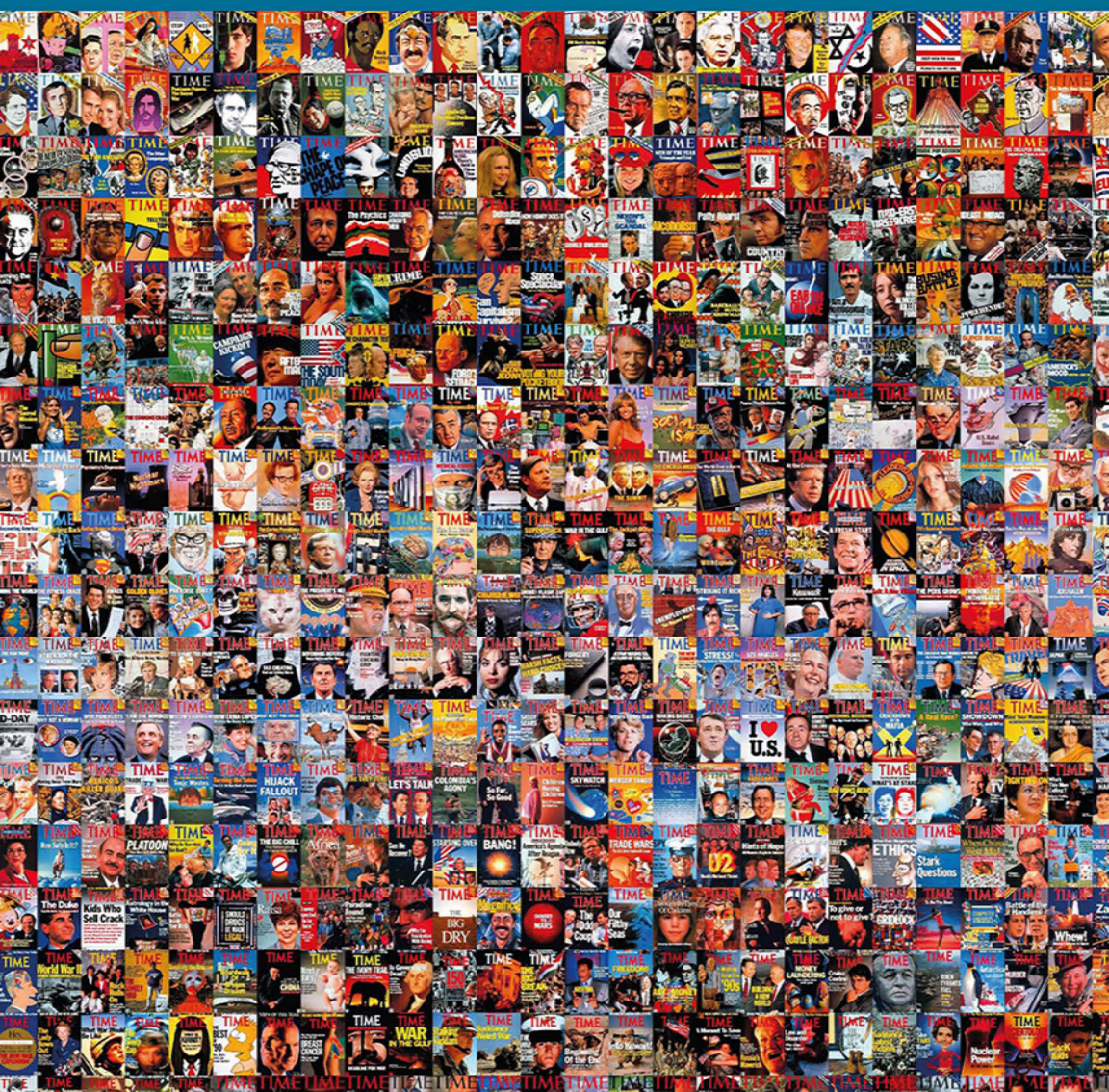
The Digital Humanities Coursebook

An Introduction to Digital Methods for Research and Scholarship

Johanna Drucker



The Digital Humanities Coursebook

The Digital Humanities Coursebook provides critical frameworks for the application of digital humanities tools and platforms, which have become an integral part of work across a wide range of disciplines.

