Introduction

Databases are at the heart of the modern world. The need to structure and collect information in an accessible and analytically useful way has been with us throughout human history, but modern technology has given a wide range of new possibilities for making these processes of collection, storage, and analysis more dynamic. The database – a computer system for storing data in a structured way that can be easily sorted, searched and manipulated – is one of the most important technologies we have for information manipulation and management.

Researchers in social sciences and the humanities do not make use of these tools as often as their counterparts in the sciences. Data about humans and their societies has a number of features that create specific issues with the process of categorising and structuring data; there are significant potential pitfalls when humanities data is mishandled, leading to inaccuracies or even serious ethical issues with research. Thus, despite both the centrality of processes of categorisation to how human societies understand the world and the potential utility of making information better tractable and searchable, scholars studying the human world often refrain from utilising database methods.

This course aims to help bridge these gaps. In the following sections and activities, we provide humanities and social science graduate students both with a basic overview of the technical skills needed to query and produce a database, and also a range of information on the challenges faced when dealing with humanities data and how these may be solved or avoided. Finally, we also provide some information on the planning and construction of database projects, from setting project scope to long-term data storage and the ethical issues faced when dealing with humanities data.

Definitions: What is a database?

A database is a structured set of data recorded with the minimum of redundancy, to satisfy simultaneously multiple users selecting and using data in a timely manner.

A database should not occupy too much space on the disk of the computer. This is why redundancy of data should be avoided: this means that the same information should be stored only once. This saves time and minimizes risk of error.

The database approach to storing information occurred due to a threefold evolution:

- Large volumes of data, centralized or distributed, which must be accessible in good time. High volume data collection and analysis was developed significantly in the modern era by states for census, military and taxation purposes, and by private actors for scientific and business purposes. In all these cases there are strong incentives to speed up data handling to achieve relevant objectives efficiently.
- Equipment evolution (performance increase, component integration, cost reduction, etc.) Increases in computing power and system performance allowed the handling of progressively larger datasets electronically.
- Software development (operating systems, client server architectures and networks). The development of progressively more accessible programs and tools allowed wider use of database systems and architecture.

For example, imagine a commercial database: the data may relate to customers, products, orders, order lines, and so on. The requests are very varied:

- What is the list of products that have been ordered by customer number "CL10"? ; RESULT.
- Who is the customer for order number "C004"? ; RESULT.
- What is the date of the last order of the customer with the name "AMS"? ; RESULT.

A database is the system used for finding these sorts of pieces of connected information, via a system for *querying* (that is, asking formal and structured questions of) the data.

Examples of Humanities Databases

Databases can be used for a number of different important functions in the humanities. This section will discuss in very broad terms some of the ways they can be used.

The first sorts of databases many humanities and social science postgraduates will come across in their research are **databases for collecting resources**. These, for example, are what most modern online library catalogues or databases of academic papers consist of. In that case the records in the database store a reference to (and perhaps also the text of) a particular source, as well as an array of information and category links: the database structure, often accessed via an online interface, allows you as the user to query and navigate these. Online search engines work on a similar principle.

We can, however, make more in-depth uses of databases than this for research purposes. A database **collecting material** – that is, below the level of whole resources – within a particular research project can be invaluable. This might be collecting text, images, parts of images, or references of some kind – for example, collecting samples of referenced information about particular entities like people or places. If the material is sufficiently consistent in its type, the database may then allow analyses by comparing different sub-sections of the material or taking particular sections of it for comparison.

A particular variant of databases that collect material that is worth mentioning are **databases of texts**. These databases may contain entire texts, or in many cases might contain text stored on a word-by-word basis, allowing additional information to be attached to each word through the database's structures. Such a system can be especially important for linguistics or literary analysis of texts, allowing for analysis of changing frequencies of word use, analytical comparisons between different variants of a text, and assessments of word patterns and structures in the language.

A database can also be used to produce a **model** of a system, population, or period that is being studied. This may in many ways be very similar to collecting material, but in this case the aim is to collect and formulate material such that it produces a consistent idea of what the world might look, or have looked, like. This is important because it adds additional constraints and possibilities: a consistent model of the world may have problems if it contains contradictory data, but it can be used for some analytical or predictive research work that a simple collection of material cannot.

Do I need a database?

This course is focused on research databases – that is, those that contain material, text, and models – and how to construct and use them in your own work. As a first step to considering doing so, you must consider whether a database is the correct tool for arranging and examining your material.

Databases are generally best used when dealing with significant quantities of data, with data of a variety of different types or about a connected range of different entities, and with data that need to

be dynamically sorted and manipulated. If your data fits some of these categories, a database may be a very helpful way to store it. If you have smaller or simpler quantities of data, or do not need the capacity for sorting and manipulation that databases offer, then there might be simpler ways to do the things you need for your project.

Consider, too, what your other options are. For example, if you have information on a group of 2000 people but you do not need to consider their connections or include any types of entity other than people, or if you wish to store some information about books but it does not need to be connected to extensive data about the authors or other entities, in these cases a spreadsheet may be all you need. Spreadsheets, often found as .xls, .xlsx or .csv files, form a single table, and can be used for a wide range of analyses and calculations: they generally assume a limited range of similar data entities and types.

However, in a slight change on the above examples, let's now imagine that the 2000 person group imagined in the previous paragraph will all be associated and connected to one another, and to different districts they live in which we also have information on. We now have a set of data that needs to spread across multiple tables, of districts and people, and which needs to have strong internal connectivity. This more connected set of data would be better stored in a database. The same might be true of our books if we wanted to connect them to entities like authors and publishers, and so on.

Database technologies

There are a range of systems for creating and accessing databases. These can be broadly grouped into different database models, and then more specifically there are a range of different programs for managing, creating and using databases available.

Most of the technologies discussed in this course are of an overall family of database types called **relational databases**. These are the major standard for most modern uses of databases, and present databases as a system of interconnected tables. Relational databases offer clear and precise structuring for data, and are very well supported with standard methods for their use. Other database types do exist, however. These include graph databases, which store data as a set of connected nodes and edges, increasing the focus on the connectivity of data. Most graph database technologies and management system are however somewhat newer and less well standardised.

Relational databases share a standard system for retrieving particular bits of data, a query language called SQL which we cover in more detail later in this course. There are then various database management system (DBMS) computer programs which can create and access databases of this kind. Our examples use **Microsoft Access**, which is a common program for creating and accessing relational SQL databases, but other programs and systems like MySQL, SQLite, and PostgreSQL also use the SQL language and relational structures, so the technical skills discussed in this course can transfer across a range of different database management tools.

Conclusions

In this first section we've introduced in some more detail the concept of a database and some examples of how humanities researchers might want to use them. We've also discussed how to

decide whether you need a database, and given a very brief overview of the names of some of the programs and technologies used.

The next sections of this course will take you through a range of the basic concepts and terminology of databases, starting from first principles, and also discuss some of the issues of turning humanities source material into data that can be stored and used in database formats. There is a short accompanying reading list with the course, which includes a number of academic discussions of particular topics that may be of interest. Each stage of the course has some exercises available, which you should do before moving on to the next section to get the most out of the materials.

Exercises

1. Find three examples of a database from your field containing research data. You may look for datasets referenced in papers you have read, ask colleagues or supervisors for examples, or find them with library and internet searching tools.

   If you cannot find three examples, please select some examples from the database resources list included with this course.

2. Write, for each database, some thoughts on why a database might have been an appropriate solution to research problems in that case, and discuss your thoughts with a classmate or colleague.