON: https://www.w3schools.com/js/js_json_intro.asp

Chapter 3 References

World Wide Web Consortium. (2004, February 10). Resource Description Framework (RDF): Concepts and Abstract Syntax. https://www.w3.org/TR/rdf-concepts/.

Joint Steering Committee for Development of RDA. A Brief History of AACR. http://www.rda-jsc.org/archivedsite/history.html#:~:text=AACR2%20was%20adopted%20by%20the,and%201985%20(published%201986) (accessed December 2020).

Pomerantz, Jeffrey. Metadata, MIT Press, 2015. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/unc/detail.action?docID=4397948. Last accessed on November 21, 2020.

Chapter 4: Types of Metadata

So far, many of the examples of metadata shared in this book have been about only one type of metadata, descriptive metadata. In this chapter, you will learn more about what descriptive metadata is as well as learn about other common types of metadata in libraries such as structural and administrative metadata.

Section 1: Introduction to types of Metadata

In this section, I will discuss the three main types of metadata: descriptive metadata, structural metadata, and administrative metadata.

Important vocabulary in this section
Descriptive metadata _____ Structural metadata _____ Administrative metadata _____

Jenn Riley in “Understanding Metadata: What is Metadata, and What is it For?: A Primer” (2017), provides a neat outline of different types of metadata. Riley’s table, recreated and modified below, illustrates the significance of understanding different types of metadata: different metadata support resources in different ways.

Metadata Type	**Example categories of information**	**Primary Uses**
Descriptive metadata	Title; Author; Subject; Genre; Publication date	Discovery; Display; Interoperability
Structural metadata	Sequence; Place in hierarchy	Navigation
Technical metadata (administrative)	File type; File size; Creation date/time	Interoperability; Digital object management; Preservation
Preservation metadata (administrative)	Checksum; Preservation event	Interoperability; Digital object management; Preservation
Rights metadata (administrative)	Copyright status; License terms; Rights holder	Interoperability; Digital object management

Recreated from Understanding Metadata: What is Metadata, and What is it For?: A Primer *by Jenn Riley, 2017.*

Descriptive

In the last chapter, most of the examples of metadata shared were of *descriptive metadata*, which focuses on characteristics of resources for the purpose of finding and understanding them (Riley, 2017). Examples include but are not limited to metadata fields such as title, author, related dates, size, genre, and subject matter. Riley's NISO article suggests descriptive metadata is useful for "discovery," by which Riley means the ability to find, identify, and explore resources. Riley also notes that descriptive metadata is used for interoperability, the ability to go between one system and another without information loss, and for display purposes. Descriptive metadata can also provide information that explains a resource, for example, what a resource is (e.g., it is a book; it is a novel) and what a resource is about (e.g., about women; about Ohio).

The majority of metadata users of a library encounter are a mixture of different types of metadata with descriptive metadata being the most dominant. Descriptive metadata describes a resource's characteristics including what the resource looks like (or in some cases sounds, feels, smells, or tastes like), and what kind of information is contained within a resource.

Structural

The word "metadata" seems to have entered the lexicon in 1968 when Philip Bagley, a computer scientist, coined the term. As a computer scientist, Bagley was not talking about information about library resources. Instead, Bagley was referring to descriptive and structural information about data elements within databases. Although the concept of structural metadata originally referred to databases, this concept is now also applied to library (and archive, museum, etc.) resources, particularly digital resources.

In libraries, *structural metadata* captures information about how a resource is constructed such as how many parts it has and relationships between parts, for example, sequence and hierarchy. Riley notes that structural metadata is

used to navigate different parts of a resource (2017), much in the way a map would allow you to navigate to and between different parts of a city. Because of this, structural metadata helps to explain how a resource is organized so people and computers know how to use it. For digital resources, structural metadata can help a computer program understand how a resource should display based on its structural metadata. Some examples of structural metadata might include, but are not limited to the order of the pages in a book, the organization of a book into chapters, or the relationship between folders and subfolders in a file system.

Administrative

Just as the “administrative” branch of an organization might focus on the overall management and care of an organization, *administrative metadata* is concerned with the use, management and preservation of a resource. Administrative metadata can help to organize and locate resources as well as help to explain the origins of a resource. Administrative metadata is an umbrella term and can be divided into a few main categories of metadata. In his book *Metadata* (2015) Jeffery Pomerantz outlines five types of administrative metadata: technical, preservation, provenance, rights, and meta-metadata. To that list I would add methods metadata and use metadata / access controls. Some of these types of administrative metadata will be discussed further later in this chapter.

To summarize, there are three main types of metadata: descriptive metadata, structural metadata, and administrative metadata; each type performs different functions of metadata. Descriptive metadata describes a resource's characteristics such as what it looks like and what it is about. Structural metadata describes a resource’s parts and their relationship to each other. Administrative metadata captures information about the use and management of a resource. Most metadata found in libraries is a combination of all three.

Section 1 Quiz Questions

What are the three types of metadata?

a) Administrative, descriptive, and discursive
b) Structural, administrative, and descriptive
c) Productive, structural, and aboutness

Which would be an example of structural metadata?

a) Page order
b) Hierarchical relationships
c) Number of parts to a resource
d) All of the above

Which is not a use for structural metadata discussed in this section?

a) Explain how a resource is used
b) Navigation
c) Discovery

Section 1 Discussion Question

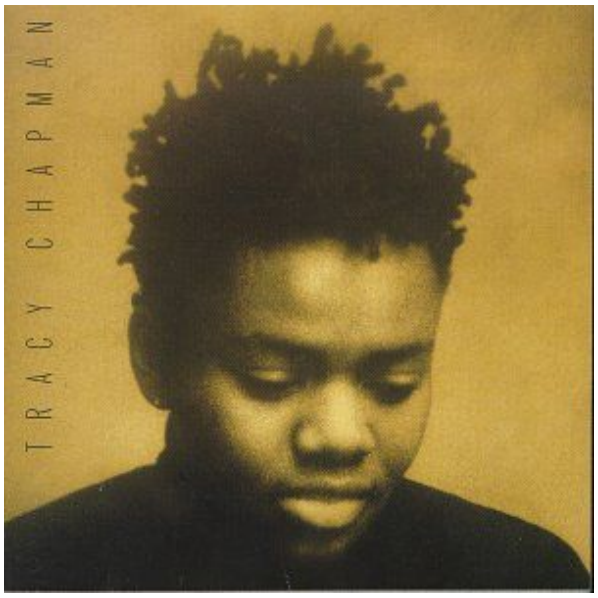
In what ways does each of the three types of metadata relate to the definition of metadata you learned in Chapter One, Section One?

Section 2: Structural Metadata

This section focuses on structural metadata and how it contributes to the use and management of resources.

Important vocabulary in this section
Sequence _____
Hierarchical relationship _____
Parent-child relationship _____
Sibling relationship _____

Structural metadata provides information about how a resource is structured, including information about the number of parts that make up the resource and how those parts are related to each other and the whole (part-to-whole-relationship). One familiar example of structural metadata is *sequence*, for example, the order of songs on an album.



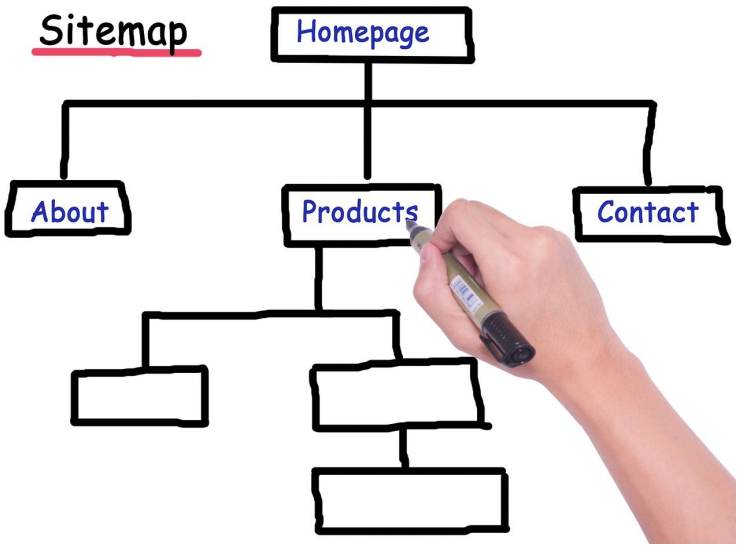
1. "Talkin' 'bout a Revolution" – 2:40
2. "Fast Car" – 4:57
3. "Across the Lines" – 3:25
4. "Behind the Wall" – 1:50
5. "Baby Can I Hold You" – 3:14
6. "Mountains o' Things" – 4:39
7. "She's Got Her Ticket" – 3:57
8. "Why?" – 2:06
9. "For My Lover" – 3:12
10. "If Not Now…" – 3:01
11. "For You" – 3:10

Cover for album "Tracy Chapman" by Tracy Chapman. Retrieved May 2020 from https://en.wikipedia.org/wiki/File:Tracy_Chapman_-_Tracy_Chapman.jpg

On the back of this album *Tracy Chapman* is a list of songs. The song titles themselves and the length of the songs are descriptive metadata, but you'll

notice these songs are numbered and listed in a certain order. The numbers and the order the songs are listed in are structural metadata. This is metadata that tells the person interested in the album what order they can expect the songs to play. If this were a CD or digital playlist, similar structural metadata would be embedded in the CD or song files to tell the computer in what order to play the songs.

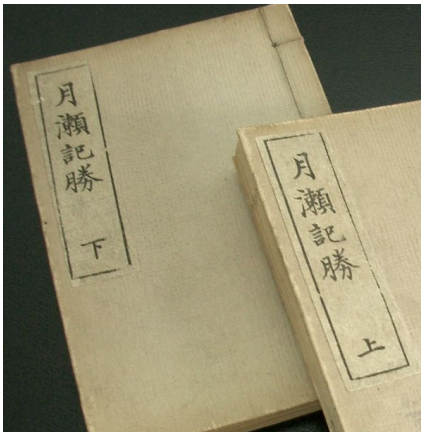
Another example of structural metadata is metadata about the *hierarchical relationships* between parts of a resource. Hierarchy in and of itself is not structural metadata. Hierarchy can exist with or without metadata. Structural metadata, however, can describe hierarchy in order for people or computers to understand how resources can or should be navigated. A sitemap is one example of this kind of metadata. Sitemaps are data about websites that tell you how a website is structured.



In the image above (from pixabay.com), a site map explains how the pages on a website are organized. At the top of the website hierarchy is the home page. Beneath the homepage are three pages: about, products, and contact. The relationship between the homepage and the three pages below it is called a parent-child relationship. Similar to a family tree, the homepage sits at the top of the hierarchy and is the parent to the “about,” “product,” and “contact” pages. Because “about,” “product,” and “contact” share the same parent, they

have a sibling relationship. This kind of visual map is structural metadata that can explain to a person how a site is organized in order to help them understand how to better use the site or manage it. This metadata can also be stored in a computer language so that a computer can interpret it and better understand how to display, store, and manage the website and its resources.

Structural metadata can also describe relationships between parts of a resource beyond sequence and hierarchy. For example, think of a postcard that has been digitized (scanned or digitally photographed so that there is a digital representation of the postcard). In the physical world, a postcard usually has information on two sides: on one side often has an image or design; the other side has a message, an address, and a stamp. There isn't an order to a postcard–some people will consider the message side to be the front, others the opposite. What is important is that the resource contains information on two sides. For a digital representation of a postcard, structural metadata can specify to a computer that two files (i.e., one for each side of the postcard) that together make a whole postcard and that these represent a two-sided object.



Describing the structure of a resource can be challenging because it forces one to interpret and characterise the structure of a resource. In the example above, a postcard has two sides, but which is "front" and which is "back" is a matter of interpretation. Another example: the structure of a traditional Japanese book, with the opening facing to the left as depicted in this image (Wikimedia Commons, 2007), could be misinterpreted and mischaracterized by someone unfamiliar with Japanese books. In each case, one's assumptions about a resource's structures can affect the way resources are presented to users and which narratives and viewpoints are reinforced.

To summarize, structural metadata provides information on how a resource is structured and how parts of a resource relate to one another. Structural metadata might include information about the sequence or order of parts, hierarchical relationships within a resource, or other specific relationships between a resources parts.

Section 2 Quiz Questions

Which of these was not an example of structural metadata from the section you just read?

a) A site map for a website
b) The name of songs on an album
c) The order of songs on an album

Structural metadata provides information about how a resource is structured. Which of the following is the best example of this based on what you've read?

a) The number of pages in a book
b) The relationship between six digital images that make up the six sides of a digitized cereal box.
c) The title of a website

Section 2 Discussion Question

In the postcard example above, why is structural metadata so important in a digital library? What do you think would happen if there was no structural metadata for that example?

Section 3: Types of administrative metadata

In this section, I'll focus on four types of administrative metadata: rights metadata, preservation metadata, provenance metadata, and technical metadata. Each of these administrative metadata types serves a special function to support the use and management of a resource.

Important vocabulary in this section
Rights metadata _____
Preservation metadata _____
Provenance metadata _____
Technical metadata _____

Rights metadata

Rights metadata outlines the allowable **uses** of a resource. It is based on legal restrictions and allowances such as who has the legal rights to possess, use, alter, share, publish, and make money from a resource. Rights metadata can also include information about ethical restrictions based on rights, claims, and desires not recognized by United States law.

One common aspect of rights metadata is copyright, which is the legal right that protects tangible creative expressions of authorship (copyright.gov). The word "creative" is broad. If you write a textbook—that is considered a "creative" work. If you create a documentary—that is considered a "creative" work. Metadata about copyright might include the name of the copyright holder, the copyright status or rights statement, copyright year, or other details about copyright.

It is common to use licenses or include a license statement as rights metadata, especially for digital resources, which are easily copied, modified, and redistributed. A license is a contractual agreement that grants certain rights to a resource that is in copyright. Licenses can become complex. Resources such the Creative Commons (CC) assist people in creating

licenses for their own resources. Creative Commons offers several licenses based on key rights such as commercial rights or the ability to adapt or reuse. For example, my textbook is made available with the Creative Commons license CC BY-NC-SA 4.0. The licence notes that users must attribute content from this book to me (BY), that my book cannot be used for commercial purposes (NC), and that if anyone uses, adapts, or builds upon any part of my textbook, they must use these same license terms (SA).

Although licenses can impose almost any kinds of restrictions on copyrighted materials, they cannot be used on materials that are not in copyright or to impose restrictions by those who do not hold a resource's copyright. Many Indigenous groups and groups who have experienced colonization, plundering, disenfranchisement, and enslavement have encountered difficulty imposing restrictions on belongings that were illegally or unethically obtained long ago. Such examples include the Parthenon Marbles (aka Elgin Marbles) "given" (this is disputed) to Britain by Ottoman occupiers of Greece, Indigenous artifacts "legally" held by museums but obtained through dubious means,[5] or images and other belongings of enslaved African Americans that were the "legal" property of enslavers.[6]

Tools such as Traditional Knowledge (TK) Labels for Indigenous communities and Social Human Labels, geared at a broader audience, allow communities to provide context outside of U.S. law to educate people on respectful access and use of cultural heritage resources. Local Contexts, the initiative from which TK Labels emerged, notes that the labels "offer an educative and informational strategy to help non-community users [...] understand its importance and significance to the communities from where it derives and continues to have meaning" (localcontexts.org).