Written by an expert with twenty years of experience in this field, the book is focused on the principles and fundamental concepts for application, rather than on specific tools or platforms. Each chapter contains examples of projects, tools, or platforms that demonstrate these principles in action. The book is structured to complement courses on digital humanities and provides a series of modules, each of which is organized around a set of concerns and topics, thought experiments and questions, as well as specific discussions of the ways in which tools and platforms work. The book covers a wide range of topics and clearly details how to integrate the acquisition of expertise in data, metadata, classification, interface, visualization, network analysis, topic modeling, data mining, mapping, and web presentation with issues in intellectual property, sustainability, privacy, and the ethical use of information.

Written in an accessible and engaging manner, *The Digital Humanities Coursebook* will be a useful guide for anyone teaching or studying a course in the areas of digital humanities, library and information science, English, or computer science. The book will provide a framework for direct engagement with digital humanities and, as such, should be of interest to others working across the humanities as well.

Johanna Drucker is the Distinguished Breslauer Professor in Bibliography in Information Studies at UCLA, USA. She has published widely on topics related to visual design in digital and print forms. Her publications include: *Graphesis* (2014), *Diagrammatic Writing* (2013), *The Visible Word* (1994), and *Visualization and Interpretation* (2020).


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

The Digital Humanities Coursebook

An Introduction to Digital Methods for Research and Scholarship

Johanna Drucker

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2021
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
52 Vanderbilt Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2021 Johanna Drucker

The right of Johanna Drucker to be identified as author of this work has been asserted by her in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
Names: Drucker, Johanna, 1952– author.
Title: The digital humanities coursebook : an introduction to digital methods for research and scholarship / Johanna Drucker.
Description: First edition. | Abingdon, Oxon ; New York : Routledge/Taylor & Francis, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2020044897 (print) | LCCN 2020044898 (ebook) | ISBN 9780367566661 (hardback) | ISBN 9780367565756 (paperback) | ISBN 9781003106531 (ebook)
Subjects: LCSH: Digital humanities. | Humanities—Research. | Digital media.
Classification: LCC AZ105 .D783 2021 (print) | LCC AZ105 (ebook) | DDC 025.06/0013—dc23
LC record available at https://lccn.loc.gov/2020044897
LC ebook record available at https://lccn.loc.gov/2020044898

ISBN: 978-0-367-56666-1 (hbk)
ISBN: 978-0-367-56575-6 (pbk)
ISBN: 978-1-003-10653-1 (ebk)

Typeset in Sabon
by Apex CoVantage, LLC

Resources

DUBLIN Core https://dublincore.org/.
JSON (vs. XML) www.w3schools.com/js/js_json_xml.asp.
KML https://developers.google.com/kml/documentation/kml_tut.
Library of Congress Classification www.loc.gov/catdir/cpso/lcco/.
MARC Records www.loc.gov/marc/bibliographic/.
Mark-Up Languages (Wikipedia) https://en.wikipedia.org/wiki/List_of_XML_markup_languages.
Metadata Standards, European Union www.dcc.ac.uk/guidance/standards/metadata/list.
SGML www.w3.org/TR/WD-html40-970708/intro/sgmltut.html.
www.w3.org/MarkUp/SGML/.
TEI https://tei-c.org/.
www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.
Weather data and metadata standards 19th and 21st century versions https://mrcc.illinois.edu/data_serv/cdmp/cdmp.jsp.
www.almanac.com/weather/history/CA/Los%20Angeles/2018-07-02.
XML https://tei-c.org/release/doc/tei-p5-doc/en/html/SG.html.

5 Database design

5a Database basics

We have already seen how structured data can be made and managed in a variety of formats: spread sheets, XML files, JSON, and so on. But as data becomes more complex and relations among elements of data more elaborate, neither flat (spread sheet) nor hierarchical (XML or JSON) structures are efficient for managing or processing. The shift from a flat database to a relational one is conceptual as well as technical. Understanding when and why such a shift should take place, as well as how to implement it, is crucial to making a relational database. Very simply, a relational database is composed of multiple tables that separate different types of information to make it more efficient to manage and easier to control. This is true at any scale, but becomes more pressing in larger projects.

