ADHHI Practicum – Archives Session

Tuesday November 5th
12:30 – 3:00 pm

Hello Everyone. Thanks for being here. My name is Emma James, and I'm the Project Archivist for the Industry Documents Library.

I'm going to start us off here today. My colleagues Rachel and Lisa here will also be joining in for different parts of our session.

Today I'm going to do an overview about archives as data, and in particular the significant role that metadata has as the crux for archives in digital humanities research.

Here are our objectives for the day:

OBJECTIVES:

- Approach research using digital archival collections as data
- Content from collections can be from a range of different document types – emails, flyers, videos, and more; this results in a range of metadata.
- Discussion on the relationship to archives and metadata (processing and applied metadata standards)
- Introduction to thinking about born-digital data
- Structured and unstructured data – metadata fields/values and OCR'ed full text

The structure of our afternoon will be as follows, and we hope by the end we will have the following outcomes.

CONTENT:

- Lecture, Discussion
- Hands on trial of using IDL search engine and API

- Practicum of using Collection Builder

OUTCOMES:

- Understand the benefits and challenges of archives as data and how this might inform the way metadata is used for humanities analysis.
- Reflect on ways in which search technologies shape scholarship and research
- General awareness of UCSF digital archives and potential digital humanities applications
- Familiarity with building metadata for digital collections
- Understanding of types of metadata, controlled vocabularies, and general metadata best practices
- Introduction to using Collection Builder to build you own digital collections with home-grown metadata

Digital Archives Ice Breaker Interactive Activity –
5-10 Minutes

Presentation and Using the IDL Search Engine and API (1 hour)

What do digital archivists do and how do we prepare archives to be accessible? Metadata and processing

Archives hold both published and unpublished materials, many of which can be in a wide variety of formats. These may include manuscripts, letters, photographs, moving image and sound materials, artwork, books, diaries, and artifacts. In the current digital era, archives also hold the digital equivalents of traditional archive materials, such as scanned books or paper documents.

Even as reproducible digital objects, digital archives are often still unique, specialized, or rare, meaning access to them in the world may be limited, especially

when copyright or privacy concerns create these limitations. These digital archival collections often still require a researcher to schedule in-person visit.

However there are also a growing number of digitized and born-digital archives, without restrictions, available to anyone to use via the internet as a large document corpus, as databases, etc . These are most often collections that are considered public data.

Some examples here at UCSF of such collections are:

Digitized Japanese Woodblock Collection (Calisphere)

Industry Documents LIbrary

No More Silence dataset
(some digitized, some not)

Industry Archives Multimedia (video/audio) on the Internet Archive'

Digital archives that are publicly available to researchers are served to users as copies of the original digital object.

NEXT SLIDE

Archivists working with digital collections engage in traditional appraisal and description of these materials, but much of our work is extended towards processing for access, **metadata to enhance discovery**, and preservation for continued access.

As an infinitely reproducible object, digital archives are capable of being distributed widely and more openly without compromising the original record. This, paired with the ability to extract the text and content of digital archives through OCR technology, completes the concept of *archives as data*, and can be manipulated and analyzed as such. Once paper-based archival materials, are transformed into a data corpus that can be applied to digital tools for researchers to explore, analyze, and evaluate characteristics of collection materials at a larger scale than observing documents one at a time.

Computational methods can be applied to archives as data to, for example, calculate word frequencies in text or visual characteristics in images, identify place, event, personal, or corporate names, propose common topics within or across textual corpora, or assign sentiments from language used in a subset of documents from a collection. In addition to addressing other analyses, archives as data can support research inquiry that surfaces previously unarticulated relationships between people, institutions, places, ideas, and events that maps any of those across place and time.

Archives as data. as a concept, is a forward looking and positive incentive and rationale for high quality digital archival collections.

NEXT SLIDE

The Significance of Metadata

It is important to note the ways in which digital content from archival collections may need to be prepared or processed before it can be worked with as data. While archives as data allow for useful digital processing and innovative analysis**, they may have been subject to some adjustments in their preparation to become a dataset**. This encompasses the 'digital archives specific' work that archivists uniquely do in order to make a digital archival collection accessible as a dataset. The collections often have a whole other 'life' in the behind-the-scenes hands of the archivists, who must also pay close attention to privacy and safety concerns, ethics, and rights protections.