[5] The 1979 article "Indian Rights: Native Americans versus American Museums: A Battle for Artifacts" by Bowen Blair outlines two such case studies of negotiations for the repatriation of traditional artifacts.
[6] For example, read about Tamara Lanier's suit against Harvard over her ancestor's image: https://www.thecrimson.com/article/2020/1/22/harvard-coalition-free-renty/.

Left: TK Secret / Sacred (TK SS) label. Indicates that the resource contains secret/sacred information and has specific conditions of access and use. It asks users to discuss any potential use. https://localcontexts.org/tk/ss/1.0
Right: SH Anonymize Me Label. Indicates the creator wants to be anonymous. https://www.docnow.io/social-humans/sh-c-am.html

Provenance metadata

Provenance metadata records information about the origins of a resource. The concept of provenance will be discussed in further in Chapter 8 (Archival Metadata). Where, when, and how was a resource created? Who owned, bought, or sold a resource? Information tracking the origins and chain of custody of a resource can help manage a resource over the long term by verifying authenticity. Information about the circumstances surrounding a resource's creation can also help explain the value, significance, and use of that resource. For these reasons, there is some overlap of provenance metadata with rights metadata (e.g., provenance metadata may show a clear chain of custody, or "legal" or cultural ownership) and preservation metadata (e.g., helping to verify authenticity helps to ensure the correct information is preserved).

Provenance metadata is especially important for resources held by archives and museums. In both communities, provenance metadata helps people understand the historical value of a resource and explain that value to learners. Imagine, for example, provenance metadata for a 18th century style ball gown says that the gown came from a Hollywood studio. This metadata is a clue that the ball gown is more likely a costume from a film or TV show than an authentic 18th century gown, which potentially changes its value.

Similar to the example above, one prominent example of provenance metadata in libraries is tracking the source of a resource's acquisition. Sometimes this metadata serves to credit a donor for a gift to the library; for resources acquired from historically significant persons, tracking such information can add historical or monetary value to a resource. In the case of e-resources that are licensed from vendors, provenance metadata can help the library to manage access. For example, a resource could be available from multiple vendors that have different rules about who has access and how the resource is accessed. Resources can be matched with their rules for access more easily if provenance metadata is available to note which resource came from which vendor.

Preservation metadata

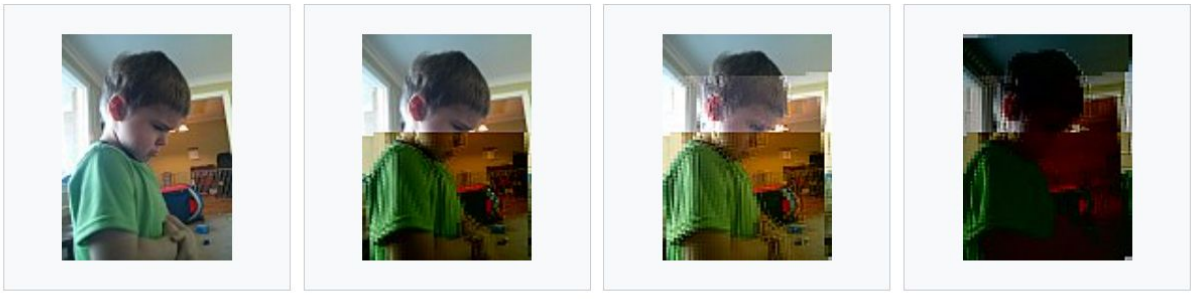
Preservation is the prevention of loss or deterioration of a resource or the information it represents. *Preservation metadata* records actions taken on a resource for preservation purposes, information that could impact preservation, or deterioration that has already occured. For physical resources, preservation metadata could include information like the summary and date of treatment for a resource as well as the name of the person who performed the treatment (e.g., "Spine repaired, November 2009—Sonoe Nakasone"). Preservation metadata is not much different for digital resources and might include the summary and date of a preservation activity as well as the name of those (people, hardware, or software) involved (e.g., "Clam AV [software] virus check, November 18, 2014 10:40PM: successful").

PREMIS (PREservation Metadata: Implementation Strategies) is a popular metadata standard used for digital preservation. PREMIS describes Objects (digital resources), Agents (people, hardware, or software performing actions), Rights (rights metadata for objects, held by agents), and Events (actions taken on objects, performed by agents) for the purpose of preservation. Often, PREMIS is used to create a log of preservation events. Below is an abridged PREMIS Event in XML reflecting the earlier virus check example.

```
<event>
        <eventType>Virus check</eventType>
        <eventDateTime>2014-11-18T10:40:57</eventDateTime>
        <eventDetail>File scanned; no virus.</eventDetail>
        <linkingAgentIdentifier>
                <linkingAgentIdentifierValue>ClamAV</linkingAgentIdentifierValue>
        <linkingAgentRole>Software</linkingAgentRole>
        </linkingAgentIdentifier>
        <linkingAgentIdentifier>
                <linkingAgentIdentifierValue>ClamAV</linkingAgentIdentifierValue>
        <linkingAgentRole>Software</linkingAgentRole>
        </linkingAgentIdentifier>
</event>
```

You may not understand what all the PREMIS fields mean, but can you locate the metadata that tells you what kind of preservation activity took place? What about the date and time? What metadata tells you which agent performed this activity?

A fixity check, a test to see if a file has changed and therefore potentially lost information, is possibly the most common preservation action. Files are made up of units of information called bits. A fixity check will check to see if a file's bits have changed. The Wikipedia article "Data Degradation" illustrates the problem of bit changes.



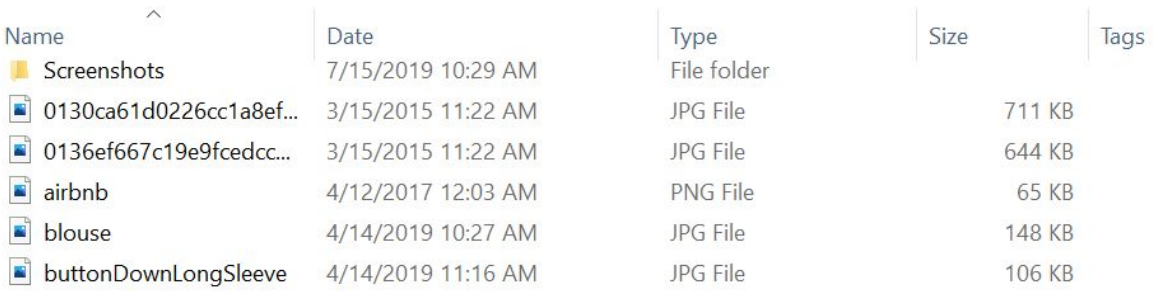
Bit rot in JPEG files, series. Shows an original image (far left), followed by subsequent images in which 1, then 2, then 3 bits from the original are flipped. Image credit: Jim Salter, October 2013, https://commons.wikimedia.org/wiki/File:Bitrot_in_JPEG_files,_0_bits_flipped.jpg (accessed 2020).

The article explains that though each of the images above contains the same number of bits (326,272 bits), a bit flip (bit changing from a 0 to 1 or 1 to a 0) can cause perceptible changes. In this example, the change in the bits results in the loss of visual information.

Some PREMIS events don't track preservation actions, but information pertaining to provenance, such as when and how a digital object is created. In these cases, provenance metadata can support the preservation of a resource by helping to verify the authenticity of the resource, thereby ensuring the integrity and preservation of its information. A full list of PREMIS event types are found on the Library of Congress's linked data service site: https://id.loc.gov/vocabulary/preservation/eventType.html

Technical metadata

Unlike some of the other types of metadata we'll discuss, *technical metadata* is unique to digital resources. Technical metadata records characteristics of a digital file, as well as contextual information about the file, such as how it was created. Technical metadata examples include but are not limited to file size (ex. 50MB), file format or MIME type (ex. PDF), and file name (ex. letter.doc). Other information that might be considered technical metadata is the name of the software program that created a file or the type of format a file was migrated from. Below is a screenshot of my computer's Pictures folder. Can you spot the technical metadata?



Name	Date	Type	Size	Tags
Screenshots	7/15/2019 10:29 AM	File folder		
0130ca61d0226cc1a8ef...	3/15/2015 11:22 AM	JPG File	711 KB	
0136ef667c19e9fcedcc...	3/15/2015 11:22 AM	JPG File	644 KB	
airbnb	4/12/2017 12:03 AM	PNG File	65 KB	
blouse	4/14/2019 10:27 AM	JPG File	148 KB	
buttonDownLongSleeve	4/14/2019 11:16 AM	JPG File	106 KB	

Above, the types (i.e., File folder, JPG, PNG), dates, file sizes, and file names are all technical metadata. This metadata was automatically extracted from my files by my operating system. Digital libraries have similar resources for technical metadata, for example, the File Information Tool Set (FITS), a set of tools that identify and extract technical metadata from files. The metadata that the FITS tools extract are then converted to and stored in FITS XML. Visit the FITS website to see examples of FITS XML: https://projects.iq.harvard.edu/fits/fits-xml

You may notice some overlap of technical metadata with descriptive metadata (e.g., size, format, name) and with provenance metadata (e.g., software program that created a file; precious file formats). Some aspects of technical metadata will also overlap with preservation metadata, checksum, for example. A checksum is an alphanumeric string that is calculated to summarize the contents of the resource. A file's checksum is technical metadata, but the metadata about fixity checks that make sure a file's checksum, and therefore the file, hasn't changed is preservation metadata.

To summarize, types of administrative metadata discussed in this section include rights metadata, preservation metadata, provenance metadata, and technical metadata, although there are some other types of administrative metadata types not covered here. Each of these serves various functions that aid our understanding, use, and management of a resource.

Section 3 Quiz Questions

Which of the following metadata fields would capture rights metadata?

a) Checksum
b) Copyright holder
c) PREMIS event

Which type of administrative metadata focuses on information about the origin of a resource?

a) Provenance metadata

b) Rights metadata
c) Preservation metadata

Which type of administrative metadata records information about a digital file's characteristics?

a) Technical metadata
b) Access metadata
c) Preservation metadata

Section 3 Discussion Question

From your understanding of administrative metadata types, discuss in what ways administrative metadata might be useful for a digital image created by scanning a photograph.

Chapter 4 Recommended Readings

Pomerantz, J. (2015). Chapter 4: Administrative Metadata and Chapter 5: Use Metadata from Metadata. The MIT Press.

Christen, K. (2015). Tribal Archives, Traditional Knowledge, and Local Contexts: Why the “s” Matters. Journal of Western Archives, 6(1).

Otto, J. J. (2014). Administrative metadata for long-term preservation and management of resources: A survey of current practices in ARL libraries. Library Resources & Technical Services, 58(1), 4-32.

Chapter 4 References

Bagley, P. R. (1969). Extension of Programming Language Concepts. United States: National Bureau of Standards, Institute for Applied Technology.

FITS: Introduction. https://projects.iq.harvard.edu/fits (retrieved December 2020).

Haider, S. (2020). Library of Congress Classification (LCC) History and Development. Librarianship Studies & Information Technology. https://www.librarianshipstudies.com/2017/11/library-of-congress-classification-history.html (accessed December 2020).

Image of “the Portable manuscript books of Tsukigase Kisho/Getsurai Kisho by Saito Setsudo in 1884” (Tsukigase-Kisho-Manuscript-Books.jpg) (September 2007). https://commons.wikimedia.org/wiki/File:Tsukigase-Kisho-Manuscript-Books.jpg (accessed December 2020 from Wikimedia Commons).

License Chooser. Creative Commons. https://chooser-beta.creativecommons.org/ (accessed December 2020).

Pomerantz, J. (2015). *Metadata*. The MIT Press.

Riley, J. (2017). Understanding Metadata: What is Metadata, and What is it For?: A Primer. NISO. http://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf (accessed December 2020).

U.S. Copyright Office. Copyright in General. https://www.copyright.gov/help/faq/faq-general.html (accessed December 2020).

SH-A Labels. Social Humans. https://www.docnow.io/social-humans/sh-a-labels.html (accessed December 2020).

Traditional Knowledge (TK) Labels. Local Contexts. https://localcontexts.org/tk-labels/ (accessed December 2020).

Wikimedia Foundation. Data degradation. Wikipedia. https://en.wikipedia.org/wiki/Data_degradation (accessed December 2020).

Chapter 5: Schemas

This chapter discusses the purpose and use of schemas, some of their features, and specific examples of schemas used in libraries.

Section 1: Introduction to Schemas

In addition to various data structures and languages, digital libraries heavily rely on schemas to create, store, and use metadata. This section will focus on defining what schemas are and discussing their purpose.

Important vocabulary for this section
Schema _____
Semantics _____
Syntax _____
Optionality _____

A schema is a specific type of standard. The International Standards Organization (ISO) defines schemas in their standard 23081.1 s3 Terms and Definitions:

"[In regards to library metadata] a schema is a logical **plan** showing the relationships between **metadata elements**, normally through establishing **rules** for the use and management of metadata specifically as regards the **semantics,** the **syntax** and the **optionality** (obligation level) of values."

As the ISO definition explains, a schema is a plan. It provides a set of metadata elements / fields / properties. It provides definitions (that's what *semantics* means in this context) for elements. They provide rules for how to arrange and use those elements properly (*syntax*). They provide information about what is required and what is optional or recommended (*optionality*).

Schemas are standards that contain highly specialized vocabulary, which are the metadata elements, subelements, and other components that store metadata as fields. Below is a table documenting a fictional schema called Sonoe's Chair Schema (SCS); take a moment to consider what makes SCS a schema in light of the definition I shared earlier.

Sonoe's Chair Schema (SCS) documentation

Property name	Definition	Rules
name	Name of the chair.	String. Unlimited character length field. Required. Not repeatable.
manufacturer	A type of creator. The person or organization responsible for manufacturing or building the chair.	String. Recommended vocabulary: Getty Union List of Artist Names. Required if known. Repeatable.
date released	The year the chair was released to the public as available for sale or viewing.	Date. Use four digit year YYYY. Required if known. Not repeatable.

What part of the SCS documentation tells you about metadata elements?

a) Property names
b) Rules
c) Definitions

There is much evidence that this is a schema based on the ISO definition from schema. The property names listed above are the metadata elements and subelements. The definitions provide meaning for the property names, so they are the *semantics*. Rules that discuss data types and ways to format and record metadata are the *syntax*. Rules that discuss if something is required or not is the *optionality*. All of this together forms a plan that someone can use to record metadata about a chair.

Try it yourself. If you were going to create metadata using SCS for a chair that was first available for sale on December 4, 1999, what would you put in the "date released" field?

a) December 4, 1999
b) December 1999

c) 1999

Some schemas contain features that will remind you of content standards and controlled vocabularies. For example, the “date released” element in SCS requires the date format “YYYY;” information on how to format data is also something many content standards provide. Some other schemas limit the vocabulary allowed for specific metadata elements like a controlled vocabulary. That said, schemas serve a unique purpose. Furthermore, schemas, controlled vocabulary, and content standards can be used together to complement each other. For example, both the content standards Cataloging Cultural Objects (CCO) and Categories of Description for Works of Art (CDWA) can be implemented using the CDWA Lite schema. Another example, Metadata Object Description Schema (MODS) often recommends Library of Congress controlled vocabulary to standardize vocabulary in certain fields.