The phrase "big data" is bandied about constantly, and it conjures images of nearly infinite amounts of information codified in discrete units that make it available for analysis and research in realms of commerce, medicine, surveillance, population research, epidemiology, and political opinion, to name just a few domains. To reiterate what has been stated earlier, all data starts with decisions about how it is made. Data does not exist in the world. It is not a form of atomistic information waiting to be counted and sorted like microorganisms in a water sample or cars on a highway. Instead, data is made by defining parameters for its creation—even the category of "cars" would need to be defined, as would microorganisms (count only living ones or also dead?). What about words on a page? Are all words to be counted (including "the" or "and") or only some? What is significant? Nothing is as simple as it appears to be at first.

Data can be big without being complex: a massive number of entries in a very simple spread sheet might list all 300 million people in America by age, or, by first name, or by height and weight. Data can be deep without being big: a spreadsheet might contain highly detailed information about a small number of entries—a hundred books and their authors, characters, plots, sentiments, vocabulary, sentence length, editions, page numbers, annotations

by multiple readers, materials of production, reception histories, and many other details. In both of these instances, the data could be accommodated readily in a spreadsheet. Each entry is linked to a single unique identifier that does not change, but in the second instance, a great deal of redundancy might appear, and in the first, age will change each year on a particular date. How should these matters be managed?

Keeping information consistent and error-free is a challenge for updates and other management tasks. The key to understanding relational databases is to think through the connection of one kind of information to another. Age changes by date, but birthdays are attached to individuals. An address for an individual might change multiple times and also be used by others. A work address might change without affecting a residential one. These might vary independent of age. Some might be in a unique relation to an individual (we share our birthdays with others, but each of us has only one). While these are unfamiliar concepts at first, examples should make this clearer.

In a designing a database, distinguishing "dependent" and "independent" relations is essential. In statistics, an independent variable is considered a cause and a dependent variable an effect: change the number of chickens in a room (independent) and see how many eggs they produce (dependent). For the purpose of structuring humanities databases, this insight helps determine what items should be in which table. Information that varies "dependently" (birthdays and individuals) should be grouped. Information that varies "independently" (residential addresses or car registration) should be in separate tables. The database construction described here is focused on consistency and efficiency for entering, managing, and querying data.

From flat to relational: an example

A spreadsheet is a single table. A relational database is a collection of tables connected to each other. One of the main reasons to break information out into multiple tables is to make it easier to manage—and to keep consistent. If I have many authors and the authors have many books, keeping the spelling and other biographical information in an "Author" table and the "Book" information in another allows them to be linked. The author information can be changed or updated in one place and therefore be less at risk of error introduction. This approach also keeps the database smaller and more efficient for queries, particularly in very large data sets. The programs that manage relational databases are called, believe it or not, RDBMS (Relational Database Management Systems). These systems manage the functions and operations of the database. They are programmed applications that build in the most common functions needed by users.

Here is an example of data in a spreadsheet format:

Table 5.1 Spreadsheet and tables example

(a) SPREADSHEET: A flat structure (every entry is one row with values in column).

Artist	*Birthdate*	*Title*	*Medium*	*Date*	*Owner*
Picasso, Pablo	1881	Portrait	Oil	1914	MoMA, NY
Matisse, Henri	1869	Dance	Collage	1922	MoMA, Paris
Carrington, Leonara	1917	I dream of you	Oil	1946	SFAI
Carrington, L.	1917?	The mask	Gouache	1951	Chicago Art Inst.
Matisse, Henri	1869	Red room	Oil	1907	Sotheby's
Tanguy, Yves	1900	Landscape	Oil	1932	SF MoMA
Walker, Kara	1969	Their house	Acrylic	1992	Whitney
Mitchell, Joan	1925	Abstract #2	Oil	1963	Private owner
Kandinsky, Wassily	1866	Composition	Oil	1917	SFAI
Tanguy, Yves	1900	Landscape	Oil	1937	High Musuem
Carrington, Leonora	1917	Dreamscape	Oil	1949	MoMA, Paris
Matisse, Henry	1869	Still Life	Oil	1923	Picasso Museum
Carrington, L.M.	1917	Interior	Collage	1951	MoMA (on loan)
Picasso, Pablo	Self-Portrait	Oil	1954	Met, NY	
Kandinsky, Vassili	1866	Landscape	Oil	1924	Tate, London

(b) TABLES: What information belongs with what and what are the connections?

