2 Data modeling and use

This chapter explores the central role of data in digital research, emphasizing that data serve as the foundational units for analysis and manipulation. It explains that structured data allows for systematic analysis, repurposing, and disambiguation, offering interpretive insights that unstructured data cannot. This chapter also discusses data models, which are intellectual tools used to extract information through processes like parameterization and tokenization, ensuring that data is discrete and machine-readable. Additionally, it addresses the ethics of data gathering, stressing that data is never neutral and must be carefully documented, especially when associated with individuals or groups, to prevent misuse, privacy violations, or exploitation for financial gain.

2a Making data

Exercise 2.1: Creating a data model

What kind of data model would you produce to study different musical tastes among your peers? What categories would be essential? Useful? How would you characterize tastes or preferences? How would you define "peers" in your group?

How-to example

To study different musical tastes among peers, you might define the peer group based on shared attributes like age, occupation, or location and build a data model that includes key categories such as demographic data (age, gender, occupation) and musical preferences (favorite genres, artists, top tracks, and listening frequency). Additional factors like music discovery methods, emotional influence on music choices, and social aspects like concert attendance or playlist sharing would provide deeper insights into musical behaviors. The model would also account for cultural and language preferences, helping to characterize each person's tastes based on genre affinity, listening habits, and the influence of mood or activities. This approach would allow for the analysis of how demographic and social factors shape musical preferences within the group.

Example data model schema

Attribute	*Description*	*Example values*
Peer ID	Unique identifier for each peer	1, 2, 3
Age	Age of peer	21, 23, 29
Gender	Gender identity of peer	"Male," "Female," "Non-binary"
Occupation	Occupation status	"Student," "Employee"
Geographic location	City, region, or country	"Los Angeles," "New York," "Germany"
Favorite genres	List of preferred genres	["Rock," "Jazz"], ["Pop," "Electronic"]
Favorite artists	List of favorite artists or bands	["Drake," "Taylor Swift"], ["Miles Davis"]
Top tracks	Most listened-to songs	["Blinding Lights," "Shape of You"]
Listening frequency	How often they listen to music	"Daily," "Weekly," "Occasional"
Preferred platforms	Music streaming platforms they use	"Spotify," "YouTube," "Apple Music"
Music discovery	How they find new music	"Recommendations," "Social Media," "Playlists"
Listening context	Activities during which they listen to music	"While studying," "At the gym." "Commute"
Emotional influence	Music preferences based on mood or emotion	["Happy," "Energetic"], ["Relaxing," "Calm"]
Concert attendance	Whether they attend concerts or live music events	True/False
Cultural influence	Influence of cultural background on their musical preferences	"Strong," "Moderate," "Low"
Language preference	Language(s) they prefer for lyrics	"English," "Spanish," "Bilingual"

Exercise 2.2: Using AI to create a data model

With Large Language Model (LLM) functionality available through tools such as ChatGPT and CoPilot, AI is contributing to the generation of data models. With thoughtfully crafted prompts, we are able to request a gender dataset example based on the demographics of California. The LLM can produce an illustrative example based on general trends derived from publicly available data, such as US Census Bureau estimates and California-specific demographic studies, like those done by the California Department of Finance. While the resultant table may be simplified for demonstration purposes, it can be exported as a csv to be expanded on or visualized. Additionally, AI can suggest alternative ways to structure the data, additional data sources, and ethical or bias blind spots. It can also provide guided workflows for how to map one data structure to another or build a data model based on specific schema standards. See the following resources for more information about AI for data modeling:

SoftBuilder. 2025. *Revolutionizing Data Modeling with Generative AI: The Future Is Now*. Accessed April 18, 2025. https://soft-builder.com/revolutionizing-data-modeling-with-generative-ai-the-future-is-now/.

Striim. 2025. *5 Key Principles of Effective Data Modeling for AI*. Accessed April 18, 2025. https://www.striim.com/blog/5-key-principles-of-effective-data-modeling-for-ai/.

Below is an expanded approach to classroom data modeling, along with considerations for context, time, and purpose created by AI.

