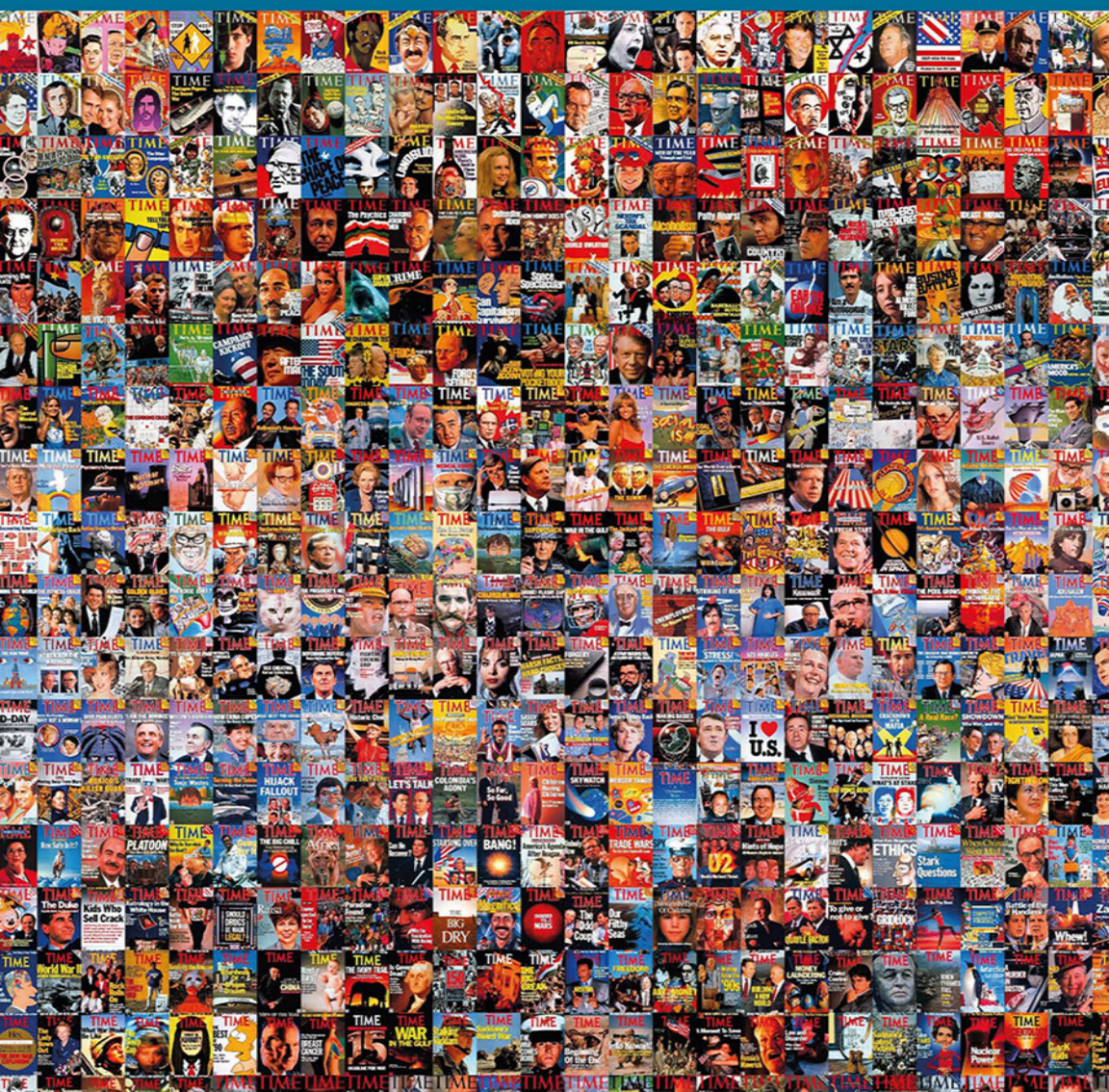
The Digital Humanities Coursebook

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

2 Data modeling and use

2a Making data

The term *data* is so ubiquitous that it seems to refer to information in any form. The word *data* derives from the Latin word *datum*, which means *given*. This suggests that data exist in the world. But in fact, data are made, produced through a process of abstraction that is called *data modeling*. Every stage of this process of making, using, presenting, or repurposing data raises ethical issues related to privacy, ownership, and use. In this section, we will look at data production practices, including appropriation of existing data, to understand fundamental concepts and concerns.