ARTIST	*WORKS*	*MUSEUMS*
Last name	Title	Name
First name	Medium	Location
Other names		
Year of birth		

Note that in the spreadsheet, variant spellings of the artists' names appear. A report generated on work by "Carrington, Leonora" would not return all of the entries relevant to this artist. In the table below, all the information about the artist is specified and associated with a unique Artist ID. Because the information is specified in the Artist table, the ID is considered a *primary key*. But it can be cited by any table as a *foreign key* and used consistently throughout. Any changes made to the information about the artist are made once and only once but become part of the entire database.

Other changes or modifications might be necessary in the tables above. Works can be listed by title—but how will similar titles by the same artist and owned by the same institution be distinguished? And is Museums really



Figure 5.1 Table structure (NS)

the right definition for owners? Should museums and owners be separated? A painting on loan from a private collection might be on display in the Museum of Modern Art, but it is not owned by the Museum. For purposes of permissions, insurance, and other logistics, this distinction is significant.

When these tables are translated into database tables, each artist entry will have a unique ID, probably a number, as will the works, museums, and owners. These IDs will be cited as the foreign keys in order to keep the information completely unambiguous, as in the tables shown. So, a citation for a painting will have an Artist ID, a Work ID, and an Owner ID. Should Museums be kept separate since they might be the site of display for a work even though they do not own it? If so, then the Work table would need a field for the Museum foreign key, as here, so it could point to the institution.

The value of the primary key is defined in the table to which the information belongs. One could make an entire table of years and point to them as the year of birth with a key and also use the same date table to reference the date of a work. But then we are conflating two kinds of dates—birth year and the year of a work. A query searching for birth year would not be able to distinguish the dates in a table. The concept is simple: primary keys define values that can be cited in any other table. This keeps the information consistent and allows it to be managed in one place. This concept is crucial to the function of a relational database, just as figuring out what belongs in what table is crucial to the design of the database. Relational databases can have dozens of tables, even more, depending on the complexity of data being managed. One feature of most RDBMS is a view that shows the tables and their relations in graphical form.

More advanced topics in database design build on these fundamentals. An "association table," for instance, can be used simply to link keys to each other. When a database has hundreds of tables and millions of entries, these structures are essential. But databases, like any file, are vulnerable to hacking, corruption, and obsolescence. Databases containing personal or sensitive information need to be secured behind firewalls and with passwords and other security measures. User roles can be assigned, so that only certain entries or tables can be accessed—allowing someone to look at names, for

instance, but not financial information. But the preservation of data for research purposes also needs protection. The records of oppressive regimes, data collected about stolen and appropriated intellectual or artistic property, records of disappeared or disenfranchised persons, etc.—all of this information is part of the history of human culture whether it is in analog or digital formats.

Database design process

Whether you build a database in a software program like Access, Filemaker, MySQL, or any other DBMS, the principles just demonstrated are essentially the same for all relational databases. A relational database consists of tables of information related to each other.

Workflow for designing a database:

1) Design the content model (defining the information in the database and its organization), define the fields for data entry, and design the relationships.
2) Next, build a form-based entry for putting data into the database. This might be organized very differently from the database in order to make it more useful or coherent.
3) Manipulate and use the data through searches/queries, reports, and other methods, including visualizations, reports, and analyses.

Other forms of database structures exist, such as object-oriented databases. In a relational database, the data is stored separately from operations or functions. In object-oriented databases these might be combined.[1] The contents and functions of databases are all computational; data are and must be machine readable at the basic level where they are stored. RDBMS do this behind the scenes where, for instance, datatype is specified (character data, integers, Boolean operators (or in other systems URLs), binary strings, and shape files or other data formats). With databases, as with other systems and components, the decisions about which format and what design to use will depend on the research agenda.

The basic principles of database management and design are modularity, content type definition or data modeling, and relations, and then the combinatoric use of data through selection and display. Databases are powerful rhetorical instruments. A database is a logical structure and it is configured in the computer. Configuring the database is an intellectual task and has to be done to suit the needs of the project to which it is being put. Databases exist for years, even decades, and may have multiple contributors and innumerable users.

Particularly in institutional contexts, considerations of security and authentication are factors in database design. Most beginning digital researchers are not going to build a database from scratch, but instead, will

use a program like Filemaker Pro, Microsoft Access, or an open source or non-proprietary version. To reiterate, all follow the same procedures: modeling the tables (fields and content type) and relations among them, creating forms for data entry, and performing queries and creating reports. [See Exercise #1: Modeling a database.]