If schemas are sets of elements, what makes a schema different from just any set of elements you might invent when storing data in a spreadsheet, xml, json, etc? The answer: semantics and syntax. Schemas provide definitions for elements and rules for how to use them. Violating the rules and definitions of a schema won’t land you in jail, but doing so may limit the usefulness of your metadata.

Pretend both your catalog and another library’s catalog use Sonoe’s Chair Schema (SCS) to create descriptive metadata records. We would assume both your institutions’ metadata is interoperable—meaning you could use each other’s metadata in your respective catalogs without a problem. There’s a problem, however: the other library is using the SCS “name” field for the name of the manufacturer, whereas your library follows the SCS definition for “name,” which is the name of the chair. When it comes time to share records, your ability to sort, filter, or search by the “name” field is affected because your records have names of chairs and the other library’s records have names of manufacturers.

Like any plan, a schema is created to make life easier for those creating metadata. The definitions, rules, and element names make it so people can create metadata that is consistent without having to reinvent them for each record. Like any standard, however, schemas can reinforce cultural and structural biases and hide those behind goals of efficiency and consistency. For example, most popular, international, library metadata standards were created by western, English speaking countries: what structural or linguistic biases or inequities have resulted from that? We should continually evaluate schemas, examine the result of biases, and be open to proposing changes to schemas or using different schemas. Although we will learn more about two schemas (DCMI Metadata Terms and MODS) later in this chapter, it is also important to note that hundreds of schemas exist for different types of resources, professions, institutions, cultures, and communities because it is difficult to find one plan or one schema to satisfy all needs.

To summarize, schemas are sets of elements and other vocabulary that have definitions and rules for how to use them. Schemas are a type of standard that can provide consistency for metadata fields names and rules. Like other standards, schemas have the power to codify cultural biases, so it is important to continue to examine those biases and their effects.

Section 1 Quiz Questions

According to what you've just read, which group of three things listed does a schema provide:

a. Semantics, syntax, optionality
b. Metadata elements, rules, a plan
c. All of the above

How does a schema differ from metadata elements that you would come up with?

a. Schemas have rules and definitions for how they are used that must be followed
b. There is no difference
c. Schemas also contain subelements

Section 1 Discussion question

Why is it important to understand the rules and definitions outlined in a schema?

Section 2: Namespaces

This section discusses what namespaces are and how they are used with schemas.

Important vocabulary in this section
Namespace _____ URI _____ Namespace identifier _____ Prefix _____

A namespace is like an identifier for a schema, which is a set of vocabulary representing properties (elements), classes (types of resources), other terms. Namespaces group schema vocabulary together and contextualize terms.

You *could* think of namespaces like family names, except namespaces are unique, whereas family names are not. Still, pretend you live in a small town in which everyone's family name is unique. While there may be multiple people with the same first name, a person's family name identifies a person, distinguishes them from someone else, groups them together with other family members, and provides other contextual information surrounding a person's identity. Which Sonoe stole bread from the bakery? Ishihara Sonoe comes from a strict family, so she wouldn't risk it. Yamamoto Sonoe moved to the next town over…must have been that Nakasone Sonoe — always getting away with bad things! In this example, the family names of the three Sonoes not only distinguishes them from each other, but provides context about who they are based on their family associations.

Namespaces are somewhat similar. Namespaces help to distinguish a "name" element from the Sonoe Chair Schema (SCS), for example, from a "name" element in another schema. This is important because the "name" elements in each schema may have different definitions and specific relationships to other elements in their respective schemas. Namespaces are

not required for every schema, but a schema without a namespace is kind of like a person without a family name.

Namespaces are often represented as *URIs* (Uniform Resource Identifiers), which are also sometimes URLs (Uniform Resource Locators). A URI is an identifier (URI); URLs are URIs that also act as a locator or address on the web. Ideally, URLs are "resolvable," which means when you click on them, they take you to a page on the internet with relevant information. Namespace URIs don't have to be resolvable, but making URIs resolvable is considered good practice to help people use and understand the namespace.

Here are some examples of namespaces that are URLs.

https://purl.org/dc/terms/
https://schema.org

Again, not all namespaces are resolvable URIs, but these examples are. If you click on each URI, they would resolve to a webpage with information relevant to their respective schemas. The part of the namespace listed above is known as the *namespace identifier* — this is the root of the namespace — the "family name."

The full name of any single vocabulary term in a schema would contain the namespace identifier. For example, the Dublin Core Terms "creator" property (metadata field) is represented by http://purl.org/dc/terms/creator; the "about" property from Schema.org would be https://schema.org/about. As with the namespace identifiers, these URIs don't necessarily have to be resolvable.

If everytime you wanted to refer to an element from a schema, you had to include the element name and the namespace identifier, this could become cumbersome. For that reason, namespace identifiers are often referred to using an abbreviation called a *prefix*. To establish a prefix in any situation, you would first come up with a nickname for your namespace. For Sonoe's Chair Schema, it might be "schair." You'd then assign the namespace as the

value of “schair.” In English, you might say, “I’m going to use “schair” to mean http://example.org/sonoeChairSchema/.” In various computer languages, it might look more like a math equation: schair=“http://example.org/sonoeChairSchema/”.

You’d then be able to use “schair” as a shorthand. Instead of writing http://example.org/sonoeChairSchema/name you could simply write “schair:name.”

Prefixes are nicknames and can be used for schemas or other resources. One example of where prefixes are often used is when a metadata description or record is using vocabulary from multiple schemas or vocabulary sets. In that case, the prefix is a quick way to distinguish which schema or vocabulary set each element is from. When I talked about RDF in the previous chapter, I provided examples of how to create an RDF triple using different languages or serializations. In the image below of an RDF triple using the Turtle language, prefixes are used as a shorthand for the schema.org schema as well as for wikidata’s schema. Additionally, a special prefix “bluest” is used as a nickname for *The Bluest Eye* to make it easier to distinguish.

```
Turtle
@prefix bluest: <https://www.wikidata.org/wiki/Q1069956>.
@prefix schema: <http://schema.org/>.
@prefix wikidata: <https://www.wikidata.org/wiki/>.

<bluest> schema:author <wikidata:Q72334>.
```

To summarize, namespaces identify, group together, and contextualize vocabulary from a schema. Namespaces are represented by URIs, which sometimes also act as resolvable URLs. Namespace prefixes shorten namespaces to make them easier to use when referencing vocabulary within a schema.

Section 2 Quiz Questions

What do namespaces provide for a schema?

a) An identifier

b) A way to group together schema from one vocabulary
c) Scoping information that contextualizes the vocabulary
d) All of the above

A prefix is required when working with a namespace
a) True
b) False

A namespace is required when working with a schema
a) True
b) False

Based on what you learned, which part of the following URI represents the namespace identifier: https://purl.org/dc/terms/?
a) terms
b) https://purl.org
c) https://purl.org/dc/terms/

Section 2 Discussion Question

Based on what you've learned about namespaces, why might having a namespace for a schema be better than not having one?

Section 3: Dublin Core Metadata Initiative (DCMI) Metadata Terms

In this section, you will learn about a schema called DCMI Metadata Terms, including how it has evolved, is used, and its role in digital libraries.

Important vocabulary for this section
DCMI Metadata Terms / Dublin Core Terms / DC Terms
_____
Properties (Dublin Core Terms) _____
Elements namespace _____
Terms namespace _____
Scope _____
Range _____
Class _____

Dublin Core Metadata Initiative (DCMI) Metadata Terms—Dublin Core Terms (DC Terms) or Dublin Core (DC) for short—is a schema that supports mostly descriptive metadata. I will refer to the schema as DCMI Terms, DC Terms, Dublin Core, and DC interchangeably for the remainder of the chapter and book. Although its current configuration uses Resource Description Framework (RDF), it is language and content standard agnostic. This means Dublin Core can be used in conjunction with a variety of languages, data structures, and content standards.

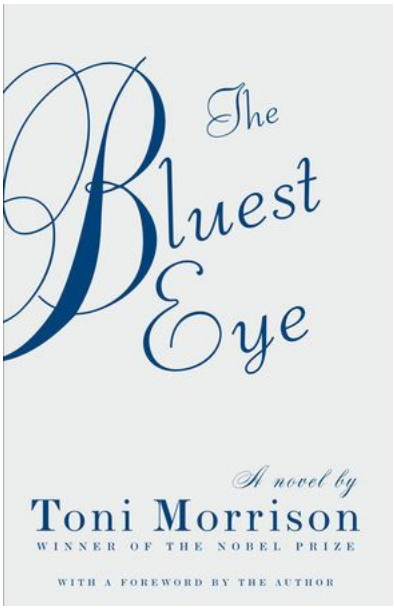
DC Terms consist of multiple subsets of vocabulary. The first set of DC vocabulary was originally developed around 1995 as the original “Dublin Core,” consisting of 15 elements called *properties* (metadata fields or categories). The original 15 properties are **contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, type**. These 15 fields are currently defined as properties in the DC *elements namespace* (http://purl.org/dc/elements/1.1).

As a schema, DC provides definitions and rules for how each of the properties and other terms within the schema are used. It is important to carefully read the definitions and rules before using the vocabulary in DC. The documentation for these definitions and rules is found here: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
The "Index of Terms" section on this website is where you'll find all vocabulary in DC. The terms are organized into a table (recreated below). The original 15 properties are listed in the second row of the table: "Properties in the /elements/1.1/ namespace." Click on one of the original 15 properties to see the definitions and instructions for how to use the property.

Index of Terms

Vocabulary Encoding Schemes:	DCMIType, DDC, IMT, LCC, LCSH, MESH, NLM, TGN, UDC
Syntax Encoding Schemes:	Box, ISO3166, ISO639-2, ISO639-3, Period, Point, RFC1766, RFC3066, RFC4646, RFC5646, URI, W3CDTF
Classes:	Agent, AgentClass, BibliographicResource, FileFormat, Frequency, Jurisdiction, LicenseDocument, LinguisticSystem, Location, LocationPeriodOrJurisdiction, MediaType, MediaTypeOrExtent, MethodOfAccrual, MethodOfInstruction, PeriodOfTime, PhysicalMedium, PhysicalResource, Policy, ProvenanceStatement, RightsStatement, SizeOrDuration, Standard
DCMI Type Vocabulary:	Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage, Text
Terms for vocabulary description:	domainIncludes, memberOf, rangeIncludes, VocabularyEncodingScheme

Recreation of "Index of Terms" table from the DCMI Metadata Terms documentation page, retrieved September 2020 from https://www.dublincore.org/specifications/dublin-core/dcmi-terms/



Look at this image (copyright Knopf) of the cover for *The Bluest Eye*. Using some of the original 15 DC Terms, let's create metadata for this resource.

Which property might you use to capture the following metadata: "The Bluest Eye"?

a) Date
b) Title
c) Relation

Which property might you use to capture the following metadata: “Toni Morrison”?

a) Creator
b) Publisher
c) Subject

Based on the information on the cover, what metadata would you record in the *language* property?

a) English
b) French
c) Spanish

The following DC metadata uses three of the original 15 Dublin Core properties to create a metadata record for *The Bluest Eye* based on the information available on the cover:

Title:	The Bluest Eye
Creator:	Toni Morrison
Language:	English

* * *

Another set of DC vocabulary is defined in the *terms namespace* (http://purl.org/dc/terms/). The terms namespaces contains a repeat of all the original 15 properties, plus another 40 other metadata fields that were added since 1995. But why include the original 15 DC properties in two separate namespaces?

Although the original 15 properties are captured in both the *elements* and *terms* namespaces, the properties are sometimes *scoped* differently in each namespace. For example, the “subject” properties in both namespaces have the same definition (“a topic of the resource.”), but only the *terms* namespace specifies in the comments that a URI is recommended. This level of specificity was not available under the earlier *elements* namespace. Some

properties in the *terms* namespace add further specificity and require a *range*—a limit to the type of information that is allowed in a field. A range might specify a data type or category of value.

Term Name: creator	
URI	http://purl.org/dc/terms/creator
Label	Creator
Definition	An entity responsible for making the resource.
Comment	Recommended practice is to identify the creator with a URI. creator may be provided.
Type of Term	Property
Range Includes	• http://purl.org/dc/terms/Agent

Term Name: Agent	
URI	http://purl.org/dc/terms/Agent
Label	Agent
Definition	A resource that acts or has the power to act.
Type of Term	Class
Instance Of	http://purl.org/dc/terms/AgentClass

Two screenshots of DCMI Metadata Terms documentation. Left: The "creator" property in the terms namespace requires a range of "agent." Right: "Agent" is defined as a class of resource that acts or has the power to act.

DC contains additional subsets of vocabulary not used as metadata fields: classes, vocabulary encoding schemes, syntax encoding schemes, type vocabulary, and terms for vocabulary description. You can see these vocabulary sets in the recreated Index of Terms table shown earlier. These terms add specificity and context to how metadata fields (a.k.a. properties) are used. For example, one of these subsets, *class*, provides vocabulary for types of resources or concepts used within a metadata field. Some Dublin Core Terms properties, such as *creator* require that only the class "agents" can be recorded inside the field. As a class, "agent" is a type of resource or concept. Specifically, an "agent" is a resource that can perform actions, for example, a person is a resource that can perform the action of creating.

Because of its relative simplicity, DC has become widely adopted as a digital library standard. Many Content Management Systems (CMS), most notably ContentDM, use Dublin Core as the basis for metadata creation. Its broad and generic terms (e.g., "creator" rather than "author," "photographer," "artist," etc.) have led to Dublin Core being used to support interoperability. The OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), for

example, uses the original 15 DC properties to help digital libraries share metadata with others.

Despite the original goal for DC to be used for any web-based resource, it still shows a strong bias for text-based resources and to some extent published resources.[7] Further, as a standard created by English speakers from the United States, it is important to examine how cultural biases that prioritize written over oral documentation and published over unpublished resources have shaped DC Terms. Many communities have created their own schemas based on DC Terms to account for the schema's limitations that are more suited to their specific needs. Examples include Darwin Core for natural history materials, PBCore for AV materials, and DC Educational for educational materials.

To summarize, DC, officially known as DCMI Metadata Terms or Dublin Core Terms for short, is a schema containing broad, generalizable metadata fields that support mostly descriptive metadata. DCMI Metadata Terms features vocabulary from multiple namespaces, most notably the element and terms namespace, and is RDF compliant.

Section 3 Quiz Questions

What is the recommended way to learn about the definitions and scope of vocabulary in DC Terms?

a) Read the entire documentation website end to end
b) They are self explanatory, so select the term that seems closest to what you are describing
c) Go to the Index of Terms documentation page and click on the vocabulary to learn more

[7] The original scope of Dublin Core was Document-like objects (DLOs) consisting primarily of text. Weibel, Stuart L. 1995. "Metadata: The Foundations of Resource Description." D-Lib Magazine, 1,1 (July). Available online at: http://www.dlib.org/dlib/July95/07weibel.html.

