

# Data Architecture and Interoperability

---

Juniper Johnson and Benjamin Gray  
INSH 2102: Bostonography  
Prof. Parr and Prof. Nelson  
Spring 2023



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
- Consider how data formats impact digital preservation

Class materials available at:

<https://bit.ly/sp23-parr-insh2102-data>



# 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 across disciplines, what databases have you used? Which are you most familiar with?

Answer this by accessing this poll:

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



# Opening Discussion

What are some features of databases that are similar regardless of content? What are formats are you familiar with?

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



# Data Interoperability + Accessing Data



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 in ways across platforms.

- Data interoperability is dependent upon **data standards**
- Focuses on the relationship of metadata + data structure for discoverability and parsability



# 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 name





# Different Querying Structures

**Queries** interact with database structures through querying languages, and can be formatted to yield a high-resolution data output. The formatting 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



# 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 question.

## 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?



# APIs + Web Scrapping

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 by using software because of the large amount of data typically involved.



# 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.



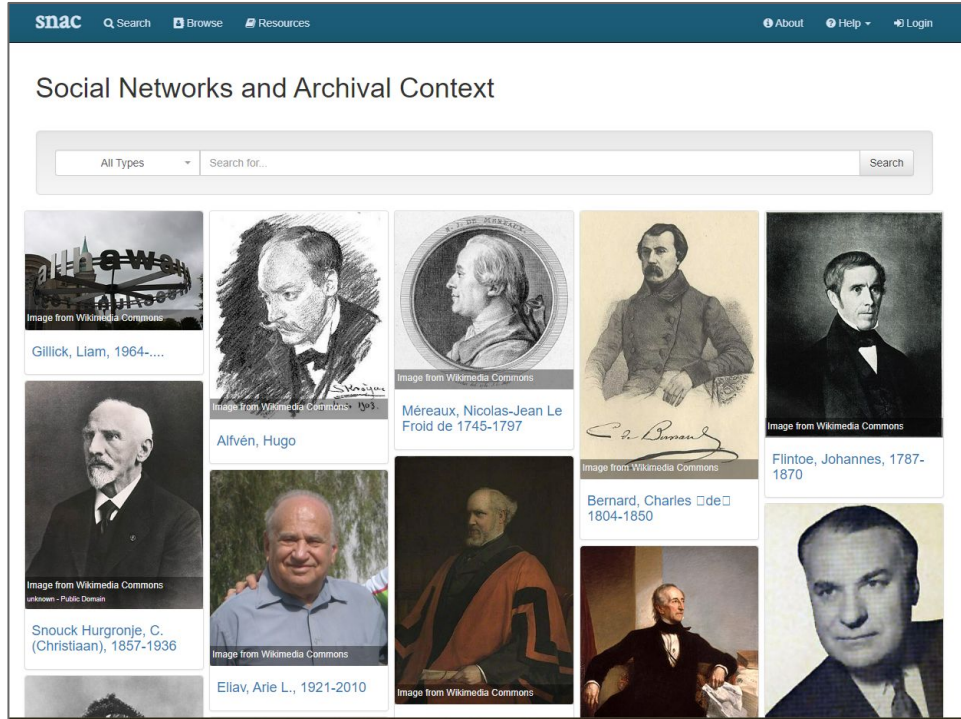
# Activity: Database Querying



Northeastern University  
*NULab for Texts, Maps, and Networks*

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

# 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?



# Database Activity: SNAC Queries

With a partner next to you, open up each database on your computers:

<https://snaccooperative.org/>

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.