Collections management and content management

Content management systems and collections management systems are both common in the management of cultural heritage materials. Collections management is just what it sounds like—a way to describe and manage collections. The tasks for collections management include everything related to tracking objects or artifacts, making records (metadata again!), and keeping updates for acquisitions, conservation treatments, loans, use, and so on. By contrast, a content management system is designed to create and present content, often in a web-environment. Both of these systems have database back ends, and common content management systems include Omeka (created by humanists), WordPress, and Scalar (also designed for humanities and media work). These will be discussed in greater detail in the Web Presentation section. The point is that in both cases, the back-end organization of data is managed in a database.

Table 5.2 SQL Query Note that the SQL query is the information that follows "SELECT" in the lines below.

```
// Let's see what's in the Artists table.
SELECT * FROM Artists;
// Let's say you want to know all artists in the database that are from a particular
   location.
SELECT * FROM Artists WHERE Location='California';
// Numerical data can be filtered or calculated (SUM, AVG, COUNT)
SELECT * FROM Works WHERE Price > 10;
// Dates on the other hand need to be treated as strings, denoted with quotations,
   as long as the data is consistent.
SELECT * FROM Works WHERE Date BETWEEN'1978–01–01' AND'1996–
   01–01';
// Note the similarity in structure between this and and Location query.
SELECT * FROM Works WHERE Medium='Ceramic';
// Join connects two tables based on the primary and secondary key between two
   tables. For example, this query provides the artist first and last name from the
   Artist table and joins it with the work title from the Work table. The two tables
   join on the Artist ID, which is the primary key in the Artist table and foreign
   key in the Work table. Note also that SQL queries can span multiple lines; only
   semicolons indicate the end of a SQL command.
SELECT Artist.ArtistID, Artists.FirstName, Artists.LastName, Work.Title FROM
   Artists
INNER JOIN Work ON Artist.ArtistID=Works.ArtistID;
```

Whatever platform or application is used to create a database and manage it, the crucial intellectual issues are expressed in the data model. The conceptual feature of relational databases that differentiates them from flat spreadsheets and hierarchical XML structures is that they consist of linked tables. The information in these tables is referenced by other tables and can be called on in queries and in making reports. The conceptual structure and identity of the data to be stored in the tables is crucial.

Tables, forms, reports

The basic components of a relational database are somewhat more complicated than those of a spreadsheet, but they are conceptually similar. A **table** structured to store information. A **form** is a screen designed to enter information into the database. It is the front end of the tables. Multiple **forms** can be built on top of the same tables. A database of bibliographical information might have a front end in which readers put comments, or rankings, or add keywords or tags. But another form might be reserved for researchers who are establishing the biographical details of the same authors in a consistent manner and linking them to various authoritative biographical databases and what are known as "name authority files" that help disambiguate one author from another. **Reports** are the output of queries to the database. A user might ask what an author published in a particular year, or with a particular publisher. If a user is going to be able to query across a span of years, then the database has to be set up to accommodate this or the user will have to query each individual year. A database management system allows the database to be used for generating queries and reports, visualizations and displays, export functions, and other operations, including being accessed live on the Web. The most common database language is SQL, which stands for Structured Query Language.

Database and database management systems are intellectual instruments that pass as technical ones. Is there such a thing as a "feminist" table? Or a "racist" form for data entry? Can format embody bias, or only content? The structure of data is always fraught with values and the very act of defining categories can create exclusionary results. Data collection is often shaped by biases and data can be constructed to support almost any position on the ideological and political spectrum. Lack of differentiation or nuance can be equally damaging. Issues like national identity are problematic since countries, boundaries, and their names change over time, in some regions more often than in others. The breakup of the former Yugoslavia, for instance, created new independent states. Recognizing the historicity of categories and considering ways in which these details can be incorporated into the data structure is important for ethical reasons as well as for research purposes.

The power of relational databases resides in the ability to control information carefully for consistency, but also, for analysis across fine-grained features of the data. Attention in the design phase to what pieces of

information are dependent on which should be carefully tested against scenarios for future query and use. This can be done on paper without ever building anything digital, and should be. Once the database construction is underway, tools within the applications, such as visualization of tables and relations, will be extremely helpful in seeing the shape of the conceptual model and the argument it contains. Tools for building databases with visual means exist as well, for those whose logical faculties work better in graphical form.

