Lecture Notes for Week 2, Class 2

Readings:

- Manovich, Lev. "Database as Symbolic Form." *Convergence: The International Journal of Research into New Media Technologies*, vol. 5, no. 2, June 1999, pp. 80–99.
- Pomerantz, Jeffrey. "Introduction." *Metadata*, The MIT Press, 2015, pp. 1–18.
- Gebru, Timnit, et al. "Datasheets for Datasets." *Communications of the ACM*, vol. 64, no. 12, Dec. 2021, pp. 86–92.
- **Optional**, complimentary reading for further interest:
    - Hoffman, Gretchen. "How Are Cookbooks Classified in Libraries? An Examination of LCSH and LCC." *Proceedings from North American Symposium on Knowledge Organization*, vol. 4, no. 1, 2013, pp. 100–11.

Agenda

1. Data's narrative

- In our previous session we discussed where the notion of 'data' comes from, what it originally meant, and how it has evolved over time. We also discussed the ways in which humanists think about data, and how that's distinct from the ways in which scientists and social scientists think about it.

- Today, we will discuss the ways in which data can tell a story -- we are slowly but steadily making our way to the labs in the near future where data will be wrangled, distorted, modelled, and visualized. But before we get there, we need to understand HOW these narratives that data can tell are constructed.

- Manovich tells us a story about the database. He argues that the database is a new symbolic form of the 20th century. How different is it from the narrative form? His conclusion is that the database is a new form of cultural expression. Although it will sometimes present itself as a narrative, it is fundamentally different from a narrative.

- Consider, for example, the box of the cd *Encarta* in the 90s had a narrative form, but the cd itself was a database.

- Pomerantz's introduction to his book on Metadata also suggests that metadata is something else than a narrative. Although it can be used to tell a story, and this is also a dangerous thing, he warns -- giving the example of the NSA and Edward Snowdon.

- Pomeranz is on the same page as Manovich: metadata has the ability to tell a story, but it does not. It is fundamentally different from a narrative. The example of the NSA and the phone records on page 2 illustrates this: a story is being created from the metadata, but it is not the metadata itself that is the story: one caller calls the hardware store, then ...

- Next week, we will visit Special Collections, where we will see a collection of postcards. What is the story of the postcards? When does a pile of postcards become a story? And most importantly what are the categories of data that we need to capture in order to tell that story? What is the narrative that we want the postcards to tell?

- The story of the postcards is not in the postcards themselves, but in the metadata that we attach to them. The metadata is the narrative of the postcards.

2. Understanding metadata

- Descriptive metadata: describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.
- Administrative metadata: provides information to help manage a resource, such as when and how it was created, file type, and other technical information.
- Structural metadata: indicates how compound objects are put together, for example, how pages are ordered to form chapters.
- Preservation metadata: contains information needed to archive and preserve a resource.
- Operational metadata: describes the processes and methods used to create, manage, and use a resource. for example, it may be necessary to emulate a specific application and operating system environment in order to interact with a digital file.
- Rights metadata: information about rights held in and over the resource.
- Provenance metadata: history of the resource.
- Contextual metadata: information about the circumstances in which the resource was created.
- Use metadata: information about how the resource has been used.
- Metadata is important because it helps users find resources, it helps users understand the resources, and it helps users manage the resources.
- A pile of postcards is just that... a pile of postcards. But if we add metadata to the postcards, we can start to make sense of them. We can start to tell a story about them.
- atomisation of the library record -- card catalogs. Can

3. Metadata standards

- Metadata standards are essential in all disciplines. But they are often highly contested and even controversial because they embody value judgments either implicitly or explicitly. Do all works, for instance, have “creators” or “authors”? Stories and songs from Indigenous communities might not. As concepts of metadata change, older standards come up for debate.
- MARC, which stands for Machine-Readable Cataloging, has been a cornerstone in library cataloging for over 50 years. However, it's important to understand that standards which were innovative half a century ago might not be sufficient for today's needs.
- Influential paper: *MARC must die*: it is a 50-year-old standard that is not fit for purpose in the 21st century. It is a relic of the past. What are some reasons why MARC must die? (I'm so lucky that nobody in this class is called Marc.)
- Problems with MARC:

1. Technological Limitations: MARC was developed in an era of limited computing capabilities. Its structure is rigid and not well-suited to modern, flexible data models like those used in contemporary web technologies.