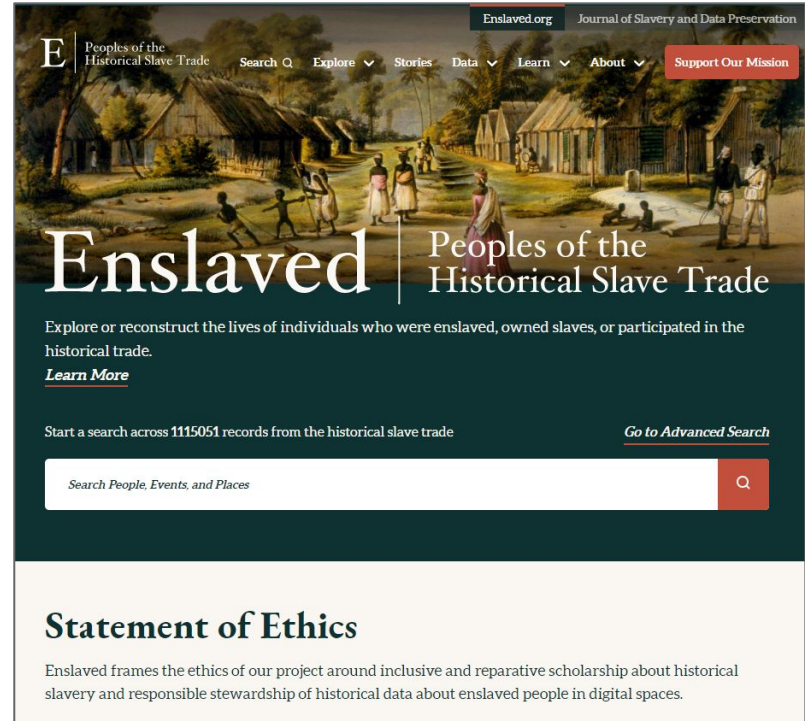
As you are searching, consider these questions:

- 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:

<https://enslaved.org/data/>

Data Documentation: <https://docs.enslaved.org/index.html>



# 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 compares 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.

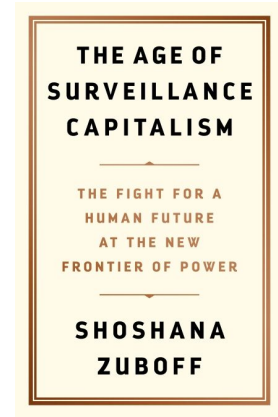
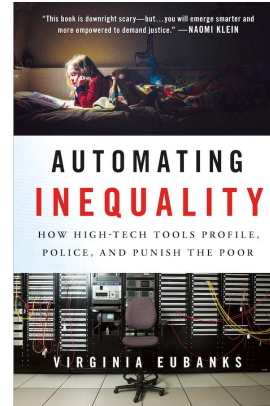
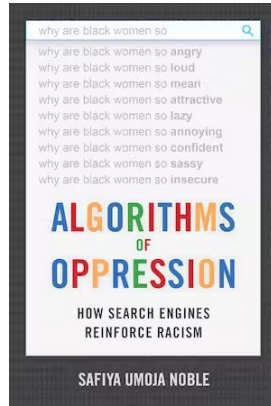
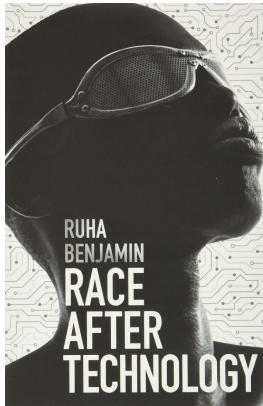
**“What is counted—like being a man or a woman—often becomes the basis for policymaking and resource allocation. By contrast, what is not counted—like being nonbinary—becomes invisible...”**

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.



# 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, and why
- Developing column naming conventions to be more compatible with machine readability: lower cases and underscores, rather than multiple words and characters that often have specific functions within programs (e.g. comma)
- README files outline the metadata and make it easier for other people to understand and apply the dataset
- Check out [NU Library's Guide for Data Management](#)



# Metadata Standards

- There are different standards for data management and metadata depending on your area of research 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.