If you are creating metadata for a resource, which of the following sets of DC vocabulary provide metadata fields for you to use?

a) Terms for Vocabulary Description
b) Properties
c) Classes
d) DCMI Type Vocabulary

Can you use “Event” as the name of a metadata field when using DC?

a) Yes
b) No

Section 3 Discussion Question

Why do you think the original 15 elements are found in two different namespaces? Why do you think a new namespace was created rather than updating the original namespace?

Section 4: Schemas and XML

This section will discuss how schemas can be implemented with XML and how that is distinct from XML schemas.

Important vocabulary in this section

XML schema _____
XSD _____

As we learned in the last chapter, XML has no predefined elements. We can make up any elements and attributes needed to encode and store data as long as it conforms to the rules of well-formed XML. While this flexibility of XML lowers the barrier to creating a metadata record in XML, it increases the possibility of inconsistency when different people or institutions invent their own XML tags. For example, one institution might use the following elements to describe a chair:

```
<record>
        <chair>Eames Lounge and Ottoman</chair>
        <maker>Herman Miller</maker>
        <invented>1956</invented>
</record>
```

Whereas another institution might use these other elements:

```
<chair>
        <title>Womb chair</title>
        <artist>Eero Saarinen</artist>
        <year>1946</year>
</chair>
```

A third institution may use another set of elements, and so on.

What's the solution? A schema! If schemas are sets of elements and other vocabulary, and XML is a language with no predefined elements and other vocabulary, then these two can be used together to create standardized XML metadata. Now see what happens when I catalog the same two records using the Sonoe's Chair Schema (SCS) from Section One.

Institution 1

```
<chair xmlns:schair="http://fake.org/sonoeChairSchema">
        <schair:name>Eames Lounge and Ottoman</schair:name>
        <schair:manufacturer>Herman Miller</schair:manufacturer>
        <schair:firstYearAvailable>1956</schair:firstYearAvailable>
</chair>
```

Institution 2

```
<chair xmlns:schair="http://fake.org/sonoeChairSchema">
        <schair:name>Womb Chair</schair:name>
        <schair:creators>
                <schair:manufacturer>Knoll</schair:manufacturer>
        </schair:creators>
        <schair:firstYearAvailable>1946</schair:firstYearAvailable>
</chair>
```

The records now have standardized elements. You may notice a couple things in this example. One, the namespace for SCS is included. Two, the records still look a little different from each other. Institution two has an element called "creators" and a subelement called "manufacturer." Why? Although the schema helps to define elements and tells you what kind of data you can put in them, it doesn't tell you how to structure them in a specific language like XML. For example, the schema doesn't tell you whether or not

to make “manufacturer” an XML subelement. For that kind of information, you will need one more thing: an XML schema.

You don’t have to use an XML schema in order to create XML metadata using a schema. The XML examples above use SCS elements to create an XML record, but do not use an XML schema. I created SCS as a schema that can be used with any language, not just XML, so my schema doesn’t have specific rules about how to create an XML record. If we wanted to create rules on how to implement Sonoe’s Chair Schema in XML, we would need to outline those rules in an *XML Schema Document (XSD)*.

XSD is a special language that defines XML elements, attributes, and other vocabulary and outlines how they should be ordered, structured, and formatted in XML. XSD are created to provide further standardization for how elements in a namespace schema are implemented. In an XML schema, there are two sets of rules to remember: the rules of well-formed XML and the rules of the XSD.

One real world example is the implementation of Dublin Core in XML. You can create Dublin Core metadata on a piece of paper, in a spreadsheet, database, XML, JSON—whatever. Dublin Core had a long history, however, of being implemented in XML as a publishing mechanism. So, how do we implement Dublin Core in XML? The answer varies. Dublin Core can be implemented in XML with or without the use of a supporting XSD, the way we created XML for Sonoe’s Chair Schema without the use of an XSD. Best practice, however, would suggest using an XSD to guide the creation of Dublin Core in XML. The DCMI suggests a few different XSD on their website: https://www.dublincore.org/schemas/xmls/. The Open Archives Initiative (OAI) has their own XSD as well: http://www.openarchives.org/OAI/2.0/oai_dc.xsd. Additionally, some institutions or Content Management Systems (CMSs) may create their own XSD in order to enforce consistent coding of Dublin Core in XML and to enable records to be automatically validated.

To summarize schemas and XML can be used together to create more consistent, standardized, and interoperable XML metadata. When XML is used with a schema you must adhere to the rules of well-formed XML as well as the definitions and rules of the schema. Furthermore, XML Schema Description (XSD) can provide additional guidance on how to implement a schema specifically using XML.

Section 4 Quiz Questions

If you want to create an XML record for a schema, you have to use XSD.

a) True
b) False

XML Schemas use XSD to specify rules for how to implement a schema specifically in XML.

a) True
b) False

Dublin Core is an XML schema

a) True
b) False

Section 4 Discussion Question

What are the advantages and challenges of following an XSD when creating metadata in XML?

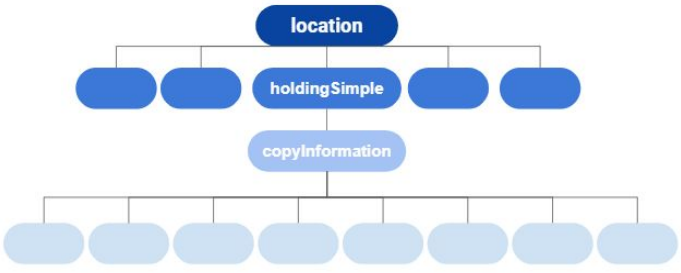
Section 5: Metadata Object Description Schema (MODS)

This section will provide an overview of how Metadata Object Description Schema (MODS) was developed and how it is used.

Important vocabulary in this section
MODS _____
Container element _____

MODS is a schema developed in 2002 by the Library of Congress to record mostly descriptive metadata for resources in XML. MODS is based on a subset of fields from MARC21 (MAchine Readable Cataloging), but uses human readable XML elements rather than numerical fields. As a result, metadata creators can make more complex and detailed descriptions using MODS than is possible using Dublin Core.

As an XML schema, MODS provides definitions and rules for a set of elements, subelements, attributes, and other vocabulary, and also specific instructions on how to use these vocabulary to create a metadata record in XML. XML uses elements, subelements, and attributes to tag information. In MODS, you will see the use of each of these three concepts.



MODS is highly hierarchical. There are **20 top level elements**, about 11 of which have sub-elements that in turn have their own sub-elements. To share an extreme example of this hierarchy, a top level element called “location” has a subelement called “holdingSimple,” which has a subelement called “copyInformation,” which has eight subelements.

In MODS, elements that have sub-elements are sometimes called *container elements*. Container elements do not store text directly. Instead, they store subelements. This type of nesting is a common feature of XML.

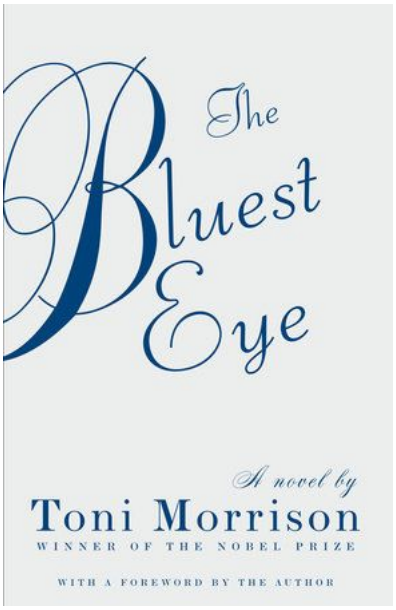
The purpose for hierarchy in MODS is to keep multiple pieces of metadata about the same characteristic of a resource together. For example, if I wanted to record metadata about the author of *The Bluest Eye*, I would use the MODS top level element “name” to store Toni Morrison’s given name, family name, and role as creator in separate subelements. In MODS, “name” is a container element. Storing Toni Morrison’s given name, family name, and role in separate sub-elements makes it easier to sort, filter, and search by these fields separately, but because all three of these subelements are contained within the “name” element, the relationship between “Toni” + “Morrison” + “creator” is preserved. Below is an example of what this metadata might look like in MODS.

```
<name type=“personal”>
        <namePart type=“given”>Toni</namePart>
        <namePart type=“family”>Morrison</namePart>
        <role>
                <roleTerm>creator</roleTerm>
        </role>
</name>
```

Notice that the MODS above uses attributes. XML uses attributes to further refine how a metadata field is used. For example, MODS provides the attribute “type” for the “namePart” element so that you can tag name parts such as given name, family name, life / death dates, or an honorific. MODS does not require that you use the type attribute, but it helps distinguish between multiple “namePart” elements. MODS also does not require you to separate a name into multiple parts, but it is an option for those who want this level of granular control over the metadata. Finally, notice also that the “type” attribute has a different purpose for the “name” element than it does for the

“namePart” element. These distinctions and rules are painstakingly outlined within MODS’s XSD.

MODS documentation features information about elements, subelements, attributes, and accepted values for each of these. Each updated version of MODS is posted to the schema’s website Home Page (https://www.loc.gov/standards/mods/), and each updated user guide is posted here: https://www.loc.gov/standards/mods/mods-guidance.html. The “Top-Level Elements in MODS” provides definitions and rules for MODS concepts and links to subelement and attribute information.



As we did with Dublin Core Terms, look at this image of the cover for *The Bluest Eye*. Using some MODS elements, subelements, and attributes, let’s create metadata for this resource.

Which top level element and subelement best capture the following metadata: “The Bluest Eye”?

d) <titleInfo> and <subTitle>
e) <titleInfo> and <title>
f) <name and description>

Which MODS elements best capture the following metadata: Toni Morrison?

d) <name> and <namePart>
e) <firstName> and <lastName>
f) <name> and <role>

Which attribute and attribute value specifies that “Toni Morrison” is the name of a person?

a) type=“personal”
b) type=“family”
c) type=“corporate”

The following is a MODS metadata record for The Bluest Eye based on the information available on the cover and using elements discussed above:

```
<mods>
        <titleInfo>
                <title>The Bluest Eye</title>
        </titleInfo>
        <name type=“personal”>
                <namePart>Toni Morrison</namePart>
        </name>
</mods>
```

Notice that <mods> is used as a root element for the record. MODS allows one of two root elements: <mods> or <modsCollection>. <mods> is used when creating a single record in a single document; <modsCollection> is used to collect multiple MODS records together into a single document or collection. <modsCollection> might also be useful for a resource that has many distinct parts that need their own metadata, for example, creating MODS descriptions for each photo in a photo album.

MODS is still widely used in digital libraries, but its complexity and the profession’s move towards linked data have prompted communities such as the “Samvera MODS and RDF Descriptive Metadata Subgroup” to translate MODS concepts to RDF. As you learned in Chapter 4 on data structures, Resource Description Framework (RDF) uses the subject-predicate-object structure to describe resources. In RDF, the same MODS metadata above might be expressed using the triples “This resource has the name Toni Morrison,” “This resource has the title The Bluest Eye,” and “Toni Morrison has the name type personal.” In these examples, “name,” title,” and “name type” still need to be defined somewhere; even in RDF, MODS can provide definitions and rules for metadata elements or properties.

As a schema based on MARC, MODS shows a bias towards published library materials with top level elements such as titleInfo, tableOfContents, and classification, which may not be as relevant or appropriate for other cultural heritage resources, for example, museum objects or artifacts. Further, as an XML schema, the use of hierarchy in MODS makes clear value statements about which types of information are most important or dominant.

In summary, Metadata Object Description Schema (MODS) is a highly hierarchical XML metadata schema based largely on MARC. It contains 20 top level elements, many of which have sub-elements that have sub-elements of their own. Although MODS is widely used in digital libraries and repositories, there is a current trend towards MODS elements being used to support structured metadata using RDF.

Section 5 Quiz Questions

If <titleInfo> is a container element, which of the following statements is not a correct use of the element.

a) <titleInfo><title>The Bluest Eye</title></titleInfo>
b) <titleInfo>Star Wars</titleInfo>
c) <titleInfo><subtitle>Episode1</subtitle></titleInfo>

In MODS, what does “top level” element mean?

a) An element that is at the top level of the hierarchy, beneath the root element.
b) An element that is the most important element of the entire schema.
c) The root element of MODS.

Look at the MODS documentation and the first example of MODS provided in this section. Which values are allowed for the type attribute when it is used with the <name> element?

a) Personal
b) Family

c) Creator

Section 5 Discussion Question

Looking at MODS's hierarchical structure, can you find examples of metadata sub-elements that you think should be their own top level elements?

Chapter 5 Recommended Readings

Dublin Core Documentation:
https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

MODS (version 3.7) Documentation:
https://www.loc.gov/standards/mods/mods-outline-3-7.html

W3School.com XML Schema Tutorial:
https://www.w3schools.com/xml/schema_intro.asp

Chapter 5 References

DCMI Usage Board. DCMI Metadata Terms. Retrieved December 2020 from https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

ISO 23081-1:2017(en), Information and documentation — Records management processes — Metadata for records (2017). Retrieved December 2020 from https://www.iso.org/obp/ui/fr/#iso:std:iso:23081:-1:ed-2:v1:en.

Library of Congress. Outline of Elements and Attributes in MODS Version 3.7. Outline of elements and attributes in MODS version 3.7: MetadataObject Description Schema: MODS (Library of Congress). Retrieved December 2020 from https://www.loc.gov/standards/mods/mods-outline-3-7.html.

Library of Congress. *MODS: Uses and Features*. MODS: Uses and Features (Metadata Object Description Schema: MODS). Retrieved December 2020 from https://www.loc.gov/standards/mods/mods-overview.html.

Morrison, T. (2007). Book cover from The Bluest Eye. United States: Knopf Doubleday Publishing Group. https://www.google.com/books/edition/The_Bluest_Eye/12_KUGLXigMC?hl=en&gbpv=0 (accessed November 2020).

Weibel, Stuart L. 1995. "Metadata: The Foundations of Resource Description." D-Lib Magazine, 1,1 (July). Retrieved December 2020 from http://www.dlib.org/dlib/July95/07weibel.html.

Chapter 6: Crosswalking

The work of metadata is not only about creating metadata about resources, but also managing metadata within, across, or between systems. In this chapter, you will learn about crosswalking, which is one important technique for managing metadata from multiple or unfamiliar origins.

Section 1: Introduction to Crosswalking

This section will introduce you to the concept of crosswalking and crosswalks as they pertain to metadata.

Important vocabulary in this section
Crosswalk (noun) _____ Crosswalk (verb) _____ Mapping _____

A *crosswalk* is "a *mapping* of the elements, semantics (meaning), and syntax (grammar) from one metadata schema to those of another" (Hodges, 2001). Sometimes libraries need to share metadata with other libraries, for example, because they share a catalog. Not every library will use the same metadata standards when creating metadata, so crosswalks that translate metadata standards are necessary for bringing together metadata from different systems. Because one goal of crosswalking is to assimilate metadata from multiple schemas, crosswalking also involves an expression of power.

You can think of a crosswalk like a translation dictionary for schemas. Just as a translation dictionary lists vocabulary and grammatically correct phrases in one language and translates them into another language, a crosswalk is a tool that lists all the vocabulary (elements, subelements, attributes, etc.) and grammar from one schema and translates those to another schema.

