

Data Architecture and Interoperability

Sara Morrell, Kasya O'Connor Grant, and Dipa Desai

HIST 7251: The Digital Archive

Professor Jessica Parr

Spring 2024, January 25th



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Workshop Agenda/Objectives

- Identify similar data structures across databases
- Discuss ethics of digital data collection, management, and archiving
- Explore different querying languages, interfaces, and metadata standards

Class materials available at:

<https://bit.ly/sp24-parr-hist7251-data>



Opening Activity: Database Poll



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Opening Activity: Databases

Regardless of discipline, databases are an essential component of scholarly research and organization management and access.

In your research and classes, what databases have you used?
Which are you most familiar with?

Answer this by accessing this poll:

<https://bit.ly/HIST7251-sp24-database-poll>



Database Poll: Results

Let's take a look at the results for the poll:

<https://bit.ly/HIST7251-sp24-database-results>



Opening Concepts

What are some features of databases that are similar regardless of content? What are the record formats (item, artifact, etc.) and interface layouts you are familiar with?

- **Record:** group of related data held within the same data structure, or an object that contains more than one value.
- **Metadata:** set of data (fields) that describes and gives information about other data.
- **Query:** a request for data or information from a database
 - Action query: Perform an action on the data (delete, add, change)
 - Select query: Retrieve data
- **Interface (GUI, API, or SQL):** mechanism that allows two systems to meet and interact or where users' queries interact with the database.



What is Data Interoperability?



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

What is Data Interoperability?

Data Interoperability refers to the way in which data is formatted to allow for diverse datasets to be merged, aggregated, or accessed across platforms.

- Data interoperability is dependent upon **data standards**
- Designing for data interoperability means thinking intentionally about the relationship of metadata and data structure to increase discoverability and parsability



Metadata: Standards + Data Mapping



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Metadata

- What is **metadata**? Information on the dataset—who, what, where, when, how, why, etc
- Metadata standards help to implement consistent naming conventions so data is more compatible with machine readability: using multiple words and special characters in a name can sometimes prevent a computer from reading it, therefore, names often use one word or connect words with underscores.
- README files outline the metadata and make it easier for other people to understand and apply to the dataset
- Check out [NU Library's Guide for Data Management](#)



Metadata Standards

- There are different standards for data management and metadata that allow data to be *interoperable*.
- The ability to convert different types of data to formats that can be read by different users and interfaces facilitates greater access and use.
- Metadata can make missing information visible, and create opportunities for us to make data more inclusive.
- Examples of disciplinary metadata standards:
 - [Darwin Core \(DwC\)](#)
 - [NeXus](#)
 - [Data Documentation Initiative \(DDI\)](#)



Project Metadata Documentation

One way to understand more about a project or database is to explore any available documentation, including **metadata application profile**, **taxonomies**, and **ontologies**.

Metadata Application Profile: a document identifying (often with examples) the metadata used by a domain, project, or application and how it is used.

Taxonomies: a formal structure of classes or types of objects within a domain.

Ontologies: a subset of taxonomies with information about behavior of entities and relationships between them.

Example: [Metadata Application Profile for the Digital Transgender Archive](#)



Controlled Vocabularies

Controlled vocabularies, a way to standardize input of categories: choose from a list prepared in advance.

example_dataset

File

Edit

View

Insert

Format

Data

Tools

Add-ons

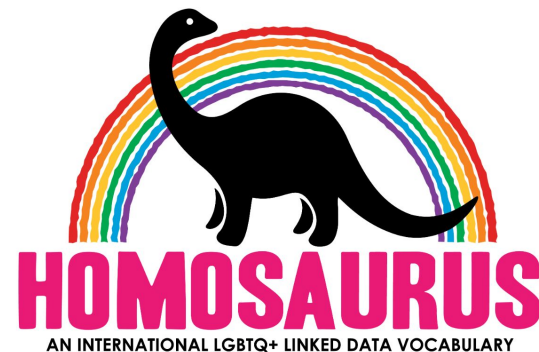
Help

Last edit was 3 minutes ago

Mapping Database Interoperability

Data mapping matches fields of information from one database to another. Data from different sources can describe similar data points with different descriptors. Ex: A database based in Europe may write dates as *day/month/year*, where a US database may write dates as *month/day/year*.