The discussion of data conjures images of columns of numbers and quantitative information. But what does it mean to talk about data in the context of the humanities, where images, texts, dance, music, and monuments might be the focus of research? Quantitative data can be produced from a wide range of sources, not just numerical ones. Anything that can be counted can be used to produce *quantitative* data. The number of words in a text, beats in a measure, pixels on a screen, and distances on a map can all be turned into quantitative data. As already mentioned, these are acts of remediation by which some feature of an already existing phenomenon is abstracted into a value. But what about more subjective values such as rating the actions of a fictional character as "good" or "bad?" Can such assessments be kept standard across different readers, data producers, or even artifacts? The data created from these interpretative judgments are termed *qualitative*.

Humanities "data" are produced by taking documents or artifacts and abstracting from them certain features that can be represented explicitly. Humanists also extract data from the records kept about those artifacts—such as library and museum catalogs or estate and auction catalogs.

Structured and unstructured data

A distinction is made in digital work between structured or unstructured data. Structured data is composed of entities that are explicit, discrete, and unambiguous—like numbers or true/false statements. Unstructured data, like natural language, is sometimes ambiguous and unclear—like a picture of a mother holding a child which can be described in many ways.

Typing on a keyboard "structures" a text as an alpha-numeric (ASCII and/or Unicode) file that can easily be processed as data. The distinction of one letter from another or from a number structures the data at the primary level. But the words that the text contains might still be ambiguous or vague.

Data structures can also be expressed through formal organization. A table, for instance, is a data structure that determines the values of the information according to their position. A spreadsheet is a table (or collection of tables) that is a common way to structure data. Data in a spreadsheet can be saved in specific formats such as. xsl (Excel),. csv (comma separated values), or. tsv (tab separated values). These are all forms of structured data that can be easily imported into and exported from a spreadsheet and used in other programs. Other ways of structuring data will be addressed ahead, such as markup, or use of other file formats designed for particular fields or tasks. Most programs that deal with data will allow save or export functions to produce a variety of standard outputs.

The distinction between structured/unstructured data has ramifications for the ways that information can be used, analyzed, and displayed. The term unstructured data is generally used to refer to texts, images, sound files, or other digitally encoded information that has not had a secondary structure imposed upon it. Analog objects such as literary texts in print or manuscript and historical records on cards or ledgers are generally unstructured data, though they often contain features that can be translated into structured data. The distinction between a poem's title and its stanzas, or the addressee and recipient of correspondence can be specified explicitly. Structured data has an explicit form that has been deliberately designed, generally for a specific purpose (such as managing financial records or keeping track of populations). Unstructured data has implicit form but may not be organized in a standard way. Structured data can be processed and manipulated more readily than unstructured data, but each has features that are of interest and value.

To illustrate the difference between unstructured and structured data, look at these two examples:

Table 2.1 Structured and unstructured data

Unstructured data:

Among the Doe family members, Jane was best known for her affinity for Washington and all things related, including the mementos and archives, some of which had been equally dear to her grandmother, who had cherished her family history, half a century earlier.

Structured data:

Family	Children	Location	Birth year
Doe	Jane	Washington, DC	1978
Doe	John	Seattle, Washington	1982
Doe	Jane	Eternal Rest cemetery	1908
Doe	Washington	Plymouth, MA	1777

In the first sentence of the unstructured example, how would the word "Washington" be identified as a geographical location versus a family name? The context of the word in the sentence suggests that it refers to an ancestor, not a place. But as we see in the structured data example, the word becomes explicit when put into a labeled column. Note, however, that the distinction between city and state is not made clear in the way the table is organized. The *Location* column contains several kinds of information and it is not standardized by type or format. This kind of ambiguity in structured data formats is resolved by data cleaning, to be discussed ahead. With training, AI engines can distinguish place names from personal ones and are used to parse (automatically analyze) unstructured data. When designing structured data, disambiguation and clarity of data type and format should be kept in mind.