To illustrate this comparison between a crosswalk and a dictionary, look at the following English to Brazilian Portuguese dictionary (disclaimer: I do not speak Brazialian Portuguese).

English → Portuguese Dictionary

English	Portuguese

Good Morning	Bom Dia
Family	Família

What is the Portugues equivalent to “good morning”? “Bom dia.” What is Portuguese equivalent to “family”? “Família.”

Now let’s look at the following MARC to Dublin Core Crosswalk:

MARC → DCMI Metadata Terms (Dublin Core) Crosswalk:

MARC	Dublin Core
520	abstract
100	creator

What is the Dublin Core equivalent to MARC 520? “Abstract.” What is the Dublin Core equivalent to Marc 100? “Creator.”

Although it’s a subtle difference, the direction of translation can matter for a crosswalk just as it can for a translation dictionary.[8] Below are two dictionaries: one is an English to Brazilian Portuguese Dictionary, the other is a Brazilian Portuguese to English dictionary. Notice some differences?

English → Portuguese Dictionary

English	Portuguese
Children	Crianças (people under 18) Filhos (children of a parent)
Family	Família

[8] See “Metadata Interoperability and Standardization – A Study of Methodology Part I: Achieving Interoperability at the Schema Level” section 4.4 “Switching-across” by Lois Mai Chan and Marcia Lei Zeng for more on the directionality of crosswalking. D-Lib Magazine, Volume 12 Number 6, June 2006.

Português → Inglês Dicionário

Português	Inglês
Crianças	Children
Família	Family

Look at the English to Portuguese Dictionary. If you are an English speaker and you want to know how to ask your Brazilian colleague about their children, who are in their 20s, which of the two Portuguese words for "children" would you use? "Filhos." Although either word could be used if this person had young children, as adults, your colleagues' children are better described using the word "filhos."

But what if you had tried to use the second dictionary—the Portuguese to English Dictionary? As you can see, going from English to Portuguese isn't quite the same as going from Portuguese to English? If you used the second dictionary to ask your colleague about their adult children, it would come out a bit funny using the word "crianças," as that word is used to refer to people under 18.

Similarly, the direction of a crosswalk can make a difference. A MARC to Dublin Core crosswalk might not give you the same kind of information as a Dublin Core to MARC crosswalk. Below is an example.

DCMI Metadata Terms (Dublin Core) → MARC Crosswalk:

DCMI Metadata Terms	MARC
abstract	520, indicator 3
creator	100 (individual people) 110 (corporations) 111 (meetings)

MARC → DCMI Metadata Terms (Dublin Core) Crosswalk:

MARC	DCMI Metadata Terms
520	abstract
100	creator

As you can see, both these crosswalks translate the schemas MARC and Dublin Core, but the information provided depends on which direction the schemas are translated. Let's say you have a Dublin Core creator that is a corporation and you want to find out what the MARC equivalent of "creator" is. You would want to use the Dublin Core to MARC crosswalk because you're starting with what you know or have (Dublin Core) and translating it to another schema. Using that crosswalk, you can see that the translation from Dublin Core to MARC depends on if the creator is a person, corporation, or a meeting. Since you have a corporation as the creator, you will know to use MARC 110.

Were you to use the other crosswalk, the MARC to Dublin Core Crosswalk, there would be crucial information missing for you to correctly translate Dublin Core creator to MARC 110.

As I mentioned earlier, crosswalks can be an expression of power. Earlier, my disclaimer noted that I don't speak Brazilian Portuguese. How would you know my translations are correct? Am I missing shades of meaning in the language that are culturally significant? Once published, a crosswalk presents an interpretation of equivalencies between schemas that can seem as authoritative as a translation dictionary published by Webster. It's important, however, to consider who is doing the mapping and what their relationship to each schema is when creating or using a crosswalk.

To summarize, a crosswalk is like a translation dictionary. It translates the vocabulary of one schema (that vocabulary includes the elements, fields, categories, attributes, terms, etc.) to that of another schema. Like translation dictionaries, crosswalks can be directional, so you'll want to make sure when

using a crosswalk or building a crosswalk that it is clear which schema you are starting from and which schema you are translating to.

Section 1 Quiz Questions

Which of the following statements is most accurate based on this reading?

a) Crosswalks map or match elements from one schema to another based on definition and language
b) Crosswalks map or match elements from one schema to another based on definition and rules of use
c) Crosswalks map or match elements from one schema to another based on the name and rules of use

Based on the crosswalk example in this reading, what is the mapping (translation) for MARC 520 in Dublin Core?

a) Creator
b) Creator (corporations)
c) Abstract
d) 520, indicator 3

Section 1 Discussion Question

What are some examples or possible reasons for needing to create or use a crosswalk?

Section 2: Challenges of Crosswalking

Sometimes the full meaning of a word can get lost in translation between two languages. There are similar challenges to crosswalking or mapping from one schema to another. Although a crosswalk attempts to find equivalent values between two schemas, it isn't always possible to create a perfect translation. This section discusses some of the challenges of crosswalking or mapping concepts from one schema to another and considerations for meeting those challenges.

Important vocabulary in this section
One to one relationships _____ Associativity _____

I will discuss four challenges of crosswalking. Although this is not a comprehensive list of challenges, as might be found in the seminal NISO article "Issues in Crosswalking Content Metadata Standards," (St. Pierre and LaPlant, 1998) the four challenges below illustrate common barriers to crosswalking.

Challenge 1: Elements may be structured differently, resulting in a lack of *one to one relationships*.

In the following example, we'll look at Dublin Core and MARC metadata.

in a Dublin Core record:	in a MARC record:
Publisher: Knopf Date: 1950	264 1 |aLondon: |b Knopf, |c 1950

In the example above, information is structured differently in each record. The highlighted portions of the record are the metadata fields or elements in each

schema: Dublin Core “publisher” and “date” on the left and MARC “264” on the right. In this example, there isn’t a one to one relationship between Dublin Core and MARC: there isn’t one field in Dublin Core and one field in MARC that contain the exact same kind of information for publication information. What takes two fields in Dublin Core takes only one field in MARC.

The 264 MARC field also uses additional symbols (delimiters) to separate some of the information in the field into subfields or subelements. The publisher information is recorded in the 264 subfield |b and the date information is recorded in the 264 subfield |c. When creating a crosswalk, referring to such subfields provides specificity to the translation. A Dublin Core to MODS crosswalk might record this specificity as follows:

DCMI Metadata Terms	MARC
publisher	264 |b
date	264 |c

Challenge 2: Information loss: some concepts don’t exist in other schema

The information recorded in the 264 subfield |a has no equivalent value in Dublin Core. Although there is a Dublin Core element that records information about locations (“coverage”) that field is used for locations that are topics or subjects of the resource, not for the location of publication. There is no field in Dublin Core that has the same meaning or is used for the same purpose as MARC 264 |a. As a result, the place of publication information cannot be recorded in a similar field in Dublin Core and may be lost.

Challenge #3: Loss of *Associativity*

The loss of associativity is a specific kind of information loss. Some schemas associate information together using links or hierarchy. For example, in the MARC record shown earlier, all the information in 264 is grouped together.

The “1” in the 264 field has a specific meaning: “publication information.” This means everything in the 264 field that follows—the location, the company name, the date—all of that is related to publication. These pieces of data are also closely associated with each because they are grouped together under the umbrella of the 264 MARC field. Dublin Core, however, does not treat publication information the same way. There is no specific date for publication, so a publication date translated into Dublin Core loses its direct and explicit association with the publisher information. In that sense, there is a subtle loss of meaning because the direct association between two pieces of data cannot be translated.

Challenge #4: Data specificity

Factors causing data specificity challenges include, but are not limited to, rules about how to format data (e.g., one schema might require dates formatted yyyymmdd whereas another might require ddmmyyyy) and issues of semantics. The word semantics means “meaning.” The semantics of an element refers to the meaning or definition of it. The following records illustrate an example of how semantics can affect data specificity.

in a Dublin Core record:	in a MARC record:
Coverage: Durham (N.C.).	651 0 |aDurham (N.C.).
Coverage: Twentieth century.	650 0 |aTwentieth century.

In Dublin Core, the semantics of the “coverage” field—that is the meaning and definition of this field—allows for multiple kinds of information to be stored. The definition of “coverage” is “The spatial or temporal topic of the resource, spatial applicability of the resource, or jurisdiction under which the resource is relevant.” The field has two meanings: a geographic place name *and* a time period or era of time, both of which must be subjects or topics related to the resource. Additional guidance in Dublin Core says you can store a third kind of information here: geographic coordinates. Because this field can contain many different kinds of information, the crosswalk from DC “coverage” to a field in MARC must note these nuances and differences of meaning.

Depending on how a person decides to use “coverage” it could have completely different mappings in MARC.

Although we’ve focused on Dublin Core and Marc examples, these issues can occur when crosswalking between any other schemas depending on their characteristics.

To summarize, there are many challenges to crosswalking, but I’ve shared four common ones. Depending on the schemas involved and how similar they are, the challenges may make it difficult to create a perfect crosswalk. Knowing the challenges, however, allows you to create a crosswalk that provides as much detail as possible to guide someone through these differences between the schemas.

Section 2 Quiz Questions

Which of the following is not a challenge mentioned in this section?

a) Loss of associativity
b) Elements not called the same thing
c) Elements are structured differently
d) Information loss

In the example of the data specificity challenge, why is it difficult to map the Dublin Core “Coverage” field to a field in MARC?

a) “Coverage” is formatted differently than the equivalent field in MARC.
b) “Coverage” has a definition that doesn’t exist in any MARC field, so there’s no way to map the fields.
c) “Coverage” can contain different types of data, so it could map to many different MARC fields depending on the type of data.

Which option below best explains the problem in the example outlined for the loss of associativity challenge?

a) One-to-one
b) Hierarchy

c) Formatting

Section 2 Discussion question

Challenge 1 discusses elements structured differently resulting in a lack of one-to-one-relationships. If there is sometimes a lack of one-to-one relationships, then how might you describe the other relationships that might result?

Section 3: How Are Crosswalks Created?

This section will discuss methods for creating a crosswalk and factors to consider when crosswalking.

Important vocabulary in this section
Absolute crosswalking _____
Relative crosswalking _____

Crosswalking can be a complex process due to several factors noted by St. Pierre and LaPlant (1998) that one must consider when crosswalking:

- ❏ the definition of each metadata element
- ❏ whether or not a metadata element is mandatory, optional, etc.
- ❏ whether or not a metadata element can be repeated
- ❏ organization of elements relative to each other, e.g., hierarchical
- ❏ the required value or data types for the element (e.g., free text, numeric range, date, or a controlled vocabulary)
- ❏ if locally-defined metadata elements allowed

To summarize the list above: one must consider mappings based on the definition and rules of use for each element. These factors create pretty rigid criteria that might result in the failure to map some elements between schemas that are extremely different. When creating a crosswalk, there is a choice or perhaps a balance between accuracy and minimizing the loss of data. This tension is reflected in the two approaches to crosswalking that Lois Mai Chan and Marcia Lei Zeng identified: *absolute-crosswalking* and *relative-crosswalking* (2006).

Absolute-crosswalking

Absolute-crosswalking involves exact or closely-equivalent mapping between elements, otherwise, elements are not crosswalked (Chan and Zeng, 2006). This method prioritizes accuracy (i.e., equivalency in both definitions and rules

of use) over retaining information. To illustrate an example of absolute crosswalking, look at the following MARC field:

264 1 |aLondon: |b Knopf, |c 1950

The 264 MARC field records information about the production, publication, distribution, manufacturing, or copyright notice of a resource; only one of these categories is allowed for each 264 field. In this example, the “1” following the “264” indicates that this 264 field is about the publication of a resource. This field also contains three subfields: subfield |a records the place of publication; subfield |b records the name of the publisher; subfield |c records the year of publication.

In an absolute crosswalk from MARC to Dublin Core, we would look for fields in Dublin Core that have the same or similar definition and rules of use as the fields in MARC. One example is the equivalency between the MARC 264 subfield |b and Dublin Core “publisher,” since the subfield |b above contains the name of the resource’s publisher. Another example is the MARC 264 subfield |c, date of publication. The MARC subfield is mapped to Dublin Core “issued” because the definition for “issued” (date of formal issuance of the resource) appropriately describes what a publication date is.

MARC	DCMI Metadata Terms
264 |a	n/a
264 |b, with 2nd indicator “1”	publisher
264 |c, with 2nd indicator “1”	issued

Unfortunately, the MARC 264 subfield |a does not mapped to any Dublin Core field in an absolute crosswalk because Dublin Core does not have a field that stores information about the place a resource was published. This decision values the accuracy of mapping (i.e., not violating the definitions and rules of

any Dublin Core field) over the loss of information stored inside the MARC 264 subfield |a.

Before moving on, note the precision of the mappings in the example above. Each MARC subfield is mapped, and because the 264 MARC field also refers to production, publication, distribution, manufacturing, or copyright notice, I specified that the MARC mappings require the “1” to indicate publication to be correct. Were this a full MARC to Dublin Core crosswalk, the crosswalk would note whether or not MARC 264 fields that contain production, distribution, manufacturing, or copyright notices have Dublin Core mappings as well.

Relative Crosswalking

Relative crosswalking maps all elements from the starting schema to at least one element in the schema mapped to, even if the definitions of the fields are not the same (Chan and Zeng, 2006). For example, imagine a schema called CLOWNS has a field called “clownsAvailable” that tells you how many clowns are available for hire.

clownsAvailable: 5

You are asked to crosswalk CLOWNS to MARC, but MARC has no fields similar to “clownsAvailable.” In a relative crosswalk, you might decide to map “clownsAvailable” to MARC 500, which is a general note field. The meaning of these two fields is different, but the mapping preserves the information from the original schema without violating rules for how to use the MARC note field, which allows almost any type of information. Retaining information was prioritized over the accuracy of the mapping, but not without consequence: there is a loss of meaning because the metadata field provided context for what “5” means.

Returning to the MARC 264 field example above, how can we use relative crosswalking to map the subfield |a to a Dublin Core element? Library of Congress’s MARC to MODS crosswalk might map both MARC 264 subfields

|a and |b to Dublin Core “publisher”.[9] Dublin Core “publisher” only stores information about entities (e.g., a person or organization) not locations. The Library of Congress seems to be taking a relative crosswalking approach in this mapping, fudging definitions to put the place of publication along with the publisher in Dublin Core “publisher.” This method prioritizes keeping the information from the MARC subfield, but not without consequence: if carried out, this mapping could affect searching and sorting.

* * *

So which method to use? It will depend on the situation and whether information loss or using metadata fields inaccurately is a bigger concern for your project. In either case, including ample detail to create an accurate and useful mapping is important. For example, in the CLOWNS to MARC crosswalk, specifying that the name of the field as well as the metadata it contains should be recorded in the new schema helps to provide a more useful translation of the metadata.

CLOWNS	MARC	Notes and examples
clownsAvailable	500	Include the original field name in the note along with the data, for example, “clownsAvailable: 5”

Similarly, a note about reordering the MARC 264 subfields |a and |b ensures that sorting by publisher, for example, will not be affected by this questionable mapping to the Dublin Core “publisher” field.