Archival collections can include a range of different materials (textual, visual, audio/visual) that may be digitized from their original, analog format, or they may be born-digital. For example, in the case of textual material that has been digitized, the images resulting from digitization need to undergo Optical Character Recognition (OCR) to extract text so that it can be analyzed as 'data'. In the case of audio files, these may need to undergo speech-to-text processes to create transcripts for textual analysis.

One other way digital archives, that are intended for the public sphere, are often manipulated or processed is through redaction or PII (personal identifying information) and PHI (personal health information). This aspect of processing does involve a series of pre-determined protocols and decisions by archivists', who

ultimately decide what elements of text might need to be fully removed (and therefore will not be searchable) from a digital object.

The same benefits for having archival documents in digital formats (easily copied bit for bit, text and metadata can be extracted, etc) also play into the challenges of long term access, discoverability, and research analysis. It is crucial to remember the constructed nature of 'data' produced by digitization processes as we grapple with the next challenge: **how to find the needles in this digital haystack of information? Now that everything has been flattened out into a machine-readable series of 1s and 0s how can humans make sense of it?**

NEXT SLIDE

This is where ***metadata*** comes in. Metadata is data *about* data. It is the information attached to data to explain who made it, what format it has been captured in, where and when it was made and other details which the authors or curators of the data want to share. Metadata predates digital data: e.g. a library catalogue records metadata about the material items in the collection. However metadata for digital data becomes even more important than for its analogue counterpart. Have you ever lost a crucial file on your computer because you called it DRAFT_1.doc or something equally non-specific? At least with a book on your bookshelf your memory may be jogged by the color, or the title and author quickly discernable on the spine.

NEXT SLIDE

Like data, metadata gives off an appearance of objectivity, of simply being an empirical recording of facts about the data (and sometimes it is). But often it is as much the result of subjective choices by humans and their variation of motives, beliefs and opinions, and training in different professions and disciplines. Even within the same profession or discipline, there can be wide diversity in standards of metadata meaning that even versions of the same original text could end up being described in different ways.

Crucially, this matters because when we make sense of digital data we are always doing so *through* a machine and not directly with our own senses. Metadata matters because what it records may result in bringing the text into sharper focus, blurring it, or hiding it from view. If readers can only retrieve what is in the catalog,

and can never browse the shelves, then metadata becomes a gatekeeper and not simply a set of observable ‘facts’ about a text.

The Library and Archive fields have worked diligently to mitigate the wild frontier of metadata variability by implementing metadata standards and schemas – which are essentially professionalized handbooks and instructions that members in the field can use as a best practice. The benefits of standardizing how institutions across the world create and input metadata allows for more universal searching across disciplines, or institutions, and minimizes subjective interpretation of field definitions (interoperability).

NEXT SLIDE

There are four kinds of metadata. I’m not going to go deeply into the differences between the four types, however the most common type of metadata you likely come across or think about is probably Descriptive metadata.

Descriptive metadata consist of information about the content and context of your data.

- Examples: title, creator, subject keywords, and description (abstract)

Structural metadata describe the physical structure of compound data.

- Examples: camera used, aperture, exposure, file format, and relation to other data or files

Administrative metadata are information used to manage your data.

- Examples: when and how they were created, who can access them, software required to use them, and copyright permissions

Preservation metadata is structured information that helps preserve and manage digital resources over time. It's a crucial part of digital preservation strategies and is used to ensure that digital information is accessible and long-lasting.

- Examples: original analog format, file format, provenance, fixity values, access rights, audit trail, encoding software, resolution, etc.

NEXT SLIDE

Metadata schema – Metadata schema outline the overall structure for the metadata. A metadata scheme describes how the metadata is set up, and usually addresses standards for common components like dates, names, and places. There are also discipline-specific schemas used to address special elements needed by a given discipline.

Because of time, I'm not going to go through the technical specifics of a metadata schema, but here are some examples of common ones (although there are so many more to speak of also) and I encourage you to go online and read through the documentation for a metadata schema to get a sense of how they work.