Reading a data *model* (as opposed to the specific data) means analyzing the categories into which the information has been organized. Even the small sample of structured data offered above suggests judgments have been made about what information to include and how to classify it. The term *Location* in the third column could mean a place of residence, or birth, or activity. By their proximity, *Location* and *Birth year* appear to be correlated. This suggests that Jane Doe was born in the Eternal Rest cemetery, which would be unusual. Many categories, such as racial identity, gender, nationality, and religious or political affiliation, are specified according to categories that are biased. If an artist was born and raised in Spain, worked in France, but maintained Spanish citizenship, what is their nationality? Are they still Spanish even if they have neither citizenship nor passport but feel a strong cultural affiliation? If someone was active in the Communist Party for ten years, but not before or after, what is their political affiliation? Human lives are of long duration and our identities are not singular, uniform, or consistent. Crucial information about family relations, kinship ties, and other important facts are not present in the table above. Modifying the data model would be necessary if this were a research project about the Doe family and their archives.

New and/or existing data

Digital research in the humanities can make use of existing data or create new data. These have different workflows to them, and each requires some critical attention.

The use of existing data allows analysis of records and materials in a format that lends itself to digital tools. For example, if we take an existing inventory of a museum collection, it will be well structured with categories of artist, location, date, medium, and dimensions. Research questions can be formulated on the basis of this data. We might be able to track certain trends. When were larger or smaller canvases painted? Are they associated with certain artists? Do Western and Asian patterns diverge or run parallel

to each other? Are gender or age significant factors in the data or merely incidental?

Work with existing data raises immediate questions about how complete it is or who made it. In the example above, does the inventory reflect the collecting patterns of a particular individual or institutional collector? Can it be trusted for insight into trends in painting over time and across cultures? Painting practices in one cultural and historical location might have been site specific so that works that could not be detached from walls and sold and might not be part of this inventory. Prohibitions against the removal and sale of cultural property might have prevented paintings from one region from being collected in another. Works by female artists might not have been identified under their own names but attributed to someone's studio.

Existing data provide information, but rarely the information about their own making, either as documentation or as metadata. The decisions that shaped the data so that it is presented as if it were a neutral statement of facts cannot be recovered easily. A useful exercise before making use of existing data is to analyze its *model*. What features of the phenomenon have been described explicitly? What might be missing? How are the fields named and what does the classification system suggest? False binarisms and categories of racial or gendered identity may be reductive, simplistic, or abusive in certain circumstances. All of these questions are part of understanding the *data model* of existing data—and of making new data.

Data modeling

As already noted, the creation of data involves modeling. This determines what will be identified as a feature, how it will be made explicit, and what format it will have. All data models embody values that carry implicit or explicit judgment and therefore often include biases. Almost all data are partial and represent some features of a phenomenon and not others. Policies of inclusion and exclusion operate to reify and reinforce biases, making them seem natural.

Here is a specific example of data modeling: begin with a classroom and ask what data is available in the situation. Anything that can be sorted, measured, or counted could be turned into data. The number of chairs, windows, desks, books, lights, pupils, or teachers present are all easily quantified, but some factors will vary at different times. If the data sample is taken at midnight, or during vacations or weekends, instead of during the school day, the population will be different. The model should account for the time of the sample and its frequency. The assumption that certain things actually "belong" to the classroom has been made in the initial list. But the books may be transient, folding chairs might have been brought in to handle an overflow attendance at a special event. These issues expose the fact that even the simplest-seeming act of data production involves a model. Our model might need to be modified depending on the purpose for which

we are collecting the data—safety, health, or usage—or to justify new furnishings. Does it matter whether the occupants are male or female, trans or queer, non-binary, and range in age from five to seventy? The specifics of the model for creating data in classroom assessment will depend on the purpose. [See Exercise #1: Creating a data model.]

When we are dealing with concepts or issues that are not immediately quantifiable, then data modeling becomes even more complex. For example, take the statement, "The discouraging climate in the classroom was not conducive to study, no matter how hard the young people worked." Determining the model to get data for assessing this statement is immediately complicated. If we identify the demographics of the "young people" in gendered terms, or in their racial or ethnic identity, we need to know if this is relevant information. The basic assumptions of sentiment analysis suggest that we can code terms as positive or negative—but according to what criteria is the "climate" described as "discouraging"? The interpretative dimensions of data modeling are evident at every turn.