Takeaway

The construction of a database for a humanities research project requires analytic rigor and sorting out what belongs with what, and what relationships these entities—and information about them—have with each other. Keep in mind the basic advantages of relational databases are to reduce redundancy, increase efficiency, and decrease errors. Analyzing your project to understand what can be put into a field or a table, and how these relate to each other is the fundamental challenge. Then, defining identities that can be referenced logically will make the relations logical as well. In spite of all of this formal structure, however, a database should be understood as an act of interpretation that makes a rhetorical argument.

Exercises

Exercise #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?

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.

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

Database design, as outlined above, involves decisions that are conceptual as much as technical. Every step of implementation involves choices that structure values and arguments into the relations in the database. These decisions may incorporate biases within the very infrastructure of the technology and

its implementation—but they are also opportunities for transformation of knowledge and its representation (Masters 2015).

Other issues also arise with humanistic use of data. Some involve high-level questions about what kind of rhetorical and/or intellectual structure a database embodies. Is it linear? Narrative? Combinatoric? Why does this matter in humanistic fields where arguments have tended to follow certain conventions? The question of whether databases offer *new* forms of argument, are only being used to support traditional arguments, or are not arguments at all is much debated.

Other concerns come from the use of what is called "legacy data." This is data that was created by someone else (individual or institution) and may come with little documentation or rationale. The decision about how it was made (gathered, cleaned, and edited) and what this process left out or changed may not be recoverable. We are left with what may be a consistent data set, but may also be a highly flawed and biased one.

The re-use of data raises ethical as well as intellectual questions. Whose data was this, what permissions were required for its production, whose intellectual property is it, and what privacy concerns or intellectual property issues might its use raise? How is it to be standardized to make it FAIR: "Findable, Accessible, Interoperable, and Repurposable," the four principles articulated by one European scholarly organization?[2]

Data that has been created through a specific research project that is particularly dense and rich may not be usable if it does not provide access points for query. The design of the database and the implementation of user, front-end, methods of search and retrieval are separate processes. Structuring an interface to a database creates a secondary argument—also shaped by assumptions about who the user might be and to what tasks or purposes the project might be put. A simple example of making a database more useful might be replacing a search box that is simply a free text field with one that is linked to controlled vocabulary in a "pick list" that orients the user to terms in the data fields. Particularly in areas where the data has been carefully structured according to specialized expertise, this will offer a way into the contents.

Finally, no single database design fits all projects and even a single project might need multiple modes of access for different communities. Levels of permission, decisions about display, questions of who has what role in an ongoing project, how to secure the project from malicious attack or spam—these are other issues for projects that have a public face and invite participation.

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

(Computing History Organization's History of Database n.d.). 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.

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 commercial 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 on an Apple computer. 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.

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. In the North American context, this argument received considerable attention when the publication of the Modern Language Association, the main scholarly organization for humanists in literary studies, dedicated a special issue of its publication to the topic. This 2007 issue of the PMLA (*Publications of the*

Modern Library Association), pitted the two formats—traditional narrative arguments and database structures—against each other. The approach generated much controversy.

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 discourse, 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, understand digital media and its specificity but also its effects.

The PMLA discussion led off with a statement by a scholar named Ed Folsom, who was co-Director of an important project, the Walt Whitman Archive. Folsom asserted that databases were a "genre" that promoted access to materials. Prior to the implementation of the Web, archival materials like original manuscripts, letters, photographs, and other primary documents were kept in institutional collections that could never be aggregated. Scholars had to visit these collections in person, and even copies or facsimiles of original materials were difficult and expensive to maintain. The transformation of scholarship in a project like the Whitman Archive was that these materials could be accessed through a single project where the database structure pointed to resources in other institutions (or, more rarely, stored digital copies as assets on a single institutional server). The database had become a familiar humanities tool, and the intellectual work of building and creating this framework provided a novel experience for humanists used to close reading, note taking, and prose arguments.

Folsom's opposition between database and narrative was accompanied by multiple responses. Some scholars appreciated the constrained and organized nature of databases as an interpretative and critical act but stopped short of seeing them as narratives. Others stressed the linear character of narratives, even in scholarship, and put them into opposition with the spatial structure of databases. Scholars compared databases to the traditional index cards and notes that track the research but disappear from view in

a final article or book. In the database, the materials remain segmented, structured, and also defined and described by the very processes described earlier.