MARC	DCMI Metadata Terms	Notes and examples
264 |a |b, with 2nd indicator “1”	publisher	Format with |b before |a, for example, “Knopf, London”

[9] This 2008 Library of Congress MARC to MODS crosswalk was created before the implementation of 264 and is based on “simple” Dublin Core, so my statement is based on the LC mapping of the older MARC 260 field and the original 15 Dublin Core fields: https://www.loc.gov/marc/marc2dc.html

To summarize, it is important to consider both the definition of the element and all the rules of how the element is used when determining if there is an exact translation in a crosswalk. Absolute crosswalking prioritizes these exact translations and does not map elements without equivalencies. Relative crosswalking prioritizes retention of information and ensures every element in the starting schema is mapped. In either case, it's important to provide as much detail as necessary to create an accurate and useful crosswalk.

Section 3 Quiz Questions

Which method for crosswalking prioritizes retention of information?

a) Relative
b) Exact
c) Absolute

What are some potential consequences of relative crosswalking discussed in this section?

a) Complete loss of information stored within a field that was not mapped.
b) Distrust between the people using the two schemas
c) Loss of context, or data issues such as difficulty sorting and searching.

Which of the following best addresses best practices suggested for creating an accurate and useful crosswalk?

a) Exact matching, even at risk of information loss.
b) Mapping that avoids information loss at all costs.
c) Detail, precision, and adding notes and examples if needed.

Section 3 Discussion Question

Imagine a situation in which you would need to create a crosswalk and describe the approach you would take to create a crosswalk. Why did you choose this approach?

Section 4: How Crosswalks are Used / What are Crosswalks Good For?

This section will talk about how crosswalks are used in libraries.

Librarians use crosswalks as reference documents when they want to understand the differences and similarities between schemas or transform metadata that uses one schema into metadata using a different schema. Crosswalks are often presented as a table in order to easily show comparisons between two or more schemas. These tables can be presented on a piece of paper, in a Microsoft Word file, Google Doc, PDF file, spreadsheet, on a webpage, and more. Below are some examples of crosswalks for popular schemas used in libraries.

Sixteen standards crosswalk:
https://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html

Marc to Dublin Core Crosswalk:
http://www.loc.gov/marc/marc2dc.html

Marc to MODS crosswalk:
https://www.loc.gov/standards/mods/mods-mapping.html

There are many potential uses for a crosswalk, but the most common use is to guide one in transforming metadata from one schema to another schema in a uniform way. Although you may not need a crosswalk to help you translate a MARC record into a Dublin Core record, for example, crosswalking on the fly is time consuming and prone to inconsistencies. Crosswalking requires time to read definitions and rules for each element of each schema. Furthermore, a person must use their judgement about which elements are equivalent; without a crosswalk to document mappings, it is difficult to make decisions and judgements consistently.

Let's look at an example. Below is an abbreviated Dublin Core record for the chapbook, *17 Gay Pesama*,[10] from a digital exhibit on Queer Narratives: Zines, Comics & Small Presses[11] at the University of North Carolina at Chapel Hill Libraries.

Dublin Core record for *17 Gay Pesama* (abbreviated) Title: 17 Gay Pesama Creator: Radmilo Durović Date: 1987 Rights: http://rightsstatements.org/vocab/InC-EDU/1.0/

Imagine you are responsible for creating MARC records for some digital exhibit items, such as *17 Gay Pesama*, but you are unfamiliar with MARC. A Dublin Core to MARC crosswalk can provide you with the documentation and guidance for how to create your MARC record. View the abbreviated Dublin Core to MARC crosswalk below, and start to compare the fields listed in the record to what is listed in the crosswalk.

DCMI Metadata Terms (Dublin Core) → MARC Crosswalk (abbreviated)

Dublin Core	MARC21	Notes
title	245 |a; use |b only if there is a subtitle	See MARC documentation for guidance on indicators
creator	100 (person) 110 (organization) 111 (conference/event)	See MARC documentation for guidance on indicators
date	264 |c	See MARC documentation for guidance on indicators
rights	506, various subfields 540, various subfields	DC "rights" typically contains a statement or link to a statement with

[10] See the full Dublin Core metadata and the resource here: https://exhibits.lib.unc.edu/items/show/6142
[11] See the exhibit page here: https://exhibits.lib.unc.edu/exhibits/show/queernarratives/zines

		various pieces of information. Although |a in either MARC field can be applied most broadly here, see MARC documentation for guidance on other appropriate subfields.

Although the crosswalk above translates MARC to Dublin Core, it doesn't tell you exactly everything you need to know to translate your record. That is because crosswalks are built to translate schemas, not individual records. For example, the crosswalk doesn't tell you what to do with the Dublin Core "creator" field, but rather gives you a choice: you can use MARC 100 if the creator is a person, 110 if the creator is an organization, or 111 if the creator is a conference. Your choice depends on which type of creator your record has, which may differ from record to record. A more complicated example is the mapping for Dublin Core "rights": which field and subfields should you use? Your Dublin Core record contains a URL in the "rights" field. Based on MARC documentation for the 506, it seems 506 subfield |u, which stores URLs, could be an appropriate choice. How might this change if the Dublin Core "rights" field in this record contained a different type of information?

The crosswalk doesn't tell you exactly what to do, but it does note which MARC fields are equivalent to the Dublin Core fields in your record and provides notes about factors to consider when implementing the crosswalk. After looking over the crosswalk, you might decide to translate the Dublin Core record into this MARC record.

MARC record for *17 Gay Pesama* (abbreviated)

100 0 |aRadmilo Durović
245 10 |a17 Gay Pesama
264 1 |a1987
506 1 |uhttp://rightsstatements.org/vocab/InC-EDU/1.0/

Crosswalks can't translate metadata from one schema to another for you. Rather, they serve as a reference tool on the equivalencies between two schemas. A person or computer program would need to take the information provided by a crosswalk and perform actions to achieve a transformation. The exception is automated crosswalks, which is a computer program that can implement metadata mappings in order to transform metadata from one schema to another.

To summarize, crosswalks are often used as reference documents, often presented as tables, that aid in the transformation of metadata from one schema to another. Depending on the complexity of the schemas on each side of the crosswalk, even the most detailed crosswalk cannot always tell you exactly how to translate each record because crosswalks translate between schemas and not specific records.

Section 4 Quiz Questions

According to this section, how are metadata mappings between schemas most often presented in libraries?

a) As tables
b) As computer programs
c) As language documents

If you have a crosswalk, you won't need to look at schema documentation.

a) True
b) False

Crosswalks should be built based on the details of a single record.

a) True
b) False

Section 4 Discussion Question

How do you know if you can trust a crosswalk?

Chapter 6 Recommended Readings

Chan, L. M., & Zeng, M. L. (2006). Metadata Interoperability and Standardization - A Study of Methodology Part I. *D-Lib Magazine*, *12*(6). https://doi.org/10.1045/june2006-chan.

Chapter 6 References

Chan, L. M., & Zeng, M. L. (2006). Metadata Interoperability and Standardization - A Study of Methodology Part I. *D-Lib Magazine*, *12*(6). https://doi.org/10.1045/june2006-chan.

Dublin Core Metadata Initiative. *DCMI Metadata Terms*. Retrieved December 2020 from: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

Hodges, G. (2001). Metadata Made Simpler. *National Information Standards Organization*. Retrieved November 2020 from https://earthref.org/ERDA/download:328/.

Library of Congress Network Development and MARC Standards Office. MARC 21 Format for Bibliographic Data: Table of Contents (Network Development and MARC Standards Office, Library of Congress). Retrieved December 2020 from: https://www.loc.gov/marc/bibliographic/.

Library of Congress Network Development and MARC Standards Office (2008). MARC to Dublin Core Crosswalk. Retrieved December 2020 from: https://www.loc.gov/marc/marc2dc.html.

St. Pierre, M., & LaPlant, W. P. (1998). *Issues in Crosswalking Content Metadata Standards - National Information Standards Organization*. Retrieved December 2020 from: https://groups.niso.org/publications/white_papers/crosswalk/.

Chapter 7: Archival Metadata

Archival studies is a separate but related field with its own history, community, philosophies, jargon, standards, and practices. Archives and libraries are often studied or practiced in shared physical or organizational spaces due to shared goals of providing access to information and some overlapping practices of storage and preservation.

Although this book focuses on Metadata in a digital library setting, many of the concepts discussed in previous and subsequent chapters can also be applied to digital archives. Furthermore, because of shared goals of digital access or digital preservation, libraries and archives in the digital world can look more similar to each other than their non digital counterparts. It is important to understand the differences between library and archival practices and metadata, however, to develop your own ideas of which or how metadata concepts in this book are applicable to archives.

This chapter will not be able to teach the nuances of archival metadata, and those interested in archives and archival metadata should look to additional resources to gain a better, broader, or deeper understanding of those sampled in this chapter.

Section 1: Important Archival Concepts

This section discusses just a few archival concepts important to understand before discussing archival metadata.

Important vocabulary in this section
Archives_____
Records (archival records) _____
Fonds_____
Respect des fonds_____
Provenance_____
Original order_____

The Society of American Archivists (SAA) says that ***archives*** are "...permanently valuable ***records***...of people, businesses, and government.[...]".[12] SAA also says an *Archives* is an organization that manages archival records or may refer to a building that stores archival records. "People" in the definition above includes individuals as well as families and communities, loosely defined. Notice also the word "records" in the definition above. In previous chapters, when I used the word "record," I was talking about a metadata record, which is a bunch of metadata about a single resource. This chapter will refer to *archival records*, which are resources (usually unpublished and no longer edited or regularly used). Examples of records (archival) include letters, photographs, meeting minutes,

[12] "The word *archives* (usually written with a lowercase *a* and sometimes referred to in the singular, as *archive*) refers to the permanently valuable records—such as letters, reports, accounts, minute books, draft and final manuscripts, and photographs—of people, businesses, and government. These records are kept because they have continuing value to the creating agency and to other potential users. They are the documentary evidence of past events. They are the facts we use to interpret and understand history." The Society of American Archivists, retrieved October 2020 from https://www2.archivists.org/about-archives

home videos, receipts—any documents created during the lives or operations of persons or groups of people that serve as historical evidence of the past.

What is the difference between a metadata record and an archival record?

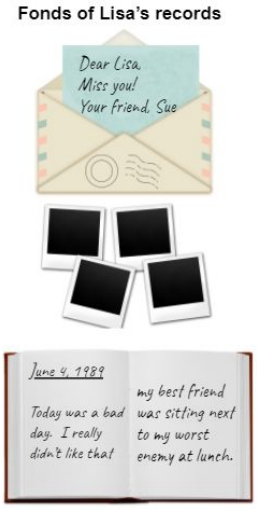
a) A metadata record is only used in libraries.
b) They are essentially the same. A metadata record contains metadata about a resource, and an archival record contains metadata about an archival resource.
c) A metadata record contains metadata about a resource, whereas an archival record is a resource such as a letter, photograph, note, etc. that serves as historical evidence.

Based on the definition of archives, which of these is a record in the archival sense?

a) A MARC metadata record
b) Meeting minutes
c) Dublin Core about a photograph

It is important to understand the difference between a metadata record and an archival record because they can both be referred to as “a record” or “records.” Pay attention to when the word “record” appears when talking about archival metadata and ask for clarification if you are not sure which type of “record” is being referred to.

Lisa



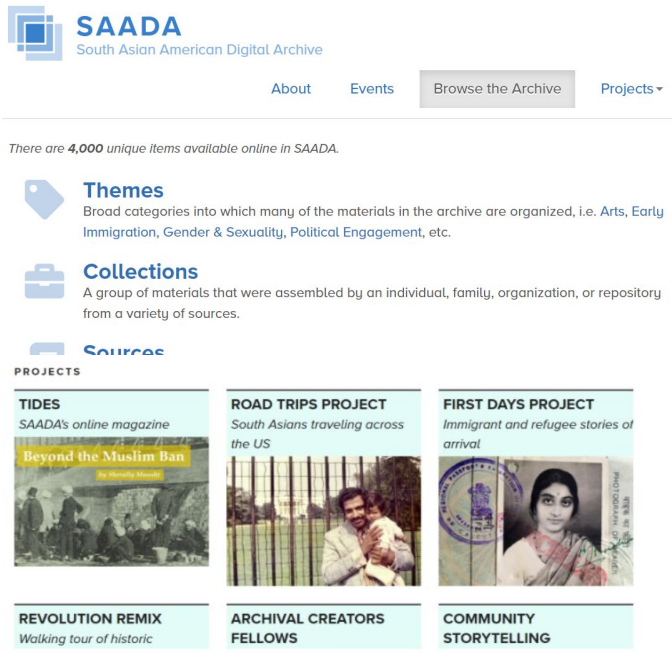
Archival records are stored in groupings for easier management an access. Arguably, the most significant grouping for archives is the *fonds*. SAA defines *fonds* as the entire body of records (e.g., all the letters, photos, notes, etc.) that is produced by a single person, family, or organization. An example of a *fonds* might be the letters, photographs, and diaries of Michelle Obama.

Another example might be the receipts, minutes, and reports of a business like Amazon.

Respect des fonds and original order

Two 19th century, European concepts are considered foundational principles for maintaining archives today: *respect des fonds* (also referred to as *provenance*) and *original order*. *Respect des fonds* refers to keeping records from the same origin (i.e., from a single person, family, or organization) together and not mixing them with other fonds. The argument behind this principle is that there is value in understanding the origins of records; that understanding where, why, and how archival records were created and who created them helps us to understand history better.

Original order refers to maintaining a group of records as close to their original organization or order. The argument behind this principle is that knowing the way a person or institution chose to organize, group, and order their records provides contextual value through relationships implied by the proximity of items. For example, if a diary includes a loose photograph of a person used like a bookmark, there may be meaning in the placement of that photo at that page in the diary.



South Asian American Digital Archive (SAADA)
Above. Browse the Archive: https://www.saada.org/browse
Below. Home Page: https://www.saada.org/.

Although provenance and original order are foundational to modern archives in the United States, they were not created as exclusive ways to group and organize records and can perpetuate inequity when record "origins" are complex.

The originator of *respect des fonds* only applied this principle to "ancient" or historic documents rather than contemporary ones (Barlett, 1992). Furthermore,

many subject or format specific archives, digital archives, and community archives have used different or multiple methods of managing archives. F.I.L.M. Archives, Inc., for example, has three “collections” in which films from various origins or fonds are grouped together; only one collection, the News 12 collection, might be considered a fonds. Similarly, the South Asian American Digital Archive (SAADA), both a digital and community archive, has examples of materials grouped in what might be considered *fonds*, as well as records from different origins grouped and organized together by themes or by project.

Respect des fonds is complicated when considering that a fonds from a person, family, or organization might contain information about or created by other people, non-family, or non-organization members. For example, my fonds might include letters that people have written to me. Although I possess the letters, I was not the creator of the letters, and they may contain details the writers never intended others to see. Jared Drake discussed these complications in terms of colonialism and imperialism in his 2016 talk “RadTech Meets RadArch: Towards A New Principle for Archives and Archival Description.” Drake points out that when *respect des fonds* was being adopted as a founding archival principle in the United States, people of color, women, sexual and religious minorities, people who are disabled, and non-Anglo immigrants were structurally or legally (or both) excluded from ownership of various forms, including ownership their records (2016). What of the people who were barred from asserting ownership or control of these records, yet whose lives are documented within?