When we are analyzing language, instability of meaning is built in. A word might mean something very different in context: "the mother of all storms" is not the same as "the mother ship" or "the weeping mother." A data model that looks for the term "mother" in a collection of texts would find all of these, but to group them into a single category and suggest that here there were three instances of the same word would be problematic.

Data models are interpretative acts. In standard approaches to data science, the tenets of "good" data practice are clearly spelled out. Data should be valid, accurate, complete, consistent, and uniform. (Elgabry 2019) But in the case of humanities data, this can be very difficult. The materials that remain from the past are by definition incomplete, sometimes ambiguous. Imagine assessing an 18th century collection of correspondence. Modes of formal address might be partial and only include a title, not a name. They might be inconsistent—sometimes dated and sometimes not. Missing data are particularly tricky, since they cannot be measured with any degree of accuracy (Onuoha 2018). Working with cultural materials to extract data requires recognition of these challenges.

Checklist for creating a data model

- 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.

When using existing data sets, one of the first exercises should be to "read" the data structure and analyze the model on which it is constructed. What cannot be represented in its scheme? In the example of the classroom above, no mention was made of seating arrangements and preferences, but for a sociologist, such information would be crucial. Different data would be required to assess how pupils chose to position themselves and what information about social identity could be deduced from this.

Parameterization

Data models depend on the concept of parameterization, or what can be measured. Some features of texts, images, recordings, and cultural phenomena lend themselves to parameterization: the number of words, the size of a canvas, the length of a song, the height of a building, or the shape of a pot are easily turned into metrics. But within cultural artifacts, many features that are fully legible to human perception are difficult to parameterize. For instance, we might be able to measure turn-taking frequency in a conversation, but how would we put a metric value on tone of voice, posture, facial expression, or other non-verbal aspects of the exchange? How are such assessments standardized from one researcher to another? We might be able to measure how many minutes a character is on a stage or a screen, but does it matter whether they are seen close up, from the back, or at a distance?

The problems in parameterization extend those of the data model and its value systems (and biases) into very real challenges with implementation. Parameterization comes down to the ability to distinguish something that can be counted according to some metric (quantity, scale, size, direction, frequency, or other quantitative metric). Some forms of human expression, while fully legible to us, are difficult to parameterize. Information in images, for instance, does not lend itself to a simple system of analysis. Language is composed of discrete units (words) governed by rules (grammar and syntax), while images are not governed by such systems. All of these systems of representing cultural phenomena depend upon the position of an observer. How far is it from London to Paris? The physical distance might remain metrically the same from the 14th to 21st centuries, but the length of time required to cover that distance would vary considerably according to many factors (class status, economic resources, gender, weather, time of year, political conditions, size of the party travelling, mode of transport, and so on).

Parameterization always requires judgment about what can be measured and how. Data are generally created for a specific purpose—governing, analyzing, controlling, supervising, or simply managing. In addition, the question of who is doing the parameterization and for what purpose has to come

into consideration. From outside of a cultural or historical framework, these judgments may create data that are skewed to the point of inaccuracy, thus lacking any validity. Parameterization relies on knowledge of a discipline or field in order to produce meaningful data.

Tokenization as a feature of data production

Tokenization is a subset of parameterization and is the act of determining what units of representation can be identified in a corpus or artifact. Words, for instance, are easy to tokenize because even in a plain text they are separated by an identifiable element, the space. Again, in images this is a far trickier matter since a figure might be separate from a ground, but it might be partially hidden, incomplete, enmeshed in shadow or otherwise difficult to distinguish as a discrete form. What can be tokenized in sound files and recordings when they are analyzed as wave patterns, instrumental parts, and overlapping resonances?

Parameterization and tokenization both rely on judgments about what is significant and how a feature can be identified and extracted from a work, artifact, or phenomena. In the analysis of a place name, for instance, location will be understood differently in a gazetteer than in a travel narrative. The tokenization of "Washington" as a place, defined as a single word, will not suffice if what is needed is knowledge about whether it was a port of departure or a destination. For this information, at least two words will be necessary, "to Washington" and "from Washington" could be distinguished from "at" or "in" the same place. But this two-word tokenization would miss a phrase like "back to my beloved Washington" whose sentiments might be as significant as the mere fact of its being a destination. These points will be taken up in the discussion of issues like the inclusion or elimination of "stop words" and other matters in the text analysis section ahead.