- Data mapping allows databases to be transferable and comparable; it also facilitates analysis. It can enhance the accessibility and discoverability of archived data.

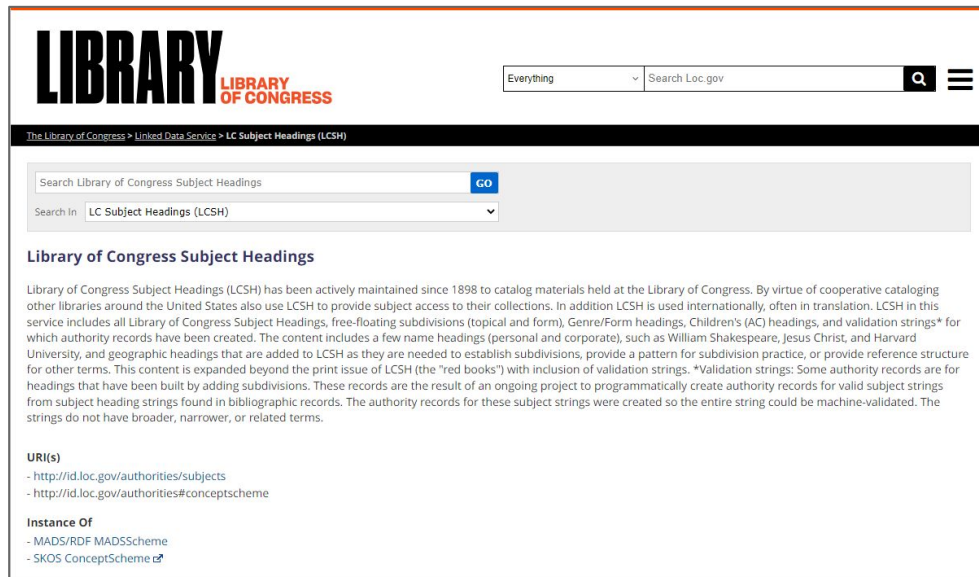


The [Homosaurus](#) vocabulary allows institutions to make LGBTQ+ resources more accessible, supplementing existing vocabularies like the LCSH (Library of Congress Subject Headings).



Library of Congress Subject Headings (LCSH)

The Library of Congress Subject Headings is a dataset of linked subject headings used to describe items in archives, libraries, and other cultural heritage institutions.




The screenshot shows the Library of Congress Subject Headings (LCSH) website. At the top, there is a header with the "LIBRARY OF CONGRESS" logo and a search bar. Below the header, there is a navigation bar with the text "The Library of Congress > Linked Data Service > LC Subject Headings (LCSH)". The main content area features a search bar with the text "Search Library of Congress Subject Headings" and a "GO" button. Below the search bar, there is a dropdown menu with the text "Search In LC Subject Headings (LCSH)". The page title is "Library of Congress Subject Headings". The main text describes the LCSH dataset, stating it has been actively maintained since 1898 to catalog materials held at the Library of Congress. It mentions that LCSH is used internationally, often in translation, and includes all Library of Congress Subject Headings, free-floating subdivisions (topical and form), Genre/Form headings, Children's (AC) headings, and validation strings* for which authority records have been created. The content includes a few name headings (personal and corporate), such as William Shakespeare, Jesus Christ, and Harvard University, and geographic headings that are added to LCSH as they are needed to establish subdivisions, provide a pattern for subdivision practice, or provide reference structure for other terms. This content is expanded beyond the print issue of LCSH (the "red books") with inclusion of validation strings. *Validation strings: Some authority records are for headings that have been built by adding subdivisions. These records are the result of an ongoing project to programmatically create authority records for valid subject strings from subject heading strings found in bibliographic records. The authority records for these subject strings were created so the entire string could be machine-validated. The strings do not have broader, narrower, or related terms.