**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.

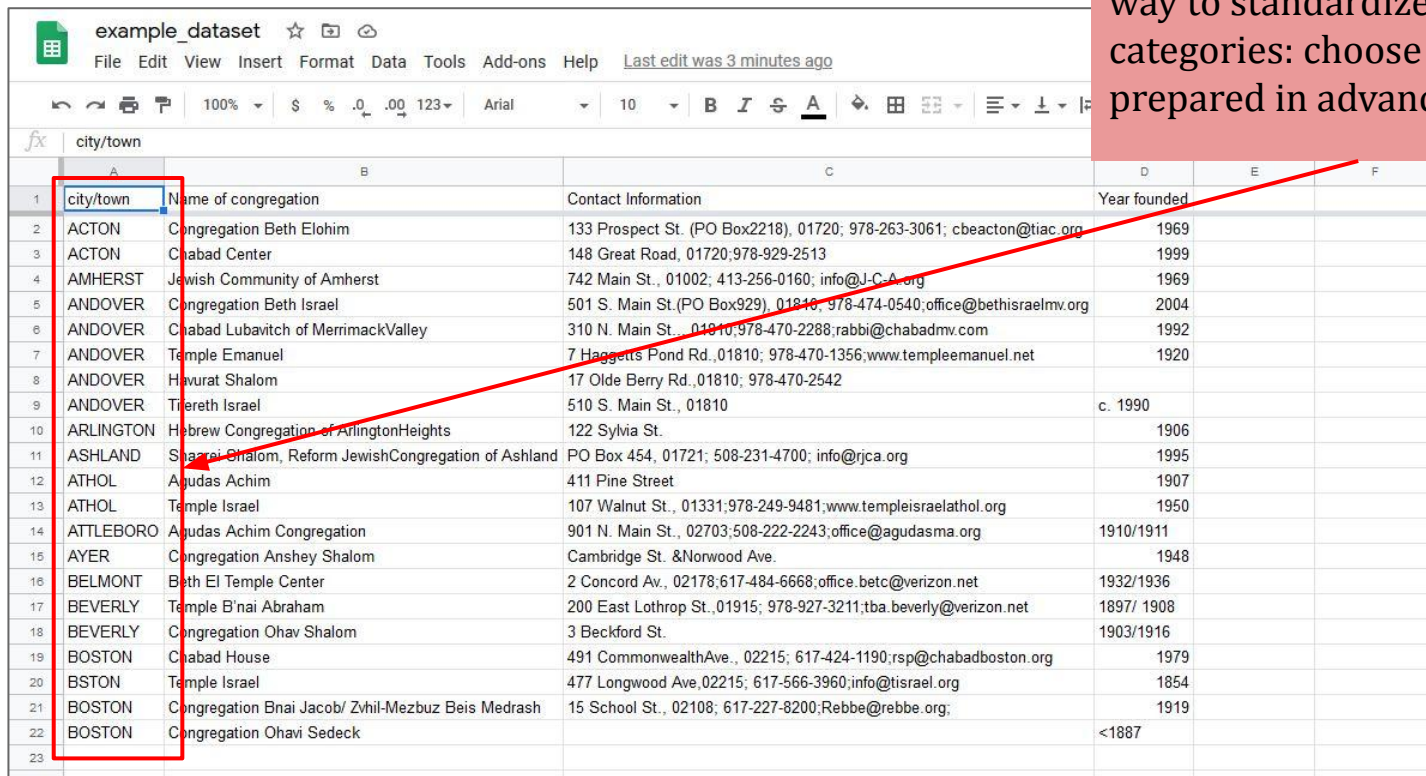
**Example:** [MAP for the Digital Transgender Archive](#)