Data vs. capta: a brief polemic

Given all that has been stated, what should be clear is that data are not "given" but made, or "captured" by a process of making categories and distinctions for measuring, sorting, or counting entities. For humanists, the challenge of finding and using existing data is considerable. One may take an existing data set, like that of a museum catalog or auction inventory and ask relevant questions of it that relate to humanities research. In that sense, the data serve the same purpose as any other primary text, as a document to be read, analyzed, and interpreted. But, as was evident in the earlier Scenarios section, in many cases, the challenge will be to make data, extract it from existing materials that are text-based or consist of other files of recordings, objects, or documents, and through the processes of modeling, parameterization, and tokenization create structured data with which to work. Working effectively with such data will be the subject of the next

section, but it should be remembered at every point that the "data" were not given, but "taken," and are actually *capta*, information captured through a process of abstraction and structuring that depends upon decisions that are already acts of interpretation (Drucker 2011).

Data ethics

Because data are always produced through an act of selection governed by decisions, they are necessarily the expression of a point of view and value system. They may not be intentionally malevolent. They are simply part of the process by which data are produced. Whether making data or using existing data, researchers need to be aware of the ethical concerns and also government regulations for use. Ethical concerns are broad and deep, but regulations tend to be local and more specific. From an ethical perspective, data collection has to be carried out in accord with the guidelines that assure that if human subjects are involved, the researcher knows the correct protocols for approval of the work (informed consent, work with vulnerable populations, anonymity, and privacy) (Floridi and Taddeo 2016). Any researcher should also be aware of intellectual property and the limits on cultural appropriation. Whose cultural materials are being used? Who has the right to grant permission for their use and/or publication? How were they acquired? Are there provenance issues and/or concerns that should be addressed in advance of use? [See Exercise #2: Ethics of legacy data.]

The production and use of data are regulated differently across nations and locales. The European Union has strong protections that require notification before data are collected. The collection, repurposing, and monetization of personal data is prohibited or limited in some areas. The California Consumer Privacy Act imposes more stringent limitations on data acquisition than most other states in the US. The "triangulation" or correlation of data sets raises serious issues with regard to individual privacy, since one or two pieces of information, once matched, can violate privacy statutes and erase anonymity. Certain kinds of data, such as medical records, are protected by particular statutes (HIPAA, the Health Information Portability and Accountability Act, and FERPA, the Family Educational Rights and Privacy Act, for instance), but again, these protections can be over-ridden when different data sets are combined. Studying the history of counterculture and underground activities, sub-cultural activities in which individuals participated pseudonymously or anonymously, for instance, can have unintended consequences if identities are revealed and connected in the present.

Data privacy is a critical concern when dealing with any information related to individuals, particularly vulnerable populations (undocumented, trans persons, children, and incarcerated individuals). In addition to understanding restrictions on gathering data or storage of data should be considered as part of research practice. Information stored on a local drive—such as desktop or a flash drive—may be less at risk of being hacked than when stored on a Cloud server. But information that is not backed up regularly is

at risk for other reasons (loss or destruction). Back-up procedures, copies, and firewalls for protection are all part of regular maintenance.

Checklist for ethical data use

- Do you know the source of the data—is it authentic and trustworthy?
- Are you aware of constraints on the use of the data (privacy or property issues)?
- Have you obtained proper permission for acquisition or use of this data?
- Does use of the data put anyone at risk?
- Is there personal, financial, or medical information in the data that can be identified through correlation with other existing data sets?
- Is the data biased, offensive, or a misrepresentation of a community or event?
- Are vulnerable populations represented in the data set?
- Document the source, data models, and production whenever possible.

Every individual researcher is responsible for understanding where the data they are using came from, how it was gathered, and what the restrictions are in use. They must work within the legal limits, but even more importantly, perform according to a responsible and ethical code of activity. Risks and consequences are real, and data collected in one circumstance may need to stay secure in order not to be put to a purpose that violates the rights of communities or individuals involved.