URI(s)

- <http://id.loc.gov/authorities/subjects>
- <http://id.loc.gov/authorities#conceptscheme>

Instance Of

- MADS/RDF MADSscheme
- SKOS ConceptScheme 



Activity: Subject Heading Comparison

To explore issues of ethics in cataloging and describing items, we will use the Homosaurus and LCSH to search for LGBTQ+ terms.

Homosaurus: [Link to Homosaurus Vocabulary Terms](#)

LCSH: [Link to Library of Congress Subject Headings](#)

Some suggested terms to search:

- Gender Dysphoria or Gender identity
- Gay liberation, Queer liberation, or trans liberation
- LGBTQ community
- LGBTQ people

Questions for Discussion:

1. What differences did you notice between the two databases?
2. What observations do you have about the way that identity is organized in the LCSH? Were there any words you noticed in particular?
3. How is identity organized in the Homosaurus project? How is this different to the LCSH?
4. What uses could you see the Homosaurus having for people cataloging or describing items?



Activity: Database Querying



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

What is a Query?

- **Search:** A non-specific search for data across data models that broadly match the key term(s) based on the search engine algorithm
 - Because algorithms are programmed by humans, they capture human biases. Consider what may be missing or misrepresented in the search output.
- **Query:** A specific request to access and retrieve data from a database
- Identifying specific data can be further narrowed by querying certain fields of information and using operators
 - Ex. Querying the library catalog by Author **AND** Title



How to use querying for research?

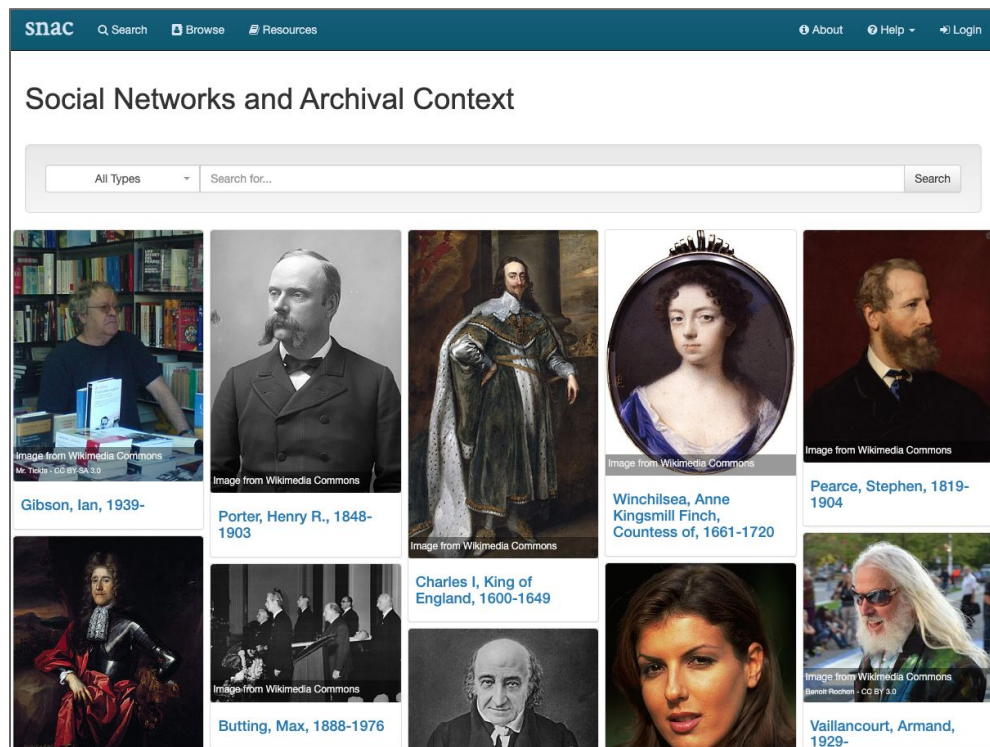
Querying is used in research to find appropriate datasets, filter subsets of information, and search across databases to get a comprehensive view of the research topic and select data that help answer your research questions.