# Controlled Vocabularies

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



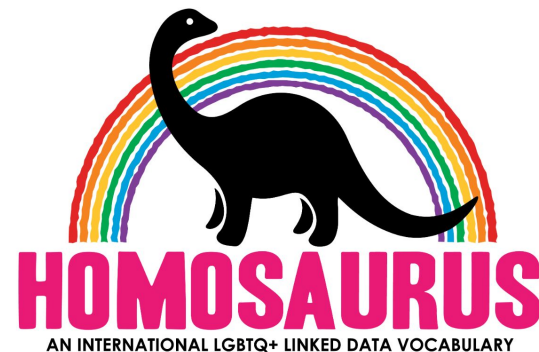
example_dataset					
File Edit View Insert Format Data Tools Add-ons Help Last edit was 3 minutes ago					
100% \$ % .0 .00 123 Arial 10 B I A					
city/town					
	A	B	C	D	E
1	city/town	Name of congregation	Contact Information	Year founded	
2	ACTON	Congregation Beth Elohim	133 Prospect St. (PO Box 2218), 01720; 978-263-3061; cbeacton@tiac.org	1969	
3	ACTON	Chabad Center	148 Great Road, 01720; 978-929-2513	1999	
4	AMHERST	Jewish Community of Amherst	742 Main St., 01002; 413-256-0160; info@J.C.A.org	1969	
5	ANDOVER	Congregation Beth Israel	501 S. Main St. (PO Box 929), 01810; 978-474-0540; office@bethisraelmv.org	2004	
6	ANDOVER	Chabad Lubavitch of Merrimack Valley	310 N. Main St., 01810; 978-470-2288; rabbi@chabadmv.com	1992	
7	ANDOVER	Temple Emanuel	7 Haggitts Pond Rd., 01810; 978-470-1356; www.templemanuel.net	1920	
8	ANDOVER	Havurat Shalom	17 Olde Berry Rd., 01810; 978-470-2542		
9	ANDOVER	Tiereth Israel	510 S. Main St., 01810	c. 1990	
10	ARLINGTON	Hebrew Congregation of Arlington Heights	122 Sylvia St.	1906	
11	ASHLAND	Shomai Shalom, Reform Jewish Congregation of Ashland	PO Box 454, 01721; 508-231-4700; info@rjca.org	1995	
12	ATHOL	Agudas Achim	411 Pine Street	1907	
13	ATHOL	Temple Israel	107 Walnut St., 01331; 978-249-9481; www.templeisraelathol.org	1950	
14	ATTLEBORO	Agudas Achim Congregation	901 N. Main St., 02703; 508-222-2243; office@agudasma.org	1910/1911	
15	AYER	Congregation Anshey Shalom	Cambridge St. & Norwood Ave.	1948	
16	BELMONT	Beth El Temple Center	2 Concord Av., 02178; 617-484-6668; office.betc@verizon.net	1932/1936	
17	BEVERLY	Temple B'nai Abraham	200 East Lothrop St., 01915; 978-927-3211; tba.beverly@verizon.net	1897/ 1908	
18	BEVERLY	Congregation Ohav Shalom	3 Beckford St.	1903/1916	
19	BOSTON	Chabad House	491 Commonwealth Ave., 02215; 617-424-1190; rsp@chabadboston.org	1979	
20	BSTON	Temple Israel	477 Longwood Ave, 02215; 617-566-3960; info@tisrael.org	1854	
21	BOSTON	Congregation Bnai Jacob/ Zvihil-Mezbuz Beis Medrash	15 School St., 02108; 617-227-8200; Rebbe@rebbe.org;	1919	
22	BOSTON	Congregation Ohavi Sedek		<1887	
23					



# 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, comparable, and facilitates analysis. It can also enhance the resolution and accessibility of archived data.