Takeaway

Data are the basic units of almost all digital work. They can be manipulated in many ways and also analyzed through a variety of methods, but data drive the world of digital research. This will be even more clear in the discussion of file formats, which are the standards through which remediation of analog materials and production of born-digital ones takes place. Structured data expresses a model of content and interpretation. Structuring data allows analysis, repurposing, and manipulation of data/texts/files in systematic ways. It also disambiguates (between say, the place name "Washington" and the personal name). Structured data is interpreted and can be used for analysis and manipulation in ways that unstructured data cannot.

Exercises

Exercise #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?

Exercise #2: 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?

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.

Elgabry, Omar. 2019. "The Ultimate Guide to Data Cleaning." *Towards Data Science*. https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4.

2b Cleaning and using data

Quantitative or qualitative data can be made useful in many ways, but one common way to store and analyze structured data is in a spreadsheet. Spreadsheets can be very simple in structure, consisting of just a few columns or rows of information, or they can spread to be enormous and contain elaborate formulae, functions, links, and operations. A spreadsheet is basically a table that can be dynamic (manipulated and changed) as well as static (a presentation). The grid format of the table dates back thousands of years to the ancient Near East and Babylonian system of accounting, mathematics, and management. The basic principle is entities/entries should be classified by type. This separation of entities by "content type" in the row and column structure keeps entities discrete, but also allows them to be manipulated for many purposes. Databases and spreadsheets are not just formal structures, they are intellectual ones expressing judgments and viewpoints through their organization and labeling.

For data to function effectively, a few conditions need to be met, some of which pose their own ethical and intellectual dilemmas. Data will need to be more or less (preferably more) uniform and standardized. For instance, dates should be given in a particular format. This sounds simple enough, but some dates might only be specific to the year, some to the month and day, others to an hour or minute. Methods of recording time and dates are often local, or historical. The start points for the calendar in Biblical, Islamic, and Mayan systems are all different. How are these dating systems accommodated? Simply subsumed into a modern western calendar? Standardization can erase specificity and cultural identity, so immediately, issues arise in making data uniform.

The basic functions of spreadsheets are data cleaning, making reports, and creating pivot tables for comparison of values. Turning data into charts and graphs will be treated later, in Chapter 6.

Terminology and practices

A good way to familiarize yourself with the vocabulary of spreadsheets: cell, row, column, range, function, report, and presentation is to work through a tutorial in any standard spreadsheet program (Microsoft Excel, for instance). In addition to learning this vocabulary, study the basic operations such as order, filter, combine, and so on. Skills for manipulating spreadsheets are essential for data driven work from the simplest to the most complex levels.

One important concept to understand and implement is that of the unique ID. This will be an alpha-numeric ID that is exclusive to an individual entity (like a personal name or publication). Imagine you are inventorying the manuscripts, correspondence, and memorabilia linked to publications by an author. The columns will be labeled as information fields, such as standard bibliographical information about publication date, publisher, length, edition, and so on. The author whose archive this is will remain the same, but authors of correspondence should be given their own unique ID. In the inventory, each item about which information is being entered should also have a unique ID. This will allow the materials to be identified efficiently and unambiguously (for instance, if there are three manuscripts for one book and four letters from one person in the same year, giving each a unique document ID eliminates confusion). The decision about how granular this should be depends on the project. For instance, if there are three manuscripts of the same book, each has a unique identifier and also some distinctive information about its place or date of production. But what about miscellany stuck into the manuscripts at random—a letter from a friend, a holiday photograph, a note relevant to travel arrangements. Should these have unique IDs? Are they separate documents? Or do they "belong" to the larger manuscript?

For whom is this information useful and how will it be used? The ephemeral documents are not, strictly speaking, part of the manuscripts. They may contain information that needs to be structured differently. A manuscript generally does not have a recipient or a sender, and correspondence does, for instance, and so representing this object in the same database (data model) as the manuscript raises questions. If a later process sorts all items in the spreadsheet by date, the additional items will need counting. If the postcard is from a friend, it will have a different author. The purpose of structured data is to make research and analysis possible without confusion. Consider this when putting information into the spreadsheet.

Functions and operations

The functional capacity of spreadsheets is enormous, but simple operations like sorting (on date, location, or other content type), counting, or filtering are fundamental. These operations are structured into spreadsheet software. But for these to work, the data have to be "cleaned" to some extent or entries that are not standardized will fall out of the report or visualization.