To summarize, being familiar with the archival concepts reviewed in this section is helpful in understanding archival metadata. In archives, the word “record” often refers to archival records, which are the documents and other evidence (e.g., diaries, receipts, memos, photographs, etc.) of people, organizations, and governments. This is distinct from a metadata “record.” Although the principles of *respect des fonds* and original order are considered foundational to archivists in the United States, this approach can negate the

origins of records and autonomy of people who are marginalized (e.g., not white, male, cis, Anglo, hetero, able-bodied, or Christian).

Section 1 Quiz Questions

Which archival principle relates to provenance?

a) Records
b) Original order
c) Respect des fonds

Based on what you've read, which explanation of original order is the most accurate?

a) It is important to know the origin of the records to understand its historical significance.
b) The way records were organized by their creators may provide helpful contextual information.
c) Archivists are required to keep records in the same order they received records in.

Which type of "record" is being referred to in the following statement? The record was created using Dublin Core and describes a digitized letter.

a) Archival record
b) Historic record
c) Metadata record

Section 1 Discussion Question

Many archival materials in older and large institutions carry archival records created during a time when a small portion of the population had the right to own property or vote. What are ways you can image this affecting archives?

Section 2: Basic Archives Metadata Concepts

In this section, you will learn some basic archival metadata concepts, many of which are further discussed in *Describing Archives: A Content Standard (DACS)*, the widely used content standard for archival metadata.

Important vocabulary in this section
Primary source _____ Arrangement _____ Series _____ Container level _____ Granularity _____ Administrative / Biographical History Note _____ Access Points _____

Emphasis on "the Collection" versus "the Item"
Whereas library metadata focuses on individual items, archival metadata prioritizes collection level description.

Library resources are typically described at the item level, meaning each resource has its own metadata. Libraries specialize in providing access to published resources, which create content or information not intended to be unique from copy to copy. Metadata for published resources similarly lacks uniqueness from copy to copy, making the metadata easier to reuse. Published resources are also assumed to be self-contextualizing, meaning all one needs to understand the resource is inside the resource. Metadata for these resources relies less on context and more on descriptive and publication information to help people find the resource.

By contrast, archival records or resources are typically described at the collection level, meaning they are described as a group of related resources. Archives specialize in providing access to unpublished or meaningfully unique

resources sometimes referred to as *primary sources*.[13] Creating item level metadata for archives is time consuming: the metadata is unlikely to be reused for other unique resources, and archives contain high volumes of items because each letter, each photo—each record—is an item. Archival records often are not self-contextualizing, so other resources help provide context for individual items such as a single letter. This characteristic of archival records and the time consuming nature of item level metadata for archives leads archivists to prioritize broad description that emphasizes context and explanations about how records within the same grouping are related to each other.

Based on what you've read so far, which answer best explains why context is emphasized in archival records?

a) Archival records are large in volume so context helps provide organization to the records.
b) Archival records are similar to library records, so context is helpful for both.
c) Archival records are not self-contextualizing, context is provided in the metadata like explaining how records in a collection are related.

Arrangement, Granularity, and Levels of Description

Arranged in 5 series: 1. Subject files concerning refugee issues, 1978-1997. 2. Project Ngoc organizational files, 1987-1997. 3. Visual and audiovisual materials, 1985-1997. 4. Artwork, 1987-1997. 5. Newspaper clippings, 1980-1998.

Text from 3.2.3 of Describing Archives: A Content Standard.

In archives, *arrangement* refers to how a body of records is organized. Metadata for archives should reflect the arrangement of the records. Archival records can be organized using different methods (e.g., alphabetical, chronological, by subject), and archivists often use the principle of original order when possible. Non-digital

[13] SAA Dictionary. Primary source definition: "material that contains firsthand accounts of events and that was created contemporaneous to those events or later recalled by an eyewitness".

archives[14] have traditionally used mutually exclusive hierarchies based on membership in a collection to organize records. The top level of the hierarchy is typically the collection (such as fonds or other methods of grouping related records); the bottom level are the items—the individual records—within the collection. Between these two levels, depending on the size and complexity of the collection, other levels of hierarchy such as *series* or subseries may be used. Even the physical containers records are stored in such as boxes and folders can themselves act as additional groupings and layers of hierarchy (*container level*) before the items.

Metadata can be used to describe any of these levels of *granularity*. What's more, archival metadata can consist of just one of these levels (single level description) or consist of multiple levels (multilevel description). Guidelines on what should be in single level or multilevel descriptions are provided in more detail in Chapter One of DACS.

Based on what you've learned about granularity, which of these options best illustrates an example of a collection level description followed by an item level description?

a) The Nakasone Papers consist of 60 letters Nakasone wrote or received during their lifetime. Letter dated Feb. 30, 1999 between Nakasone and their Australian penpal.
b) Photograph of Billy Warlock. Photograph of Carmen Elecktra.
c) Meeting minutes from Dec. 2010. News Clipping of article about offshore drilling.

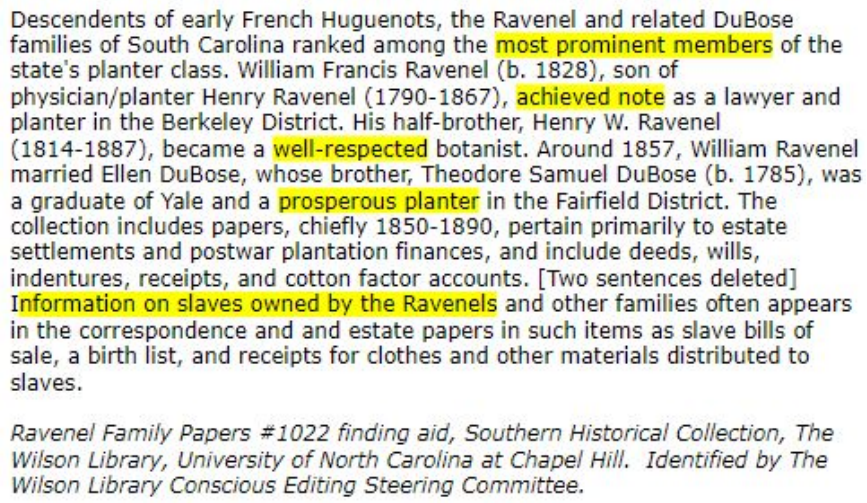
Describing creators

Describing creators of archival records can provide helpful context for understanding the records and subsequently, learning about history. For this reason, metadata for archives often provide more information about archives creators than libraries provide about creators of library resources. Although name authority records, such as those used for library materials, can and are

[14] Although many digital archives also use hierarchy to organize collections, digital technology allows archival records to act more independently. Through the combination of metadata and digital archive software, resources can be displayed as part of multiple collections or searched and browsed independently based on search terms.

often used for archives, DACS guides archivists to create an *Administrative / Biographical History* Note to provide details about a creator that might be relevant to understanding their collection.

As discussed in the previous section, the principle of *respect des fonds* can complicate and obscure creatorship. Remember that a person's fonds might consist of records that person did not create, yet the biographical history note typically centers around the person, family, or organization identified as the "creator" of the fonds.



Descendents of early French Huguenots, the Ravenel and related DuBose families of South Carolina ranked among the most prominent members of the state's planter class. William Francis Ravenel (b. 1828), son of physician/planter Henry Ravenel (1790-1867), achieved note as a lawyer and planter in the Berkeley District. His half-brother, Henry W. Ravenel (1814-1887), became a well-respected botanist. Around 1857, William Ravenel married Ellen DuBose, whose brother, Theodore Samuel DuBose (b. 1785), was a graduate of Yale and a prosperous planter in the Fairfield District. The collection includes papers, chiefly 1850-1890, pertain primarily to estate settlements and postwar plantation finances, and include deeds, wills, indentures, receipts, and cotton factor accounts. [Two sentences deleted] Information on slaves owned by the Ravenels and other families often appears in the correspondence and and estate papers in such items as slave bills of sale, a birth list, and receipts for clothes and other materials distributed to slaves.

Ravenel Family Papers #1022 finding aid, Southern Historical Collection, The Wilson Library, University of North Carolina at Chapel Hill. Identified by The Wilson Library Conscious Editing Steering Committee.

As Jarrett Drake has discussed, because the creators of fonds have historically been those with the structural and legal power and access to own and donate records, many of these administrative / biographical notes reinforce those structural inequities found in the archival record. Although the note is supposed to provide historical context for the records, Drake says "In this note, archivists often write massive memorials and monuments to wealthy, white, cisgendered and heterosexual men, including selective details about the creator that have minimal bearing on the records, and instead serve to valorize and venerate white western masculinity" (Drake, 2016).

Access points

As with library metadata, metadata for archives relies on access points to help people answer specific research questions. DACS identifies the following access points in its Overview of Archival Description section: Names, Places, Subjects, Documentary forms, Occupations, Functions and Activities.

Names, places, subjects and even occupations should seem familiar to you from learning about controlled vocabulary. Documentary forms are types and forms of documents found in archives. Examples of documentary forms include, but are not limited to letters, photographs, diaries, memos, interviews, and clippings. Functions and activities are "terms indicating the function(s), activity(ies), transaction(s), and process(es) that generated" the archival records (Overview of Archive Description in DACS). Examples of Functions and Activities might include but is not limited to something like overseeing of foreign relations program or supervision of junior faculty.

To summarize, one of the distinctive differences between library and archives metadata is the greater emphasis in archives on context and describing resources in the aggregate. Archives metadata reflects the arrangement of the archival records, which often rely on hierarchical structuring. Archives metadata can contain single level or multiple level descriptions. Like library metadata, archives metadata rely on access points. Unlike library metadata, the need for more context in archives metadata leads to more detail about the people the records are by or about.

Section 2 Quiz Questions

Which of the following is not a reason mentioned for archival metadata emphasizing collection description over item description?

a) Describing items is not important because they are only single resources.
b) Describing items is inefficient due to the uniqueness and volume of items in a collection.
c) Describing items in relation to other items and the broader context of the collection is most helpful.

What is meant in archival metadata by granularity?

a) The level (e.g., collection, series, item) at which an archival collection is described.

b) The size and amount of metadata created.
c) Which records in the collection the metadata is describing.

Which of the following is not an access point listed above?

a) Time periods
b) Functions and Activities
c) Documentary forms

Section 2 Discussion Question

How do you think an administrative history note for a government or organization and a biographic history note for a person compare? What kind of information in both would you imagine necessarily to understand archival records in a collection?

Section 3: Archives standards: DACS, EAD, and EAC-CPF

In this section, I will provide an overview of archives standards, focusing on three three standards: Describing Archives: A Content Standard (DACS), Encoded Archival Description (EAD), and Encoded Archival Context — Corporations, Persons, and Families (EAC-CPF).

Important vocabulary for this section
DACS _____
Finding aid _____
EAD _____
EAC-CPF _____

Describing Archives: A Content Standard (DACS)
DACS is the content standard for Archives that is endorsed, created, and maintained by the Society of American Archivists (SAA). Starting in 2019, DACS became an open source standard published and revised through the online, code sharing platform GitHub, which allows anyone to propose changes to the standard.

As a content standard, DACS provides rules and guidance on what kind of information one should include in archival metadata. DACS also provides guidance on where to get the information that is included in the metadata, and in some cases, guidance on how to record the information.

DACS is agnostic to encoding language, data forms and formats, and software platforms or systems. It can be used to guide the creation of MARC metadata records, records in a relational database or on a spreadsheet, or

text based inventory lists. Despite this versatility, DACS is mostly associated with the *finding aid*.[15]

There are two main sections of DACS: guidelines on how to describe archival collections (Part 1) and guidelines on how to describe people (Part 2). The first section includes chapters on metadata elements to include to help identify a collection, describe how the collection is organized, summarize isness and aboutness, explain how a collection can be accessed and used, indicate provenance and the chain of custody of the collection, point to related materials, and more. The second section focuses on the creation of archival style authority records for entities such as individual people, families, and corporate bodies with chapters on which and what form of name to use, alternative names, biographical / historical and other contextual or identifying information about the entity, related entities, and more.

DACS supports the description of persons, families, and corporate bodies as creators, because records are created through the functions and activities of people. Persons, families, and corporate bodies are each treated differently because each type of entity has distinct characteristics. DACS defines a person as a human, which may seem obvious, but is complicated when thinking of the history of Black people, Indigenous peoples, and those from different marginalized groups not treated as whole persons.

DACS defines a corporate body as “an organization or group of people identified by a name and that acts, or may act, as a unit, or an institutional position held by a person” (like a President). This definition is broad and doesn’t outline any monetary, structural, geographic, or legal requirements; corporate bodies could be clubs, businesses, non-profits, governmental units, or international organizations. DACS defines family as “two or more people related through marriage, birth, adoption, or other legal manner, or who present themselves as a family.” The definition for family also attempts to be broad by not imposing biological or legal restrictions on family. These

[15] “a description that typically consists of contextual and structural information about an archival resource” SAA Dictionary.

definitions for family and corporate body are broad enough to accommodate most groups of people, but the definitions blur when considering certain communities that may have elements of both.

Encoded Archival Description (EAD)

Although DACS does not require that archives be described in a finding aid, finding aids are the most commonly used form of metadata found in archives. Finding aids have historically used a variety of technologies including typewritten paper, word processing files, HTML, and XML. *EAD* is a schema for encoding finding aids in XML. Thanks to widely adapted archives Information Management Systems such as ArchiveSpace that automate the creation of EAD encoded finding aids, EAD is more widely used than ever before.

EAD is an XML schema that provides a library of XML tags used to mark up or tag elements of a finding aid to make it easier for a computer to process and display the information in the finding aid. You can find documentation on EAD on the Library of Congress website:
https://www.loc.gov/ead/

Look at this excerpt from a finding aid:

```
Box 1
Folder 1 photographs
Folder 2 letters
```

Above is part of a description of the containers and contents of a collection. Box 1 contains one folder of photographs and another folder of letters. Although a person is able to read this and discern meaning from it, a computer would view this description as a jumble of text. EAD uses XML to tag parts of this text as metadata elements and their values so that a computer recognizes this as structured data that can later be processed in a

number of ways (e.g., sorted, formated). The same text might look like this encoded in EAD:

```
<c02 level="file">
        <did>
                <container localtype="box">1</container>
                <container localtype="folder">1</container>
                <unittitle>photographs</unittitle>
        </did>
</c02>
<c02 level="file">
        <did>
                <container localtype="box">1</container>
                <container localtype="folder">2</container>
                <unittitle>letters</unittitle>
        </did>
</c02>
```

Can you see which line of code explains the box and folder numbering?

Based on what you know of the container description, what information does <unittitle> provide in this example?

