# 5 Database design

The creation of a database for humanities research requires careful analysis to structure data logically, minimize redundancy, and ensure efficiency. A key challenge is determining how entities and their relationships should be organized within tables and fields, all while respecting cultural and ethical standards. Though databases often rely on established classification systems, they should also be seen as interpretative tools that reflect a particular perspective. Databases can take various forms, such as relational, hierarchical, or graph-based, each with unique structures. Ethical considerations, especially in handling legacy data and ensuring accessible design, are critical throughout the process.

## 5a Database basics

***Exercise 5.1: Modeling a database***

Analyze your research to create tables of information that can be managed effectively. What data belongs in which table? What benefits are there to this organization and structure?

*How-to example*

Imagine you inherited a vinyl record collection from a family member, and you want to better understand what type of music is in the collection. Organizing the collection into a database or catalog would be a great way to analyze and understand its contents. Here's a structured approach to achieve this:

1. Start by creating a table that captures essential details about each record. This will give you a snapshot of the diversity in terms of genres, artists, and eras. (Fields: Record_ID [unique identifier], Title, Artist, Genre, Release Year, Label, Country of Release)
2. Then create additional tables to populate certain fields.
    a) Create a table for the Genre field. You will want to create a table that defines different genres (e.g., Jazz, Rock, Classical, Funk). Fields in this table may include Genre_ID, Genre_Name, and Description.
    b) Create a table for Artist information. This table stores artist details, helping you explore the diversity of musicians, their cultural backgrounds, and their prominence in different genres and eras. Fields in this table would include Artist_ID, Name, Genre_ID, Nationality, and Era.
    c) Depending on your interest, you may also choose to have a Condition and Rarity table that can help you identify particularly valuable or unique items in your collection. Fields in this table include Record_ID (unique identifier), Condition (Mint, Good, Poor), Rarity Level (e.g., Limited Edition, First Pressing).

This organization and structure would allow for you to analyze genre distribution, decade or era clustering, artist popularity and diversity, and rarity outliers. This structured approach not only gives you insight into the types of music in the collection but also helps you appreciate its historical and cultural value. This exploration may reveal patterns and surprises, deepening your understanding of the inherited collection.

## Recommended readings


Geeks for Geeks. 2019. "Difference between RDBMS and OODBMS." www.geeksforgeeks.org/difference-between-rdbms-and-oodbms/.

Microsoft (Written for Access, but the Principles Are Applicable to Any Database). https://support.microsoft.com/en-us/office/database-design-basics-eb2159cf-1e30-401a-8084-bd4f9c9ca1f5.

Pina, Eduardo, José Ramos, Henrique Jorge, Paulo Váz, José Silva, Cristina Wanzeller, Maryam Abbasi, and Pedro Martins. 2024. "Data Privacy and Ethical Considerations in Database Management." *Journal of Cybersecurity and Privacy* 4 (3): 494–517. https://doi.org/10.3390/jcp4030024.

Ramsay, Stephen. 2004. "Databases." In *The Companion to Digital Humanities*, edited by Susan Schreibman, Ray Seimens, and John Unsworth, Ch. 15. Oxford: Blackwell's. www.digitalhumanities.org//companion/.

## 5b Database issues: legacy data, ethics, use

***Bonus content: a note on the history of databases***

While spreadsheets and tabular records have been part of human culture for thousands of years, the relational database is a surprisingly new innovation ("History of Databases" 2011). Computer scientist Edgar Codd is given credit for the idea of the "relational database." His work in 1970–72 defined the underlying structure of relations—that instead of flat or hierarchical structures, efficiencies and benefits could be implemented through relational organization (Brown 2002). Codd's insight was fundamental—why not cross-reference tables that contained different parts or types of information instead of having everything stored in a single set of rows and columns? This insight was revolutionary (Codd 1970). Codd worked at IBM, as did his fellow researchers, Raymond Boyce and Donald Chamberlain, who invented the programming language for queries, originally named SEQUEL, now referred to as SQL.

The breakthrough for industry applications came around 1979, however, when Dan Bricklin and Bob Frankson invented VISICALC (a contraction of the words "visible calculator"). Known as the "killer app," it transformed personal computers from things used for hobbies to essential tools for business purposes. The idea of a program that could calculate spread sheets automatically by altering the value of variables gave every business from insurance to manufacturing ways to project financial scenarios instantly. As a labor-saving application, it was an instant success for its capacity to recalculate spreadsheets. One major addition to these technologies has been object-oriented databases (mentioned earlier), which combine operations and entities in their design. In addition, various databases that do not rely on SQL to query and retrieve information, such as graph databases (which underlie networks and support their visualizations), have joined their predecessors (Panwar 2020). What is remarkable is how few new database structures have been added, and how robust (long-lived) and viable these forms of data structure have been.