Examples of commonly used metadata schemas:

- Dublin core
- MODS (Metadata Object Description Schema)
- OLAC (Open Language Archives Community)
- VRA Core (Visual Resources Association)
- TEI (Text Encoding Initiative)
- PREMIS (Preservation Metadata Implementation Strategies)

NEXT SLIDE

Controlled vocabulary - p**rovides a consistent way to describe data**. Examples of controlled vocabularies include subject headings, thesauri, ontologies, and taxonomies. Using a controlled vocabulary will improve your data's findability and will make your data more shareable with researchers in the same discipline.

Examples of controlled vocabularies:

General Purpose

- Library of Congress Authorities for subject headings
- United Nations Educational, Scientific and Cultural Organization (UNESCO) thesauri

Arts & Humanities

- Art & Architecture Thesaurus
- Rare Books and Special Collections Vocabulary
- Thesaurus of Musical Instruments

Health Sciences & Medicine

- International Classification of Disease (ICD)
- Logical Observation Identifiers Names, and Codes (LOINC) (for tests, observations and measurements)
- Medical Subject Headings (MeSH)
- Rx Norm (for drugs)
- SNOMED-CT (for clinical terms)

Sciences

- Astronomy Thesaurus
- NASA Thesaurus
- National Agricultural Library Thesaurus
- The Getty Institute Thesaurus of Geographical Names

Social Sciences

- Ethnographic Thesaurus
- Thesaurus for Economics

NEXT SLIDE

Streamlining the format and manner in which metadata exists in order to establish provenance, reliability, and traceability of content even dates back to the medieval tradition of Hermeneutics – the O.G. of metadata standards. The *accessus ad auctores* functioned as a kind of basic metadata standard for medieval religious scholars, but was also used for medical and philosophical texts.

Find any item online (could be a webpage, a YouTube video, a digital book, or a digital paper).

Can you answer these questions?

1. Who is the author (*quis/persona*)?
2. What is the subject matter of the text (*quid/materia*)?
3. Why was the text written (*cur/causa*)?
4. How was the text composed (*quomodo/modus*)?
5. When was the text written or published? (*quando/tempus*)?
6. Where was the text written or published (*ubi/loco*)?
7. By which means was the text written or published (*quibus faculatibus/facultas*)?

NEXT SLIDE

Places Where We Engage with Metadata for Research

We will briefly touch on the common ways we engage in metadata for research, but we do not need to spend a lot of time on this because for most of us, we are familiar with these tools.

The abundance of digital information has led to the development of an array of new search and discovery technologies. One of the most powerful of these is the **'search engine'**, an information retrieval system allowing users to query a collection of digital data and receive 'answers'. We're familiar with the search engines which allow us to retrieve information from the worldwide web, but they are also a ubiquitous part of our interactions with digital data in many other places, including our computers and mobile phones.

Search engines contain two key elements: an **index** and a **query** system. The index has much in common with a traditional library catalogue or the index in a book: it provides a list of items and their locations (which folder the file sits in for a database, or their url (web address)). The query system matches human questions to items in the index. You could think of it as an automated librarian or archivist, and the success of the most powerful search engines relies heavily on the query system appearing to know what users want, almost before they articulate it themselves.

Then there are web search engines that specifically operate using programs called crawlers, which retrieve information form items published online. Crawlers create the index which users query the search engine.

Metadata lies at the heart of search engines and building the index. It's a shortcut to the contents which allows the search engine to present likely matches to the users' queries in the split-second timings that we are used to.

We must keep in mind that web searching, while it appears magically fast and accurate, is often based on filters and ranking criteria that attempts to best-match your search to results. As such, there is potentially some inherent bias involved in the algorithms that assist our searches. In order to **search critically**, it is

important to unlearn what commercial web search has taught us. That means asking questions about the metadata, the index, and query system, and trying to understand how those tools work in order to make them work for you. This helps researchers work out what the search engine might not be showing us, and what the sources of its biases might be.

This is not only essential when using web search engines, but also using digital search on the past – such as with digitized collections in historical and cultural archives.

Another example of how search engines shape how we see historical document collections is connected to the difficulties encountered by Optical Character Recognition (OCR) tools, which turn historical typefaces or hand-written text into machine-readable text. The variability in effectiveness and success in OCR tools makes the contents of these documents potentially more opaque to machines than modern documents.

Similarly, if metadata on digital or digitized historical documents is incomplete, includes terms that are too niche, or has poor OCR text, the use of search engines for research becomes limited in capacity.