1. **Identifying available data**
   The first step in data modeling is to determine what data is available or relevant in the classroom. This will include both **static** data (that does not change frequently) and **dynamic** data (that changes based on context or time).

   Potential data points in a classroom:

   - **Physical environment:**
     - Number of **chairs**, **desks**, and other furniture
     - Number of **windows**, **doors**, and **lights**
     - **Temperature** in the room (fluctuates with time of day, season, and HVAC system)
     - **Health and sanitary conditions** (cleanliness, presence of hand sanitizers, etc.)
     - **Accessibility** (e.g., wheelchair ramps, size of aisles, braille on signs)
     - **Age and condition of infrastructure** (quality of walls, plumbing, electrical systems)
   - **Occupants:**
     - Number of **students** present (varies based on time, attendance)
     - **Gender** breakdown (male, female, non-binary, trans, queer, etc.)
     - **Age range** of occupants
     - **Teachers** and **staff** present
     - **Purpose** of the occupants (students, maintenance workers, visitors, event attendees)
   - **Temporal factors:**
     - **Time** of day (affects population, activities, lighting needs, etc.)
     - **Day of the week** (weekday vs. weekend, school holidays)
     - **Season** (impacts heating/cooling, daylight, etc.)
     - **Frequency** of the sample (how often data is collected or observed: hourly, daily, weekly, etc.)

2. **Contextual variability**
   Data in the classroom will vary depending on context. For example, the model must account for changes based on the time of day, week, or even year. Additionally, transient items (e.g., folding chairs) may skew data if not accounted for properly.

   Contextual factors to consider:

   - **Events vs. regular school day**: Classroom population and layout may differ drastically during events (e.g., guest lectures, school functions) as compared to normal class hours.
   - **Transitory vs. permanent items**: Some items like folding chairs, extra desks, or books may not permanently belong to the classroom but appear temporarily. A flag should be incorporated into the model to denote whether items are transient or fixed.
   - **Purpose-specific data**: The data you collect may differ based on the specific use case:
     - **Safety assessments**: Would include factors like room capacity, exit accessibility, or hazards (sharp objects, toxic materials).

- **Health checks**: Might focus on sanitary conditions, air quality, and ventilation.
- **Resource usage**: Focus on how the room is utilized (e.g., furniture condition, occupancy rate, lighting energy consumption).

3. **Purpose of the data model**
The purpose of the model is vital as it determines which data points are necessary, how frequently they should be collected, and how the data is processed. Different scenarios require different approaches:

Example purposes and related data:

- **Health and safety evaluation:**
  - **Air quality, room temperature**, and **lighting conditions**
  - Number of **accessible exits** and **capacity**
  - Presence of **sanitation materials** (hand sanitizer, masks, etc.)
  - **Gender** breakdown if it affects safety concerns or regulations
- **Classroom usage and efficiency:**
  - **Occupancy rates** (number of students, teachers, etc.)
  - Number of **transitory items** (folding chairs during events)
  - **Energy usage** (lighting, HVAC based on time and season)
  - **Accessibility** issues, including the quality of technology for special needs
- **Justification for new resources (furniture, equipment):**
  - Age and condition of current **furniture and equipment**
  - Frequency of **student and teacher usage**
  - Changes in **occupancy patterns** (gender, age, and number of students)
  - Room **capacity needs** based on typical and peak usage times

4. **Data model design**
Now, let's translate these ideas into a simple **data model**. This can be represented as a structured table or database schema.