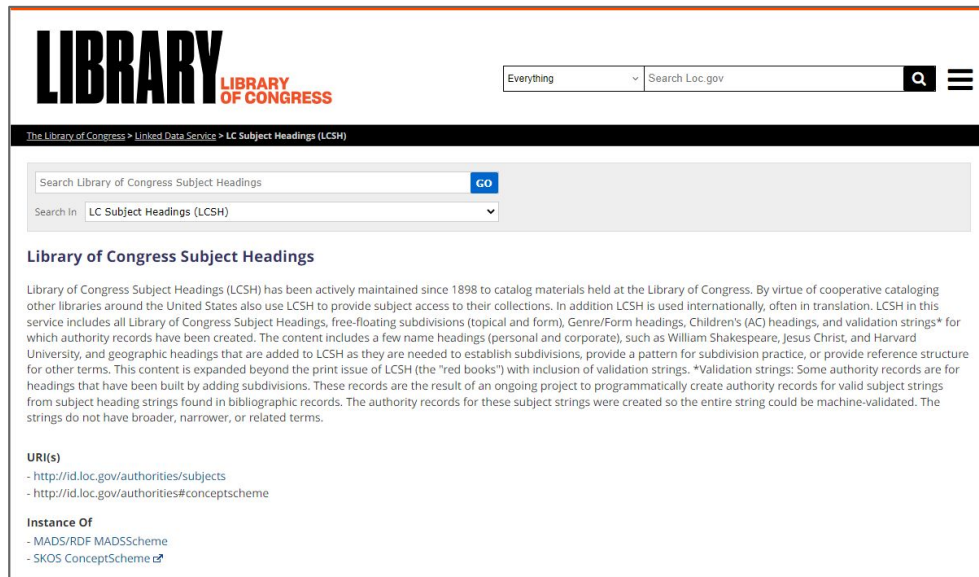


The [Homosaurus](#) vocabulary allows institutions to make LGBTQ+ resources more accessible, supplementing existing vocabularies like the LCSH (Library of Congress Subject Headings).



# 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 box with the text "Search Library of Congress Subject Headings" and a "GO" button. Below the search box, there is a dropdown menu labeled "Search In" with the selected option "LC Subject Headings (LCSH)". The page title is "Library of Congress Subject Headings". The main text describes the LCSH service, stating it has been actively maintained since 1898 to catalog materials held at the Library of Congress. It mentions that the service 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 [e](#)



# 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: <https://homosaurus.org/v3>

LCSH: <https://id.loc.gov/authorities/subjects.html>

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? Where 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?



# 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 >> <a href="#">Browse Alphabetical List</a> >> <a href="#">Format Descriptions as XML</a>		
Format Descriptions		
<b>Still Image</b> <ul style="list-style-type: none"><li>• <a href="#">SVG_1_1</a></li><li>• <a href="#">TIFF_6</a></li><li>• <a href="#">All still image format descriptions</a></li></ul>	<b>Sound</b> <ul style="list-style-type: none"><li>• <a href="#">WAVE</a></li><li>• <a href="#">MP3_FF</a></li><li>• <a href="#">All sound format descriptions</a></li></ul>	<b>Moving Image</b> <ul style="list-style-type: none"><li>• <a href="#">MPEG-4_FF_2</a></li><li>• <a href="#">AVI</a></li><li>• <a href="#">All moving image format descriptions</a></li></ul>
<b>Textual</b> <ul style="list-style-type: none"><li>• <a href="#">PDF/A family</a></li><li>• <a href="#">DOCX/OOXML_2012</a></li><li>• <a href="#">All text format descriptions</a></li></ul>	<b>Web Archive</b> <ul style="list-style-type: none"><li>• <a href="#">ARC_IA</a></li><li>• <a href="#">WARC</a></li><li>• <a href="#">All Web archive format descriptions</a></li></ul>	<b>Datasets</b> <ul style="list-style-type: none"><li>• <a href="#">DBF</a></li><li>• <a href="#">HDF5</a></li><li>• <a href="#">All dataset format descriptions</a></li></ul>
<b>Geospatial</b> <ul style="list-style-type: none"><li>• <a href="#">ESRI shape</a></li><li>• <a href="#">GeoPackage_1_0</a></li><li>• <a href="#">All geospatial format descriptions</a></li></ul>	<b>Generic</b> <ul style="list-style-type: none"><li>• <a href="#">ASE</a></li><li>• <a href="#">RIFF</a></li><li>• <a href="#">All generic format descriptions</a></li></ul>	



# 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.

For more information on digital preservation:

<https://library.duke.edu/using/policies/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](mailto:nulab.info@gmail.com)

**Developed by Juniper Johnson**  
Digital Integration Teaching Initiative  
DITI Research Fellow

**Dipa Desai**  
Digital Integration Teaching Initiative  
DITI Research Fellow

Slides, handouts, and data available at <https://bit.ly/sp23-parr-insh2102-data>

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

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

Schedule an appointment with us! <https://calendly.com/diti-nu>