***Bonus content: discussion of debates about databases***

Debates in the digital humanities in the 2000s addressed fundamental definitions of database forms in relation to traditional scholarly formats, such as essays and narratives. More recently, discussions of race and power, exclusionary practices, and politics of the academic world and knowledge work have also made critical contributions, but few have focused on database structures, only design and implementation.

The idea that databases were the new, current, and future form of knowledge and that they would replace narrative in the study of history, the creation of literature, or the development of artistic expression was asserted by several digital humanists in the 1990s and early 2000s.

Among the assertions was that databases were non-linear while narratives were linear, that processes of selection resulted in fixed narrative modes while processes of combination are at the heart of database "logic." The potential for multiple readings of information, even of interpretative data in structured fields, seemed to suggest a radical shift in methods of working with humanities materials. The arguments had a strong techno-deterministic tone, suggesting that changes in ways of thinking are the direct result of changes in the technology we design and use. Counter arguments suggested that combinatoric work and content models are integral elements of human expression and have been since the beginnings of the written record, which can be dated to five or six thousand years ago in Mesopotamia.

The distinction between database structures and narrative forms is real, but are they in opposition to each other or merely useful for different purposes and circumstances? Why make such strong arguments on either side? At stake seems to be the definition of what constitutes human expression and the rules and conventions according to which it can create the record of lived and imaginative experience. But also at stake is an investment in the ways we value and assess new media and their impact on traditional methods of scholarship.

***Exercise 5.2: asking humanist questions***

Go to http://data.gov and look at the data sets under Agriculture. What humanities questions can you ask of this data? Why would a data set like Fruit and Vegetable Prices (https://catalog.data.gov/dataset/fruit-and-vegetable-prices) be useful for humanistic research? What other information would you want to link to this dataset once you have formulated your project?

*How-to example*

The "Fruit and Vegetable Prices" dataset could be valuable for humanistic research because it provides insights into the relationships between food costs and broader social, cultural, and economic factors. For example, it can help researchers explore how pricing impacts dietary choices across socioeconomic groups, shedding light on health inequities and access disparities. Geographic variations in pricing can highlight regional economic inequalities, while historical price trends can offer a lens into changing cultural practices and values surrounding food. Furthermore, the dataset is a powerful tool

for critiquing agricultural policies, examining how subsidies or trade agreements influence consumer prices, and assessing the ethical implications of food deserts and accessibility.

To maximize the utility of this dataset, linking it with complementary information is essential. Demographic and income data can illuminate the socioeconomic dimensions of food affordability. Health datasets can provide a basis for examining the relationship between food costs and dietary health outcomes, such as obesity or malnutrition rates. Agricultural production data, such as crop yields and farming practices, can reveal the interplay between food supply dynamics and pricing. Additionally, geographic and environmental data can contextualize price fluctuations by considering climate conditions, while historical and trade data can show how globalization and trade policies have shaped local food markets. Together, these connections allow for a nuanced exploration of how economic and environmental factors intersect with culture, ethics, and social structures.

***Exercise 5.3: working with legacy data***

Europeana and King's College are just two examples of projects that have absorbed legacy data. Seek out a resource on a cultural site that is not familiar to you, such as the Kyoto National Museum in Japan, and see if you can determine what is legacy data and what is new material. https://knmdb.kyohaku.go.jp/.

*How-to example*

The Kyoto National Museum's Collection Database offers a wealth of information on its holdings, including artworks and artifacts ranging from prehistoric times to the modern era. This resource reflects a mix of legacy data and newly cataloged materials. Legacy data often includes items that have been part of the museum's collection for decades, such as artworks classified as "National Treasures" or "Important Cultural Properties," which have detailed historical records, like the 12th-century Buddhist painting *Bishamon-ten (Vaisravana)*. This artwork is a part of the museum's longstanding collection, indicating its legacy status. In this example of a landscape painting on a folding panel (https://knmdb.kyohaku.go.jp/eng/687.html), there are two descriptions, demonstrating that they may be preserving both the former description and the new description for the items.

Conversely, new material in the database might include digital records or acquisitions made in recent years, often involving modern data categorization methods or enhanced imaging technologies to make the artifacts more accessible to the public and researchers alike. The presence of parent/child relationship within the category field may also indicate how additional categories have been created to detail the different types of data that has been collected over time.