Questions to consider prior to beginning research:

- What do you want to know? What terms best describe this information?
- What information is available, missing, accessible to you, etc.?
- How should queries be structured for the database(s) you are using?
- What tools/features exist on the database to aid iterative querying?
- How will you keep track of queries and data results?



Database: SNAC



SNAC: Social Networks and Archival Contexts is a free, online resource with biographical and historical information about persons, families, or organizations.

What do you notice about the homepage?



Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Database Activity: SNAC Queries

With a partner, open up the following database on your computers:

SNAC Database

Type in a few queries to get a sense of the data. For example, search historical names related to Boston: Crispus Attucks, Paul Revere, etc.

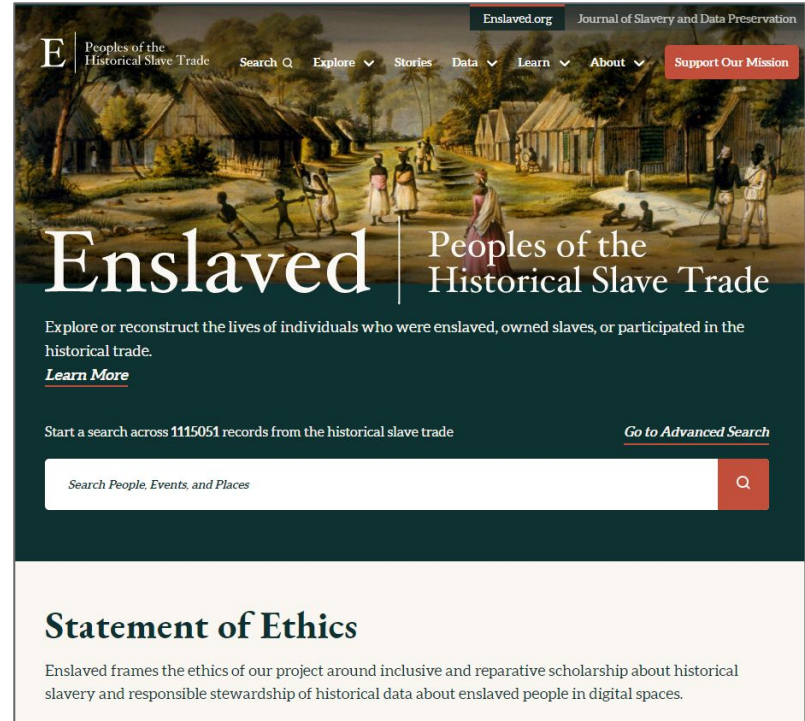
As you are searching, consider these questions:

- How is this data structured?
- How would you described this data?
- How is the data described in the database?
- Are there features to allow for easier searching?



Database: Enslaved

Enslaved: Peoples of the Historical Slave Trade is a database that pulls from datasets and databases to make searchable individuals related to the slave trade (both those who were enslaved or owned slaves).



Database Activity: Enslaved Project

Now, take a moment and search for the same people in another database:

<https://enslaved.org/>

The **Enslaved: Peoples of the Historical Slave Trade** project utilizes linked data and different datasets to create a dataset from multiple sources. More information about their data can be found here: [Enslaved Peoples Project Database](#)

Data Documentation: [Enslaved: Peoples Project Documentation](#)



Metadata Documentation

Enslaved Project

Controlled Vocabulary: a complete list of the project's controlled vocabulary created for records to help with searching, visualizing, and organizing data.