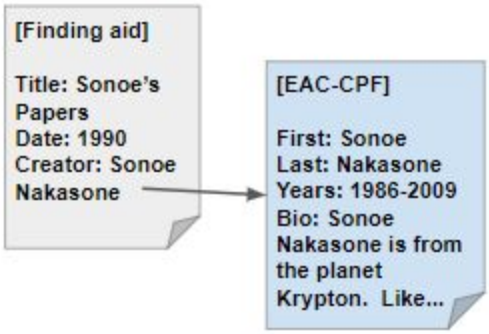
a) The title of the box
b) The title of the folder
c) The title of the collection

EAD has hundreds of tags, and just because it is a standard, it doesn't mean it is implemented the same way at every library. Many institutions and information management systems will come up with their own local best practices, guidelines, or templates to supplement standards like EAD because of the variability on how elements are interpreted and implemented.

Encoded Archival Context: Persons, Families, Corporate Bodies (EAC-CPF)

Again, DACS is agnostic to any form of metadata or language of encoding, therefore, its rules and guidelines for describing creators and creating authority records can be implemented using a variety of methods like creating MARC records or establishing RDF triples. Despite this, the Society of American Archivists adopted *EAC-CPF* as a standard that explicitly addressed the contextual nature of describing people and other creators of archival collections.

Finding aid descriptions with biographic information related within each finding aid. Note that both records would need to be revised when Sonoe dies to reflect the year of her death.

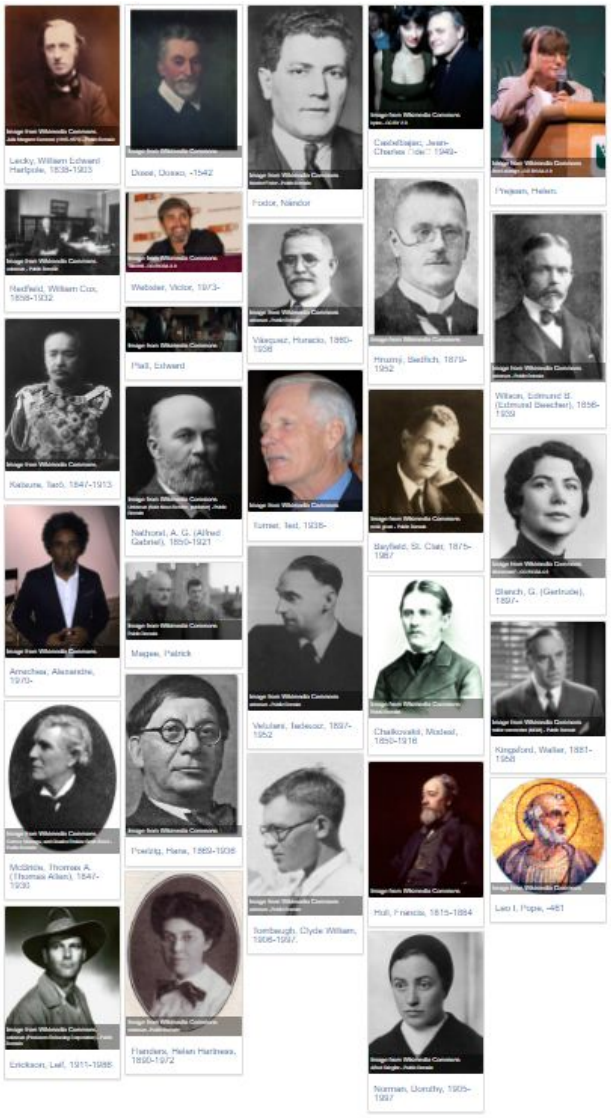


Finding aid description related to separate EAC-CPF description for Sonoe Nakasone. Note that Sonoe's EAC record can be updated independently of the finding aid.

EAC-CPF is an XML schema that provides metadata elements for implementing the rules and guidelines found in DACS. Daniel Pitti, the lead architect of EAC-CPF, argued for components of archival description to be treated as nodes, separate modular descriptions, that would then be related to each other (Pitti 2004). This would enable a creator description in EAC-CPF to contain all the contextual information about a creator and then be reused by linking it to many records or collections of records, many other creators, and many activities and functions.

Aside from the EAC-CPF website that provides access to the schema documentation, one of the most popular resources related to EAC-CPF is Social Networks and Archive Context (SNAC). Often described as “Facebook

for dead people,"[16] SNAC is "an online resource that helps users discover biographical and historical information about persons, families, and organizations that created or are documented in historical resources." SNAC is also an international cooperative that maintains the database and conversations about EAC-CPF.



As noted earlier in the chapter, the legacy of provenance or *respect des fonds* has reinforced and perpetuated racism, colonialism, sexism, and ableism by assigning creatorship or ownership of records primarily to white, able-bodied men. The majority of archives "creators," as identified by archival finding aids and creator authority descriptions, reflect this supremacy of white identity. Even a screenshot of the SNAC website on a random day is likely to showcase more men and more White or European "creators" than women, creators of color, or non-Western creators on its home search page. This isn't because SNAC explicitly supports the suppression of people who are not male or white, but because of the overwhelming numbers of white male entities recognized as archive creators in our country's most prominent archives.

[16] Honestly not sure who to attribute this unofficial tagline to. I frequently heard this description used by Tammy Peters of the Smithsonian Institution Archive in the early 2010s.

One current effort of SNAC and of archivists interested in a more diverse historic record, is to recognize that institutions like slavery prevented many African Americans with archival records from owning their own records. As a result, the SNAC community is exploring how best to represent the names of the enslaved in the context of archival authority. Although this work has reparative qualities and increases access of records to descendents of enslaved people, the idea of "authority" over names of those who did not and in death do not have the power to name themselves could itself be viewed as an act of violence.

To summarize, DAC, EAD, and EAC-CPF are three widely used metadata standards in archives. DACS is a content standard that is language and form agnostic, but is often used to support finding aid creation. EAD is an XML schema for encoding finding aids. EAC-CPF is an XML schema used to create archival style authority records for persons, families, and corporate bodies.

Section 3 Quiz Questions

What kind of standard is DACS?

a) A controlled vocabulary
b) A classification system
c) A content standard

Which of the following best explains what EAD is used for?

a) To markup finding aids in XML
b) To implement DACS in XML
c) To create authority records in XML

Based on the reading, which choice below best explains the purpose of an archival style authority record?

a) The same purpose as library style authority records
b) To have something that will be compatible with an EAD finding aid
c) To have more context about creators and link to multiple resources.

Section 3 Discussion Question

How do you think authority records of creators who were not originally recognized as creators should be treated? What considerations should be made when creating these authority metadata records?

Section 4: Types of metadata in archives

You may have already noticed that archives metadata uses various types of metadata to describe, explain, provide access to, and manage archival records. This section will highlight common types of metadata found in archives aside from the descriptive metadata and name authorities metadata, which this chapter has focused on until now.

Important vocabulary for this section
Access metadata _____
Patron _____
Accession _____

Access metadata

Unlike libraries, most archives do not allow archives users to freely browse through or retrieve their own items. Archives materials are unique or rare, so loss, damage, or theft results in greater consequences (i.e., permanent information loss). Further, as primary sources, archives content is often unfiltered, unedited, and can contain personal information. The rarity, uniqueness, and personal nature of archival records can lead archives or those donating to them to impose restrictions on materials. Restrictions may be noted in public metadata so users know under what conditions they are allowed to access materials, or hidden as administrative metadata for only archives employees to see and manage. Metadata can also be structured such that a computer could automate restrictions of digital resources based on access metadata. Noting embargos (a ban or restriction on an item) and an access control matrix are just two examples of access metadata employed for archival resources.

Patron metadata

Although libraries also collect *patron* (a user of library or archives resources or services) metadata, archives patron metadata is often more detailed. This is partially a legacy of when only researchers of a certain level or type were

encouraged to use archives collections (this is still the case at many archives where only “serious” researchers are permitted). Archives patron metadata is also detailed because archives are concerned about the security of collections. It is not uncommon to collect addresses and phone numbers as well as keep a log of which materials a patron accesses in case records go missing.

In her 2015 Code4Lib Keynote address, Kate Krauss warned of the risks collecting patron metadata poses. Rather than delete this information, it is not uncommon for archives to keep patron metadata for years. Krauss puts the collection of patron metadata into the context of data privacy and the history and current practices of surveilling marginalized communities.[17] She argues that collecting and retaining detailed patron metadata puts patrons at risk should the metadata be subpoenaed. What are the unintended consequences of making such patron metadata a requirement for using archival collections?

Accessions Metadata

Archives keep various forms of administrative metadata, one prominent type being accessions metadata. *Accession* means “to take intellectual and physical custody of materials [...]” (SAA Dictionary). The word accession is also used to refer to the materials taken in during the accession process, e.g., “We have a new accession of yearbooks donated by the Nakasone family.”

Accessions metadata varies from institution to institution, but is likely to include an identifier (i.e., an accession number) and provenance information such as who donated it, their contact information, the date, and a note or copy of any legal documents confirming the gift. It is also common to include an inventory or summary of the materials being accessioned. Some institutions will also include appraisal information. Appraisal is “the process of identifying materials [...] that have sufficient value to be accessioned” (SAA Dictionary). If during the appraisal process, the archive decided to keep only half of what

[17] Krauss points to FBI surveillance of activists from Black Lives Matter and identifying information that could out or endanger patrons who are trans as examples.

was donated and return or discarded the rest, these decisions may be recorded in the accessions metadata. Accessions metadata is important because a fonds or other top group of archival records could consist of items from multiple accessions.

Preservation

Archives collect records that they think have lasting value to society. Preservation is crucial because to achieve lasting value, records must last; preservation metadata supports this goal.

Chapter 5 has already provided an overview of preservation metadata. Here, I will share an example of how metadata supports digital preservation.

The OAIS (Open Archival Information System) Reference Model is an archives standard that outlines how digital preservation in a digital archive or library could work. Part of this model is the idea of information packages, which store metadata and resources. Think of an information package as a big bag that contains a bunch of presents. Each present in the bag (i.e., the information package) contains something special. One present might contain descriptive metadata in MODS. Another present might contain preservation metadata in PREMIS. Another present might have structural metadata. The best present contains toys (i.e., the resources), which the descriptive, preservation, and structural metadata from the other presents are about.

One metadata tool used to create information packages is a schema called *METS* (Metadata Encoding and Transmission Standard). METS allows you to encode data (i.e., resources) and metadata into one neat package. Below is an abbreviated example of a Submission Information Package (SIP) in METS.

```
<METS:METS>
        <METS:structMap TYPE="pdf">
                <METS:div ORDER="1" LABEL=”frontMatter” DMDID="DM1" ID="div1">
                        <METS:div ORDER="1" LABEL="Title page front" TYPE="page”>
                                <METS:fptr FILEID="123"/>
                        </METS:div>
```

```
                    <METS:div ORDER="2" LABEL="Title page back" TYPE="page">
                        <METS:fptr FILEID="321"/>
                    </METS:div>
                </METS:div>
            </METS:structMap>
            <METS:amdSec>
                <METS:rightsMD ID="ADMRTS1">
                    <METS:mdWrap MDTYPE="OTHER"
                OTHERMDTYPE="METSRights">
                        <METS:xmlData>
                            <rts:RightsDeclarationMD>
                                <rts:RightsHolder>

                                <rts:RightsHolderName>Sonoe</RightsHolderNa
                                me>
                                </rts:RightsHolder>
                            </rts:RightsDeclarationMD>
                        </METS:xmlData>
                    </METS:mdWrap>
                </METS:rightsMD>
            </METS:amdSec>
            <METS:fileSec>
                <METS:fileGrp ID="OBJECTS">
                    <METS:file CHECKSUM="b39fba28d527b2823b871f115f913f67"
                    CHECKSUMTYPE="MD5" ID="9999" MIMETYPE="pdf"
                    SIZE="4528808"/>
                </METS:fileGrp>
            </METS:fileSec>
<METS:METS>
```

Structural and rights metadata are presented above. The structural metadata starts on the second line of XML with the <METS:structMap> element. In a METS SIP, structural metadata for hundreds, even thousands of resources could be included in a single METS file. Below METS:structMap, is the administrative metadata section (<METS:amdSec>) that contains rights metadata (<METS:rightsMD>). Like the structMap section, the rights metadata section can contain rights metadata for many resources in the same METS file. The last portion of the METS metadata above is the file section (<METS:fileSec>). The file section stores all the digital resources.

To summarize, there are many types of metadata that help archives to manage and preserve records and make records available to archives users. This section discussed access metadata, accessions metadata, patron metadata, and preservation metadata.

Section 4 Quiz Questions

Which of the following types of metadata is included in accession metadata, according to this reading?

a) Access metadata
b) Provenance metadata
c) Technical metadata

Which type of metadata would include information about an embargo, according to this reading?

a) Access metadata
b) Provenance metadata
c) Access control matrix

Which of the following explanations of METS is most accurate based on what you read?

a) METS is a schema that requires the use of MODS and PREMIS.
b) METS is a schema to is used for
c) METS can be used to create information packages containing metadata and resources.

Section 4 Discussion Question

What kinds of patron metadata do you think archives need to keep and why?

Chapter 7 Recommended Readings

Describing Archives: A Content Standard

Drake, J. (2016, April 07). RadTech Meets RadArch: Towards A New Principle for Archives and Archival Description. https://medium.com/on-archivy/radtech-meets-radarch-towards-a-new-principle-for-archives-and-archival-description-568f133e4325

Hart, L. (2020). Conscious Language for a Jim Crow Archive. https://doi.org/10.17615/brty-6g17

Chapter 7 References

Bartlett, N., 1992. Respect des Fonds. Primary Sources & Original Works, 1(1-2), pp.107-115.

Describing archives: A content standard [DACS 2019.0.3]. (2019). Chicago: The Society of American Archivists. https://github.com/saa-ts-dacs/dacs (retrieved October, 2020).

Drake, J. (2016, April 07). RadTech Meets RadArch: Towards A New Principle for Archives and Archival Description. https://medium.com/on-archivy/radtech-meets-radarch-towards-a-new-principle-for-archives-and-archival-description-568f133e4325 (retrieved October 17, 2020).

Krauss, K. Keynote, delivered at Code4Lib 2015. (Starts around min. 22) https://youtu.be/Dd04w--7EuY?t=1306 (accessed September 2020).

Pitti, D. (2004). Creator Description: Encoded Archival Context, Cataloging & Classification Quarterly, 38:3-4, 201-226, DOI: 10.1300/J104v38n03_16

Screenshot of Ravenel Family Papers #1022 finding aid, Southern Historical Collection, The Wilson Library, University of North Carolina at Chapel Hill. Identified as an example of aggrandizement by Laura Hart, founding co-chair of The Wilson Library Conscious Editing Steering Committee. Mimicking an example from Jarett Drake's 2016 RadTech Meets RadArch [...] talk cited above. https://finding-aids.lib.unc.edu/01022/ (accessed October 2020).

Social Networks and Archival Context. https://portal.snaccooperative.org/about (accessed October 2020).

Society of American Archivists. Accession. https://dictionary.archivists.org/entry/accession.html (retrieved December, 2020).

Society of American Archivists. Appraisal. https://dictionary.archivists.org/entry/appraisal.html (retrieved December, 2020).

Society of American Archivists. Primary source. https://dictionary.archivists.org/entry/primary-source.html (retrieved October 18, 2020).

What Are Archives? Society of American Archivists. https://www2.archivists.org/about-archives (accessed October 2020).