

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



API Documentation

- When using APIs for web-scraping, it is necessary to refer to the API documentation and a link is usually found on the API homepage.
 - 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 unaccounted for differences and API 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/>
- IMDB: <http://www.omdbapi.com/>
- FBI: <https://crime-data-explorer.fr.cloud.gov/api>
 - Other Federal government APIs:
<https://api.data.gov/docs/>
- Twitter: <https://developer.twitter.com/en/docs.html>



Ethical Considerations

Contextual Privacy

- When we think about privacy online we want to think of it as contextual. 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 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: Paul Revere, John Adams, 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 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?

- Open and closed data (as a political issue)
- Issue of **ownership** over data and ownership over insights from it
 - Making data in/operable can reflect the political intentions over who can use it, and for what
 - Commodification of data such as selling data to target audiences: potential for abuse, lack of accountability, and the difference in power between the person on whom data is collected, and the data collector
- **Privacy** - identifiable information: sometimes stored personal data gets stolen; sometimes making connections between different data points enable identification even in a seemingly anonymized dataset, etc.



Data: Critical Conversations + Ethics



Northeastern University
NULab for Texts, Maps, and Networks

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

Algorithms & Big Data: *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



Algorithmic Bias

- Algorithms are *not neutral*. **People create algorithms.**
 - Algorithmic processes—and even the data itself—reflect societal biases.
- When an algorithm is written or trained using data that misrepresents the actual population, this produces **algorithmic bias**.
- Similarly, **when data reflects biased realities**, the algorithm will continue to reproduce outcomes if those outcomes are desirable (despite their harm to—or erasure of—other groups).
- Algorithms reflect social inequalities, and can serve to exacerbate them.
- Read this [Vox article](#) for more information on algorithmic bias.



Alleviating Injustice

- When we look at the data used to train an algorithm, we must ask **what kinds of data** are being counted, and what kinds of data are being *overlooked, ignored, excluded*?
- What are the consequences of counting and not counting different kinds of data on various populations, especially marginalized groups?
- Will the technology and big data-driven solution **eliminate** human bias or **amplify** it?

“Algorithms by themselves are neither good nor bad. It is merely a question of taking care in how they are built.”

Sendhil Mullainathan, Chicago Booth University

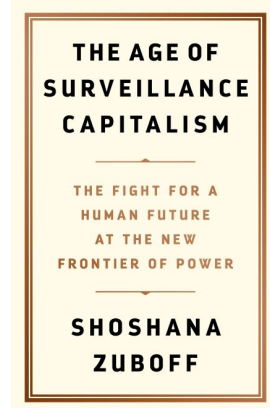
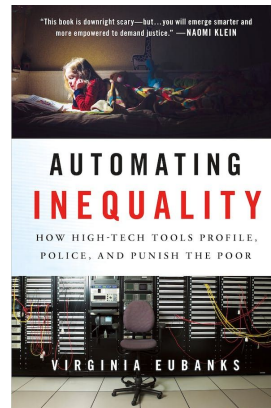
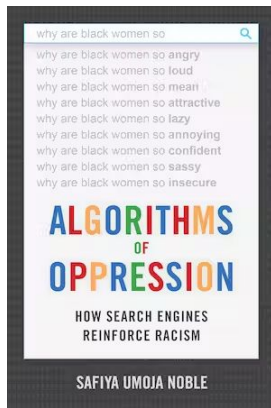
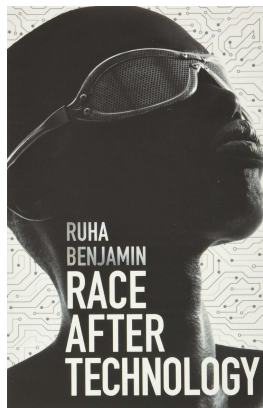
“Counting and measuring do not always have to be tools of oppression. We can also use them to hold power accountable, to reclaim overlooked histories, and to build collectivity and solidarity.”

Catherine D’Ignazio & Lauren Klein, *Data Feminism*, 2020



Method Showcase: 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 + 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 metadata as early as possible not only helps the users of your data, but helps you to plan for possible issues in building out the dataset
- Developing column naming conventions 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. These standards allow datasets 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 apparent missing information, and create opportunities for us to make data more inclusive.
- Examples of disciplinary metadata standards:
 - [Darwin Core \(DwC\)](#)
 - [NeXus](#)
 - [Data Documentation Initiative \(DDI\)](#)



Metadata Application Profiles

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: 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.



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

S

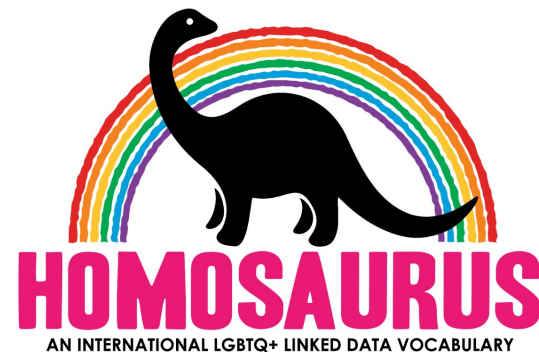
A

<

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).



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.

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

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>