Classroom data model:

Attribute	*Type*	*Description*	*Example values*
Room ID	Integer	Unique identifier for each classroom	101, 202, 303
Chair count	Integer	Number of chairs present	30, 50, 70
Desk count	Integer	Number of desks	20, 35, 45
Window count	Integer	Number of windows	2, 4, 6
Light count	Integer	Number of lights	8, 12, 20
Temperature	Float	Room temperature in Celsius	18.5, 21.0, 23.3
Occupant count	Integer	Number of people in the room	15, 30, 60
Gender breakdown	String	Proportion of male, female, non-binary, etc. occupants	"10 male, 15 female"
Age range	String	Age range of occupants	"5–10," "18–25," "30–60"
Event type	String	Type of event (normal class, special event, maintenance)	"Class," "Event," "Maintenance"
Transient items	Boolean	Whether the room contains transitory items	True/False
Accessibility status	String	Level of accessibility (good, moderate, needs improvement)	"Good," "Moderate," "Needs Improvement"
Sample time	Timestamp	Date and time when the data was sampled	"2024-10-03 14:00," "2024-10-03 09:00"
Purpose	String	Purpose of data collection (usage, safety, etc.)	"Safety," "Health," "Usage"
Room condition	String	Infrastructure quality	"New," "Good," "Worn"

5. **Considerations for data collection**

- **Frequency**: Data collection frequency depends on the purpose. For safety assessments, data might be collected once a month or after specific events. For room usage, daily or weekly data might be necessary.

- **Real time vs. static**: Some data points, such as temperature and occupant count, may fluctuate and require real-time tracking, while others, like the number of desks, might be static.

  This flexible model ensures that the classroom environment can be studied from multiple perspectives, depending on the goal of the data collection.

Use the checklist from Chapter 2 and apply it to the expanded discussion based on the classroom example above generated by AI. Assess how well the AI did. Did it meet all the criteria in the checklist? What issues might you have with using AI for creating this data model?

Checklist for creating a data model

- Determine intellectual property rights and permissions for data collection and engage in discussion with institutional owners and/or community leaders where appropriate.
- If your research involves human subjects, be sure you are familiar with the legal restrictions governing data collection and follow the IRB (Institutional Review Board) procedures.
- Identify information that can be described unambiguously in your sources.
- Create a set of labeled categories that are clear and distinct.
- Decide which categories will be specified numerically, by true/false values, or by descriptive terms.
- Assess the difficulty of keeping your entries consistent.
- Standardize the format for each data type (dates, places, and names).
- Limit the vocabulary for description except in the comments.
- Read the categories you have established to see what the data model will take from the original sources.
- Check for assumptions about categories that may embody bias.
- Check your data model for completeness and accuracy.

Exercise 2.3: Ethics of legacy data

A research project lists all of the individuals in the diary entries of a famous politician. Many of these are coded to protect the identity of the persons involved, some of whom were the politician's gay lovers. Several of these are still alive and well-known figures. A key to all of these has been provided by a scholar. Where do these belong in the spreadsheet?

How-to example

The spreadsheet should maintain anonymity. The coded entries should remain in place to protect their identities, as originally intended, and the key provided by the scholar should be kept separate and confidential to avoid compromising their privacy. Disclosing sensitive information about living individuals without their consent could have serious personal and professional consequences, so the key should only be accessed by approved researchers under strict ethical guidelines. Levels of access should be considered, including password-protecting hard drives and computers. Cloud storage may need to be avoided or employ two factor authentication to ensure security. Most importantly, legacy data should always be checked with IRB to ensure that it can be used legally, and if you are working with communities that you are not a part of, that you are taking intentional steps to make sure that community's views are represented in your approach and treatment of the data.

Recommended readings

Flanders, Julia, and Fotis Jannidis. 2012. "Knowledge Organization and Data Modeling in the Humanities." *Whitepaper*. https://wwp.northeastern.edu/outreach/ conference/kodm2012/flanders_jannidis_datamodeling.pdf.

Flanders, Julia, and Fotis Jannidis, eds. 2019. *The Shape of Data in the Digital Humanities: Modeling Texts and Text-Based Resources*. London: Routledge. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1926341.

Christine Cepelak. 2024. "An Introduction to Data Ethics: What Is the Ethical Use of Data?" *DataCamp*. https://www.datacamp.com/blog/introduction-to-data-ethics.

GeeksforGeeks. 2024. "Data Modeling: A Comprehensive Guide for Analysts." *GeeksforGeeks*. https://www.geeksforgeeks.org/data-modeling-a-comprehensive-guide-for-analysts/.