Enslaved | Peoples of the Historical Slave Trade

Controlled Vocabularies

Terms and Definitions

Version 3

March 23, 2023

Revised by Dean Rehberger, Walter Hawthorne, Daryle Williams, Catherine Foley, Alicia Sheill, Sharon Leon, Steven Niven, Kristina Poznan, and Marisol Fila

Thank you to individuals who contributed to authoring various versions of Enslaved.org's Controlled Vocabularies: Heather Bollinger, Ryan Carty, Luisa Cruz, David Eltis, Ina Fandrich, Jessica Fletcher, Henry Louis Gates, Jr., Daniel Jenkins, Seila Gonzalez Estrecha, Gwendolyn Midlo Hall, Kathe Hambrick, Paul LaChance, Jane Landers, Paul Lovejoy, Henry Lovejoy, Keith McClelland, Érika Melek Delgado, Jeff Mixter, Jim Schindling, Kara Schultz, Angela Sutton, Duncan Tarr, Bruna Tine, and Ethan Watrall.

*Feel free to ask questions at any point
during the presentation!*

Who and what is data created for?

What are some political issues when it comes to data and data creation?

- Open vs. closed data and accessibility
- **Ownership:** who owns your data or data about different individuals and how does this reflect political intentions?
- **Data Commodification:** the gathering and selling of data on target audiences for advertising, marketing, etc.
 - Potentials for abuse, lack of accountability, power differentials, etc.
- **Data Privacy:** the ways in which companies are (or are not) protecting customer or user data from outside sources, or the use of identifiable information in datasets



Data Concerns: *What gets counted counts*

D'Ignazio and Klein identify problematic data practises that cause harm:

- Lack of quantitative research on maternal mortality masks systemic problems.
- Undocumented immigrants are often (sometimes voluntarily) absent from census data, which determines levels of federal funding: a “paradox of exposure.”
- TSA scanning machines binarize bodies to attempt to uncover concealments, but can thereby mistakenly assign risk alerts.

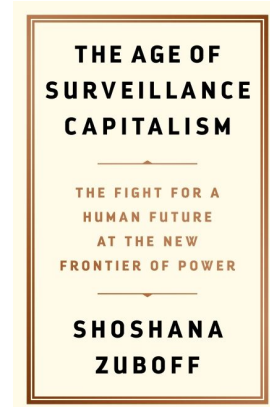
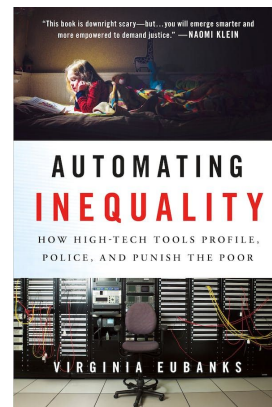
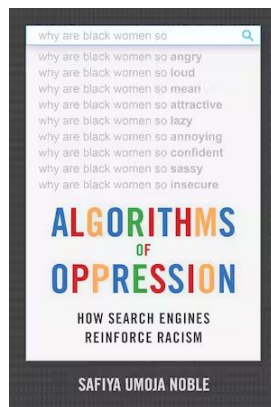
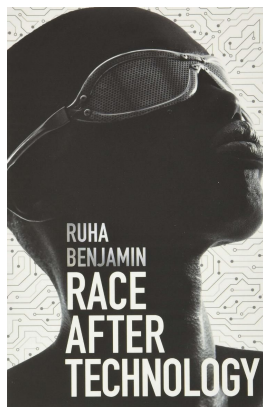
“What is counted—like being a man or a woman—often becomes the basis for policymaking and resource allocation. By contrast, what is not counted—like being nonbinary—becomes invisible...”

Catherine D'Ignazio & Lauren Klein, [*Data Feminism*](#), 2020



Critical Data Studies

Critical Data Studies: an emerging interdisciplinary field that addresses the ethical, legal, cultural, social, epistemological, and political aspects of data science, big data, and digital infrastructures.



Database Architecture + Accessing Data



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*


Different Querying Structures

Queries interact with database structures through querying languages, and can be formatted to yield a specific data output. The query structure depends on the query language syntax, and how it interacts with the database management system's interface.