Massive databases that provide access to scanned and marked-up materials now exist but using them without any narrative framework can prove challenging. A visit to the Digital Sanskrit Buddhist Canon online demonstrates this point quite dramatically, as does a look through digitized medieval manuscripts in the British Library. Without an interpretative frame, the materials appear in their original language, sometimes with metadata, often not, and are difficult to use for anyone without expertise. Data—whether in the form of original documents or their description—is an addition to and not a replacement for other forms of scholarship. The same is true for metadata—it adds whole dimensions to the scholarly endeavor that add value as well as producing controversy because of the legacy of classification.

Legacy data and its challenges

While many humanistic records of great antiquity can be considered "data," they have to be remediated to become digital data and computationally tractable. Examples of such "legacy" data sets are the Roman census records, inventories of monastic libraries, and the lists of settlement records in Iceland. (See the Coda for a brief discussion.) These can be—and many have been—put into digital form for study and analysis. Another type of legacy data would be of the sort made digitally for an unrelated purpose or project that is being used for research. This will be dealt with in the "repurposing" section.

King's College London, which has long been involved in digital humanities research, took on the challenge of assessing the data inherited from early three decades of projects. While this does not sound like a long time, the period has witnessed considerable changes in practices of digital methods (Smithies et al. 2019). The group identified about a hundred "legacy projects" that spanned "a wide spectrum from text analysis and annotation tools, digital corpora of texts, images and musical scores to digital editions, historical databases and layered maps" (Ciulla 2020). The team decided to collect considerable information about the projects—team members, funding, period of time in which the project took place, standards used, and other details. Among the difficulties addressed were the security breaches that had occurred on "unmaintained applications" (Smithies et al. 2019). Another issue was managing expectations of various team members, particularly in situations where partners from different countries and institutions were stakeholders. Preserving data and metadata in a variety of formats is expensive, but without the original interface, it may be illegible or unusable. Some of these concerns will come into focus again in the discussion of sustainability.

For the researcher, using legacy data has other concerns. Data made for specific research projects is often idiosyncratic and may not be documented. The quality control on data production may have been uneven (Wuttke 2019). Legacy data issues are particularly complex in medical humanities and social sciences, fields where the possibilities of violating original agreements about collection of information might be at risk of being broken in new projects. Any identifying information that can be "re-engineered" could put human subjects at risk or violate their privacy. In addition, monetizing any data in a use for which it was not originally intended poses ethical issues. Finally, legacy data often do not reveal the process by which they were modelled. What was eliminated in the process and what was conflated? Data have great potency and appear to have an authority as statements of facts, and so knowing their source and history is crucial.

Many cultural institutions produce elaborate datasets from which interesting research projects can be built. For instance, the New York Public Library, the Library of Congress, most museums, and the data.gov sites are rich resources. These can be made use of for humanities research by posing certain kinds of questions. The advantage of proceeding from data to humanities questions, rather than trying to make data to address a research problem, is that the data already exist. [See Exercise #1: Asking humanist questions.]

Metrics and standards

Metric standards have their own strange histories. We know that inches and centimeters are human-created standards for measure of space and dimension. But a year has a relation to a natural cycle of motion around the sun, as the day is determined by the turning earth. But what is the means by which a "minute" is determined, or a day broken into hours? Are all hours the same? Medieval monks had a system for dividing the day into twelve hours of daylight and twelve of darkness throughout the year. In summer the daylight "hours" were longer than in winter, and vice versa, but the division of units served their purposes. If we are transcribing the record of activities from a monastery in this period, how do we reconcile these differences with the standard measures of time we are accustomed to using?

Temperature data seems to be empirically derived, based on the thermal condition of phenomena under investigation. But the Fahrenheit and Celsius scales have very different units. The Fahrenheit scale is an idiosyncratic scale, rooted in the experience of the man who designed it. He defined the low end as the coldest temperature taken in the town where he lived and the midpoint as his body temperature and the high point as that at which water boils. This was later refined and made in a more precise system, but that a standard metric was created with a human reference point—he had a slight fever when defining the precise body temperature—is remarkable.

In an important sense, all metrics share this characteristic—they are created in reference to human experience—but they function as if they are value-neutral and universal.