Onuoha, Mimi. On Missing Datasets. https://github.com/MimiOnuoha/missing-datasets.

Ochsner, J. 2020. "Why Human Subjects Research Protection is Important." *The Ochsner Journal*, National Library of Medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7122250/.

Sample IRB guidelines from UCLA. https://ohrpp.research.ucla.edu/policies-and-guidance/.

Journal of Open Humanities Data. 2025. Accessed April 18, 2025. https://openhumanitiesdata.metajnl.com/.

2b Cleaning and using data

Exercise 2.4: Creating a data extraction scenario

Suppose a scholar is interested in looking at the ways the prices of first editions and manuscripts of African American authors have changed in the last century. What kinds of data can be extracted from current sales records? Create a scenario in which this occurs by tracking the current prices of the work of Langston Hughes and Alain Locke. What is missing that you would like to be able to make into data? Do you think these would be different from what could be extracted from dealers' records? Imagine what kinds of data would be in the dealers' records that might be different from in this online site: https://www.abebooks.com/.

How-to example

Scenario**:** A scholar decides to track the current prices of Langston Hughes' first edition of *The Weary Blues* and Alain Locke's *The New Negro*. By browsing online sales platforms like AbeBooks, they note that a signed first edition of *The Weary Blues* is listed at $10,000, while a rare first edition of *The New Negro* is priced at $5,000. By collecting and comparing prices over time, the scholar could analyze trends, perhaps seeing price spikes around cultural anniversaries or during shifts in the market demand for African American literature.

When a scholar explores the changing prices of first editions and manuscripts by African American authors over the last century, such as those by Langston Hughes or Alain Locke, they can extract a variety of data from current sales records. Platforms like AbeBooks offer insights into the sale price of rare books, details about the item's condition (ranging from "fine" to "poor"), and specific publication details, such as the edition and year. Additionally, the platform may list special features like signatures, handwritten notes, or limited-edition prints, all of which affect the value. Seller reputation and location also play a role in pricing, potentially influencing the perceived worth of a given item.

However, online listings often lack historical sale data, which would show how prices have fluctuated over time. Moreover, information about the provenance or ownership history, which can increase the value of rare items, is often incomplete or unavailable. Additionally, while items are frequently labeled as "rare," platforms like AbeBooks provide no standardized metric to quantify rarity, leaving this claim somewhat subjective. This absence of detailed historical and contextual data makes it harder for scholars to comprehensively track trends in the market for African American literary works.

Recommended readings

Chetcuti, Ian. 2023. "Data Cleaning: Common Mistakes and How to Do it Right." *Medium*. September 2023.

Data Space. August 2024. "Data Cleaning 101." *Data Space Academy*. https://dataspaceacademy.com/blog/data-cleaning-101-key-dos-donts-for-flawless-results.

Owens, Trevor. 2011. "Defining Data for Humanists: Text, Artifact, Information or Evidence? *Journal of Digital Humanities* 1 (1). https://trevorowens.org/2011/12/15/defining-data-for-humanists-text-artifact-information-or-evidence/.

Tóth-Czifra, Erzsébet. 2019. "DARIAH Pathfinder to Data Management Best Practices in the Humanities." Version 1.0.0. DARIAH Campus [Pathfinder]. https://hdl.handle.net/21.11159/019595b2-cca1-70ad-b1e3-d088b4409de5.

Bibliography

Drucker, Johanna. 2011. "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly* 5 (1). https://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Floridi, Luciano, and Mariarosaria Taddeo. 2016. "What Is Data Ethics?" *Philosophical Transactions* A374: 20160360. https://dx.doi.org/10.1098/rsta.2016.0360.

Posner, Miriam. n.d. "Parts of Your Data." https://miriamposner.com/classes/dh101f17/tutorials-guides/data-manipulation/parts-of-your-data/.

Vigen, Tyler. 2022. "Spurious Correlations." https://www.tylervigen.com/spurious-correlations.