- **GUI: a graphical user interface** that allows point-and-click interactions between a human user and a digital database
- **SQL: structured query language** used with the command line interface by a human user to access a database
- **API: application programming interface** that allows software to access a database



Database GUI Examples




Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more.

Icon	Value
Computer monitor	78TB
Open book	37M
Film strip	9.2M
Speaker	15M
Television	2.5M
CD/DVD	959K
Image	4.6M
Microphone	246K
Calendar	1.6M

Search

[Advanced Search](#)



HATHI TRUST
Digital Library


Search the HathiTrust Digital Library

Search words about or within the items

☒ Full-text ☐ Catalog


[Advanced full-text search](#) [Advanced catalog search](#) [Search tips](#)

[Should I search catalog or full-text?](#)



DRS Digital Repository Service

Northeastern University / University Library / Library Departments / **Archives and Special Collections**

Archives and Special Collections  *Community*

The Archives and Special Collections Department is part of the Northeastern University Library. Its goal is to document the teaching, research, community service, and administrative functions of the University and to document student life. This goal is accomplished by collecting the historically significant records of the University. The Department also preserves and makes available the records of private, non-profit, community-based organizations that document diverse and under-documented populations.



Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

SQL Basics: TL;DR

SQL: structured query language used to store, manipulate, access, and process information in relational databases.

SQL looks similar to and integrates well with other programming languages like Java.

Relational databases use stored SQL statements as instructions to maintain and manage the database. You can use SQL statements to query relational databases!

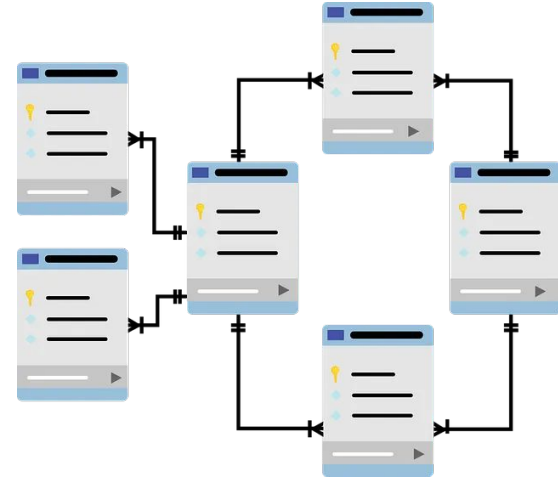


Image Credit: User mcmurryjulie from Pixabay.com

Relational databases will have a data model that shows the interactions and connections among different data within the database.



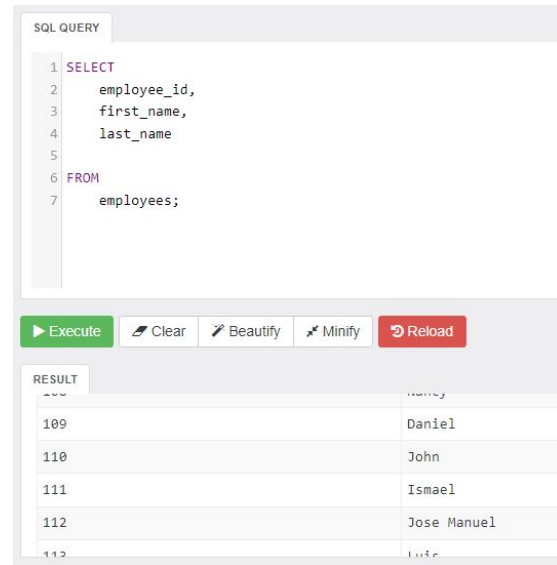
SQL Database Example + Tutorial

Go to: <https://www.sqltutorial.org/> > **Section 2: Querying**

Read over how to use SELECT to retrieve data from the example database. Go to the “TRY IT” tab in the upper right of the menu. Try writing a query. Click the ‘Execute’ button to execute the query. Edit the query to get different outputs.