Among the considerations that need to be brought to bear on repurposing data are the many ways it can be assessed. The need to support data-driven research has become a major topic of discussion in the library community (Padilla 2016). In their role as information professionals, librarians have been attentive to ways in which standards for use and assessment can aid researchers in both creating and sharing digital assets. Their contributions to this conversation are essential. Similarly, data professionals in areas of scientific and humanistic research are well aware of the ways in which data can be tampered with and misused. The larger challenges of missing data, partial data, and inaccuracies can never be fully managed since, by definition, the human record is constructed only from what remains.

Restricting access

The assumption that all data should be available, and that complete transparency is optimal, may not apply in all cases. (See Chapter 10 for discussion of access.) When the web presentation platform Mukurtu was being developed for use in digital projects that contain Indigenous knowledge, the rationale for the platform was that not all communities wish their cultural heritage information to be fully available to all viewers. Within many cultures, what can be accessed depends upon the role, class, position, age, or gender of individuals—or the caste or kinship group to which they belong. Taking such considerations into account with regard to data means shifting one's assumptions away from a single approach and recognizing the degrees of granularity that might need to be built into control of data retrieval, access, or use.

Further constraints on data access arise from other ethical considerations. Students should not have access to other students' records and confidential information should not be shared in personnel and health matters. Personal information should not be shared without consent. And individuals' data should not be monetized without their being informed. But other forms of data may be sensitive as well, and researchers dealing with controversial materials should be attentive to the harm these can do, or the risks of their being misappropriated.

Debates about digitization include its capacity to increase disparity and to level it. Does access to information promote inclusion? Does control over the creation of online resources replicate patterns of exclusion long in place? The tensions between Global North and South, for instance, around preservation of cultural heritage, have been a regular concern for UNESCO (Kulesz 2016). Their discussion of these matters takes into account that fundamental differences in attitudes about cultural production need to be

recognized, rather than simply applying a single approach across communities. Questions of rights for intellectual property are a matter of negotiation and cannot be ignored. Data structures that build different levels of permission into their design recognize the need for decisions to be made with informed understanding of the ways cultural assets might be appropriated once they are digitized.

Takeaway

In the day to day creation of data sets and databases, theoretical questions are not generally asked. Instead, they are implemented using standard metrics, categories, classification systems, and spreadsheets to make databases. Databases come in several forms, flat, hierarchical, relational, object-oriented, graph, and so on. Databases can be described by their contents, their function, their structure, or other characteristics. But questions of ethics and use arise around legacy data and its repurposing, as well as around access and retrieval by asking if these should be assumed as universal.

Exercise

Exercise #1: asking humanist questions

Go to 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 FarmPlenty be useful for humanistic research? What other information would you want to link to this dataset once you have formulated your project?

Notes

1 In object-oriented databases, a set of behaviors is merged with the data into a single object.
2 For a complete discussion of issues and recommendations by ALLEA (All European Academies), see: www.ria.ie/sites/default/files/allea_sustainable_and_fair_data_sharing_in_the_humanities_2020_0.pdf

Recommended readings

Christie, Michael. n.d. "Computer Databases and Aboriginal Knowledge." www.cdu.edu.au/centres/ik/pdf/CompDatAbKnow.pdf.
Nodegoat. 2017. "Formulating Ambiguity in a Database." *Nodegoat.net*. https://nodegoat.net/blog.s/21/formulating-ambiguity-in-a-database.

References cited

ALLEA (All European Academies). 2020. www.ria.ie/sites/default/files/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*. www.kdl.kcl.ac.uk/blog/legacy-project-datasets/.
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.
Computing History Organization's History of Database. n.d. www.comphist.org/computing_history/new_page_9.htm.
Kulesz, Octavio. 2016. "Intergovernmental Committee for the Protection and Promotion of Diversity of Cultural Expressions." *UNESCO*. https://en.unesco.org/creativity/sites/creativity/files/sessions/10igc_inf4_the_impact_of_digital_technologies_octavio_kulesz_en_0.pdf.
Masters, Christine L. 2015. "Women's Ways of Structuring Data." *Ada: A Journal of Gender, New Media, and Technology*. Issue No. 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.

Resources

British Library www.bl.uk/manuscripts/BriefDisplay.aspx.
Digital Sanskrit Buddhist Canon www.dsbcproject.org/gallery.
Mukurtu https://mukurtu.org/.
Noun project https://thenounprojects.com.