## Recommended readings

Christie, Michael. n.d. "Computer Databases and Aboriginal Knowledge." https://www.ria.ie/assets/uploads/2024/06/allea_sustainable_and_fair_data_sharing_in_the_humanities_2020_0.pdf.

## Bibliography

ALLEA (All European Academies). 2020. https://www.ria.ie/assets/uploads/2024/06/allea_sustainable_and_fair_data_sharing_in_the_humanities_2020_0.pdf.

Brown, Farmer. 2002. "A Brief History of Modern Relational Database Management Systems." www.mountainman.com.au/software/history/it1.html.

Ciulla, Arianna. 2020. "Exposing Legacy Data." *King's Digital Library Blog*.

Codd, Edgar. 1970. "A Relational Model of Data for Large Data Banks." *Communications of the ACM* 13 (6). https://history.computer.org/pioneers/codd.html.

Freire, Nuno, Enno Meijers, René Voorburg, and Antoine Isaac. 2018. "Aggregation of Cultural Heritage Datasets through the Web of Data." *Science Direct*. https://www.sciencedirect.com/science/article/pii/S1877050918316168.

Hinfelaar, Martijn. September 20, 2021. "Cross Cultural Design: Challenges and Considerations." https://wearebrain.com/blog/cross-cultural-design-challenges-and-considerations/.

"History of Databases." 2011. "Computer History Museum." https://www.computerhistory.org/revolution/memory-storage/8/265/2207.

Kulesz, Octavio. 2018. "Intergovernmental Committee for the Protection and Promotion of Diversity of Cultural Expressions." *UNESCO*. https://unesdoc.unesco.org/ark:/48223/pf0000380584.locale=en.

Masters, Christine L. 2015. "Women's Ways of Structuring Data." *Ada: A Journal of Gender, New Media, and Technology* 8.

Padilla, Thomas H. 2016. "Umanities Data in the Library: Integrity, Form, Access." *D-Lib Magazine* 22 (4). www.dlib.org/dlib/march16/padilla/03padilla.html.

Panwar, Arjun. 2020. "Types of Database Management Systems." www.c-sharpcorner.com/UploadFile/65fc13/types-of-database-management-systems/.

Smithies, James, Carina Westling, Anna-Maria Sichani, Pam Mellen, and Arianna Ciula. 2019. "Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King's Digital Lab." *Digital Humanities Quarterly* 13 (1). www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html.

Wuttke, Ulrike. 2019. “Here There Be Dragons: Open Access to Research Data in the Humanities.” *OpenMethods Metablog*. https://dhmethods.hypotheses.org/262.

Ximin Chu, Xin (Robert) Luo and Yan Chen. April 2018. “A Systematic Review on Cross Cultural Information Systems.” *Science Direct*. https://www.sciencedirect.com/science/article/abs/pii/S0378720618303422.

## Resources

- Data Privacy and Ethical Considerations in Database Management (https://www.mdpi.com/2624-800X/4/3/24)
- DBMS Learn Database Management System (Updated 30 Aug 2024) (https://www.geeksforgeeks.org/dbms/)
- Description of types of databases (https://www.geeksforgeeks.org/types-of-databases/)
- Digital Sanskrit Buddhist Canon (http://www.dsbcproject.org/gallery)
- Draw.io (http://Draw.io)
- Ian, 10 Essential Database Concepts that All Beginners Must Learn, updated July 31, 2024 (https://database.guide/10-essential-database-concepts-that-all-beginners-must-learn/)
- Indigenous Data Sovereignty and Governance. (https://nni.arizona.edu/our-work/research-policy-analysis/indigenous-data-sovereignty-governance) See also: https://indigenousdatalab.org/
- Information about visual database design (https://www.mysql.com/products/workbench/design/)
- Jongen, Tom. “How to make your data project ethical by design,” *Medium,* June 11, 2021. https://towardsdatascience.com/how-to-make-your-data-project-ethical-by-design-99629dcd2443
- Mukurtu (https://mukurtu.org/)
- Segalla, Michael and Dominque Rougiès. 2023. “The Ethics of Handling People’s Data,” *Harvard Business Review*, July-August, 2023. https://hbr.org/2023/07/the-ethics-of-managing-peoples-data
- SQL Tutorials on W3Schools (https://www.w3schools.com/sql/)
- The Noun Project (https://thenounproject.com)