What do you notice about how the data is structured and recorded? What looks familiar?

Check out the [SQL tutorial sample database model](#) to see the relationships among data.



The screenshot shows a web interface for SQL queries. At the top, there's a tab labeled 'SQL QUERY'. Below it, a text area contains the following SQL code:

```
1 SELECT
2   employee_id,
3   first_name,
4   last_name
5
6 FROM
7   employees;
```

Below the text area are five buttons: 'Execute' (green), 'Clear' (light blue), 'Beautify' (light blue), 'Minify' (light blue), and 'Reload' (red). Below the buttons is a tab labeled 'RESULT'. Under the 'RESULT' tab, there is a table with two columns. The first column contains employee IDs, and the second column contains their names. The data shown is as follows:

employee_id	first_name	last_name
109	Daniel	...
110	John	...
111	Ismael	...
112	Jose Manuel	...
113



APIs + Web Scraping

An **API**, or application programming interface, is a set of subroutine definitions, communication protocols, and tools for building software that ultimately allows applications to communicate with one another.

- An API may be for a web-based system, operating system, database system, computer hardware, or software library.

Web-scraping is the process of extracting large amounts of data from an internet source and downloading the data to a local repository.

- The scraping process can be done manually, but is usually automated through software because of the large amount of data typically involved.



API Documentation

- When using APIs, it is necessary to refer to the API documentation—a link is usually found on the API homepage.
- Why?
 - While the concepts remain roughly the same, APIs differ and the syntax for accessing data can be very different.
 - You will likely need an API key, and the links for registering for the key will be found in the documentation.
 - There may be other differences and specifics that require a close understanding of the API's structure.



Popular APIs

- New York Times: <https://developer.nytimes.com/>
- Reddit: <https://www.reddit.com/dev/api/>
- OMDb: <http://www.omdbapi.com/>
- FBI:
<https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/docApi>
 - Other Federal government APIs:
<https://api.data.gov/docs/developer-manual/>



OMDb: API Output

```
{"Title":"Breakfast at  
Tiffany's","Year":"1961","Rated":"Approved","Relea  
sed":"06 Oct 1961","Runtime":"115  
min","Genre":"Comedy, Drama,  
Romance","Director":"Blake  
Edwards","Writer":"Truman Capote, George  
Axelrod","Actors":"Audrey Hepburn, George  
Peppard, Patricia Neal"}
```

When using the OMDb API, you can access dictionaries with data for each movie.

This is a portion of the output when querying for "Breakfast at Tiffany's".

What looks familiar about how the data is structured?

Dictionaries are data containers. They describe the relationship between data elements.



NYT: Article Search API Output

```
{"abstract":"The public health researcher Abigail Echo-Hawk is a leading voice in a movement to empower Indigenous people, wielding data as a tool for racial equity.","web_url":"https://www.nytimes.com/2023/12/12/health/indigenous-data-abigail-echo-hawk.html","lead_paragraph":"“Transforming Spaces” is a series about women driving change in sometimes unexpected places.","source":"The New York Times"}
```

Using the NYT Article Search API, you can retrieve data on articles pertaining to a specific topic.

This is a sample of fields from output when querying for articles pertaining to data

What looks familiar about how the data is structured?



NYT: Article Search API Query Example

Click to see an example of the full API query results:

<https://api.nytimes.com/svc/search/v2/articlesearch.json?q=data&api-key=jbht5SsPFHcWwIUgIGrk80NSKAvfN05X>

Change the word after q= to search for a different keyword.



Ethical Considerations

Contextual Privacy

- Consider context when working with online data. What someone might be comfortable saying in one context might not be something they're okay saying to a researcher.

Keeping People Safe

- It is risky to publicize the username, profile picture, or exact text of a social media post or profile.
- To show example posts etc., you can make up your own or heavily redact them.
- Please be mindful of obtaining consent if you are scraping individual info.