We have outlined the ideals about what good metadata can do for archives and research. However, in practice, our field confronts a long legacy of 'the time before best practices' and also inherited metadata. This is especially challenging when paired with the deluge of volume of archival materials that are born-digital. We will discuss how this has been a challenge in real life for the Industry Documents Library at UCSF.

NEXT SLIDE

Overview of Industry Documents Library.

Brief Overview of the History of the IDL

- IDL is a digital archive of 15 million documents created by industries which influence public health, hosted by the University of California, San Francisco Library.

- Originally established in 2002 to house millions of documents publicly disclosed in litigation against the tobacco industry in the 1990s, IDL has expanded to include documents from the drug, chemical, food, and fossil fuel industries to preserve open access to this information and to support research on the commercial determinants of public health.
-
- The documents have been cited in more than 1,000 publications including peer-reviewed articles, books, government reports, investigative journalism pieces, and documentaries.

Preview of the website (Rachel to screen share IDL website, no slides)

- Object level description (inherited metadata)
    - Both structured data and unstructured data (OCR text) available
- Boolean full text search
- Collection level descriptions
- Bibliography
- Dataset and API (examples of API query with resulting CSV file)
    - Datasets include the structured and unstructured metadata
    - API only queries the sidecar metadata (structured)

Digital Humanities Research at the IDL

- Risi, S. & Proctor, R.N. (2019). Big tobacco focuses on the facts to hide the truth: an algorithmic exploration of courtroom tropes and taboos. *Tobacco Control.* http://dx.doi.org/10.1136/tobaccocontrol-2019-054953.

Used methods from computational linguistics to identify differences in the rhetorical strategies deployed by defense versus plaintiffs' lawyers in cigarette litigation

- Tobacco Analytics, digital humanities website to display visualizations from this research (site no longer active!). Grant had ended, funding to support site ended. Web archiving challenges.

- Hu, B., Rakthanmanon, T., Campana, B.J.L. *et al.* (2013). Establishing the provenance of historical manuscripts with a novel distance measure. *Pattern Analysis and Applications* 18**,** 313–331. https://doi.org/10.1007/s10044-013-0332-z

Applied graphic element detection and recognition to identify tobacco company logos and search for all documents with matching logos, to test image data mining techniques

Potential Digital Humanities Applications:

- Create network maps to identify and assess relationships between companies, PR firms, third-party organizations, scientists, and regulatory officials to visualize influence on research and public health policy

- Apply sentiment analysis and term frequency analysis to study how smoking-related diseases have been described in industry documents over time

- Assess industry marketing strategies targeting specific populations and how they impact health disparities

Have everyone try using the IDL API themselves.

In Conclusion - Metadata Best Practices

Good data documentation is essential in order to find, use, share, and manage data, especially over time. This includes information such as:

- The context of data (why and how data was collected)
- The structure of data (including how multiple files relate to each other)
- Quality assurance that data is complete and uncorrupted
- Information on data confidentiality, access and use conditions (where applicable)
- Identification and tracking of different versions of datasets

QUESTIONS

MATERIALS AND RESOURCES?

BREAK (15 minutes)

PART 2:

Introduction to Collection Builder (1.25 hours)
Objectives:

- **Building your own archive collected research materials**
- **Using Collection Builder as a tool to describe, organize, and view those collections, perhaps for outward viewing**

Depositing your dataset in a data repository can:

- Fulfill funder requirements for open and/or preserved data
- Share research with a larger and often disciplinary audience
- Improve online discovery
- Increase citation rates
- Preserve data for the long-term

Data repositories may require specific metadata standards and formats in order to deposit data. It is important to determine metadata standards used by a repository before beginning the metadata plan for your data.

Each person would have a GitHub account. Each person will build their own Collection Builder GH website.

5-10 objects for them to sit through and think through to do some cataloging

- No more silence
- Helen Fahl
- IDL small collection
- Japanese Woodblock prints (Lisa will add)

Three groups of documents (one for each group of people)

Build a basic database of metadata, through provided metadata and manual description

Compare how data across each member in your group may differ; if there are differences, decide on how to standardize amongst each other.

Translate initial dataset into the Collection Builder metadata template.

Follow steps to build you own visual dataset in Collection Builder.