2. Complexity and Cost: MARC is complex and requires specialized knowledge to use and interpret, leading to higher training and maintenance costs. This complexity can be a barrier to efficient data sharing and collaboration among libraries and other institutions.

3. Incompatibility with Current Web Standards: MARC does not integrate well with current web standards and technologies, such as linked data and semantic web, which limits its utility in the digital age where data interoperability and sharing are crucial.

4. Limited Flexibility: The format is not flexible enough to accommodate the diverse types of resources and metadata now commonplace in libraries, such as digital resources, multimedia, and community-generated content.

5. Globalization and Multilingualism: MARC struggles with non-Latin scripts and globalized cataloging practices, which is increasingly important as library collections become more diverse and internationally focused

4. Humanities metadata

- Dubline Core: a simple, standardized way of describing resources. It is a set of 15 metadata elements that can be used to describe digital resources such as video, images, web pages, and text documents. It is widely used in the humanities and cultural heritage sector.
- Does not refer to Dublin the city, but to Dublin, Ohio, where the standard was developed.
- Dublin Core Fields: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights.
- Dublin Core is comprised of 15 “core” metadata elements; whereas the "qualified" Dublin Core set includes additional metadata elements to provide for greater specificity and granularity.
- Here, again, we see the difference between metadata that is descriptive and fields that are operational, defining relations among parts of a collection and/or its records.

5. Classification

- If you are putting together materials that are related to each other by a particular theme — an event, site, or topic — thinking through the way they will be described is a significant feature of the intellectual work.
- In the case of an archive of queer performance, what fields in a metadata scheme allow that identity to be expressed? What types of performance need to be identified and in what vocabulary? Who defines that vocabulary? What are the politics of characterization of queer identity from within the community and from outside? Who speaks in the metadata?
- Categories that address ethnic and racial identity, sexual orientation, and gender, are particularly fraught since the act of classification can be oppressive.

- Classification is one of the most powerful forms of organizing knowledge is through the use of classification systems.
- No classification system is value neutral, objective, or self-evident. All classification systems bear within them the ideological imprint of their production. A system of identifying works of art by their creators might be inappropriate in a community where practices are tied to tradition and repetition, rather than originality and invention.
- Classification systems are closely related to metadata. The Dewey Decimal System and the Library of Congress are two different classification systems for published materials.
- The Library of Congress (LoC) has twenty-six divisions and uses the Roman alphabet as its structuring framework. It uses two of these, E and F, for the history of the Americas and only one, D, for the history of the rest of the world. While naval science and military science each have a dedicated category, caring professions do not appear, nor do activities traditionally associated with women's work, which are all subsumed under larger categories.
- Like LoC, the Dewey Decimal System begins with philosophy and religion, ends with geography and history, and reserves a special place for biography. Looking at a system like Dewey or LoC, one can discern a particular point of view about the world, the knowledge of it, and the values that shape understanding.

6. A new narrative?

- Gebru et al. are reflecting on this notion that data is not neutral! It does tell a story! It tells a story about the people who created it, the people who are represented in it, and the people who are excluded from it. It tells a story about the values of the people who created it, and the values of the people who are represented in it. It tells a story about the power structures that created it, and the power structures that are represented in it. It tells a story about the world that created it, and the world that is represented in it. It just is not a very explicit story. It is a story that is hidden in the data. Important today to bring that narrative to the surface!!!

7. Postcards' narrative

- As we conclude today's class, let's turn our attention to the Historical Postcard Collection from Princeton University, spanning from 1890 to 1960. This collection provides not just images but stories of the university's past. It includes pictures of iconic buildings, some of which no longer exist, and captures the essence of the campus environment during that era.
- Each postcard in this collection is more than just an image; it's a piece of history that tells a story. I want you to think about the narratives these postcards might hold. Consider the stories they could tell about the university's evolution, architectural changes, and even the cultural and social aspects of campus life during the first half of the 20th century.
- Remember our discussion on metadata? This is where it becomes practical. We can use metadata to organize and interpret these postcards, creating a narrative from what might otherwise be just a collection of images.
- Let's walk through the database structure. Think of each row as a different postcard and each column as a piece of information that helps us understand and contextualize it.

- As you look at this structure, consider what other types of information might be useful. How can this data help us piece together the narrative of Princeton University during this period? What stories can emerge from analyzing this collection through the lens of our metadata categories?