Databases + API: More Information

If you are interested in learning more about database architecture (including SQL) and APIs, see the following resources:

- [SQL Tutorial](#)
- [Introduction to MySQL with R \(Programming Historian\)](#)
- [Guide for Using NYT API](#)
- [Web Scraping Tutorial \(with Jupyter Notebook\)](#)



HIST 7251: Metadata DITI Session Preview



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Dublin Core

- [Dublin Core is a set of metadata standards](#) (fun fact: it was named after Dublin, Ohio, not Dublin, Ireland!)
- It was created to reject siloed cultural memory and enable radically open cultural heritage data
- It is designed to be simple and flexible, which has both advantages and disadvantages
- Contains 15 ‘core’ metadata elements (i.e., Title, Date, Subject, etc.) and additional ‘qualified’ elements to give metadata greater specificity.



Describing Archives: A Content Standard (DACS)

Describing Archives: A Content Standard (DACS) is a metadata standard for digital cultural heritage projects and archives, containing information about necessary metadata fields and implementation.

The complete standards are hosted on GitHub: [Link to Describing Archives: A Content Standard \(DACS\)](#)



Text Encoding Initiative (TEI)

The **Text Encoding Initiative (TEI)** is a consortium that collectively develops and maintains a standard for the representation of texts in digital formats.

TEI Guidelines (P5) can be found here: [Link to TEI Guidelines](#)



Data Formats + Preservation



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Class Discussion: File Formats

What are the most common file, data, or document format that you work with?

What is the most difficult one?

What is your experience with data conversion? (Word or Google Docs to PDF, XLSX to CSV, TIFF to JPG, etc.)

Why do you have to convert data formats? When is it necessary?



File Formats

File formats are specific to the type of data being recorded and stored. Some file formats can easily support multiple data types and be converted to different formats, while others are less easy to convert. In addition, certain formats are intended for long-term storage.

- Directionality of file conversion and longevity of the format
- OCR, plain text, pdf
- Export formats

Sustainability of Digital Formats: Planning for Library of Congress Collections		
Introduction	Sustainability Factors	Content Categories Format Descriptions Contact
Format Descriptions >> Format Description Categories >> Browse Alphabetical List >> Format Descriptions as XML		
Format Descriptions		
Still Image <ul style="list-style-type: none">• SVG_1_1• TIFF_6• All still image format descriptions	Sound <ul style="list-style-type: none">• WAVE• MP3_FF• All sound format descriptions	Moving Image <ul style="list-style-type: none">• MPEG-4_FF_2• AVI• All moving image format descriptions
Textual <ul style="list-style-type: none">• PDF/A family• DOCX/OOXML_2012• All text format descriptions	Web Archive <ul style="list-style-type: none">• ARC_IA• WARC• All Web archive format descriptions	Datasets <ul style="list-style-type: none">• DBF• HDF5• All dataset format descriptions
Geospatial <ul style="list-style-type: none">• ESRI shape• GeoPackage_1_0• All geospatial format descriptions	Generic <ul style="list-style-type: none">• ASE• RIFF• All generic format descriptions	



Digital Preservation

Data is often created to fit a certain format, to be used with a certain technology. The ability to convert data to different formats allows it to be accessible and usable over time.

Because data formats and types are specific to a certain technology, data can also be used for digital archaeology, tracking the evolution of data.

[Link to Duke University Library Digital Preservation Guide](#)



Questions?



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Thank you!

If you have any questions, contact DITI at nulab.info@gmail.com

Developed by Juniper Johnson, Dipa Desai, Sara Morrell, and Kasya O'Connor Grant

Digital Integration Teaching Initiative Research Fellows

Slides, handouts, and data available at <https://bit.ly/sp24-parr-hist7251-data>

We'd love your feedback! Please fill out a short survey here:

<https://bit.ly/diti-feedback>

Schedule an appointment with us! <https://bit.ly/diti-meeting>