Filtering refers to the act of sorting data by characteristics and features. If a feature has not been added into the dataset, it cannot be used. If the manuscripts by an author contain locations that were part of the "North" and "South" in the Civil War, but that information has not been specified, the data in the author's work cannot be analyzed according to these criteria. If housing records do not contain information about demographics for evictions, then activists cannot assess and act on issues of potential discrimination. The way a category is specified will determine its use not only in making data, but in the way it can get repurposed—but also, reused out of context.

Pivot tables are methods of selecting and comparing information in the data set to perform analyses. Some comparisons make sense, but others do not. Pivot tables only work when the comparisons are sensible. Comparing the page length of an author's manuscripts to publications over time might make sense, though interpreting the results would require more than a simple statement to assess the many factors involved. (Was her work edited more or less closely? Did she mature as a writer? Were publishers willing to take more chances after certain marked successes?) But comparing page length to locations cited or to the number of postcards received from friends in a particular period of time might be meaningless. Reading correlations across data requires critical reflection.

Data cleaning practices, intellectual and ethical considerations

The very concept of "cleaning" data suggests elimination and purging techniques that should be approached with caution. Some forms of standardization may be engaged without violence to information, while others impose strong acts of interpretation.

In the example of contrasting date systems mentioned previously, the data might be preserved in their original form and format, referencing the cultural contexts in which they were produced. For instance, one column using a western standard format that includes year 0, but also CE and BCE dates can be used as a shared frame of reference. However, translating all dates into that single format could be seen as erasing cultural identity. When western dates are identified as CE, Common Era, rather than AD (Anno Domini), this shifts the framework from religious to secular references. But that does not change the fact that the birth of Christ is a dividing point.

Imposing these systems on non-Christian cultures imposes a culturally specific framework that is neither neutral nor universal.

The constant trade-off between preserving the specificity of source information and standardizing information for use is always fraught with ethical questions. Validity (is the data accurate and authentic) and security (has the data been altered or changed) are also critical factors. Professional data entry practices in many fields are designed to make data interoperable by standardizing its formats, and in library and museum practices, standards are essential if information from one cataloger is to work with that of another and if catalogs are to integrate across systems. Digital humanities projects often begin with strong convictions about the interpretative requirements of their study and the need for highly customized data and metadata (to be discussed ahead). A researcher might feel every individual named in a lifetime of correspondence, or specialized description of features of artifacts, or esoteric fan-based knowledge of a set of recordings are essential. But when the time for interoperability or cross-project functionality arises, the intellectual investment in elaborate data constructions is often jettisoned, sacrificed to expediency.

Designing data from the beginning of a project to suit research needs and anticipated uses makes sense. But overly customized data structures can pose difficulties for integration with other projects or repositories. Documentation is essential, particularly when categories have been designed specifically for an individual project. Working with artifacts, archives, and cultural materials requires keeping clear understanding of whose intellectual property these are and whose characterizations of their features are being produced. With whom will the data be shared? Where will it be posted or stored? Who is at risk from its characterizations? [See Exercise #1: Creating a scenario.]

Checklist for data cleaning

- Are all dates formatted the same way?
- Are all quantities specified explicitly?
- Is spelling consistent?
- Do fields contain unambiguously unique information, or do they need to be split?
- Is the vocabulary for qualitative judgments consistent?
- Is information missing or incomplete in many fields?
- Is there repetition of information?
- Will the information be able to be filtered and sorted accurately?

Spreadsheets vs. relational databases

A spreadsheet is a powerful instrument for intellectual work in a structured data format. But keeping consistency in a large-scale project becomes difficult. One of the solutions to this is to shift from a flat database structure

(spreadsheet) to a relational database. Relational databases separate information into tables so that information that varies can be distinct from that which stays consistent. These will be the subject of a section ahead.

Takeaway

Data models are interpretative intellectual tools for extracting information from artifacts or phenomena. They make use of parameterization (counting) and tokenization (what can be defined as a discrete unit) to produce quantitative or statistical information. Data may be qualitative as well as quantitative, and gathered with subjective criteria, but for purposes of processing, the data must be discrete, distinct, and unambiguous. It must be "machine readable" to be available to be processed computationally. Many data sets are created for a particular purpose or project, but decisions about the data model should be documented. The ethics of data gathering need to take into account possible re-use, discovery, violation of privacy, and liabilities for the repurposing of data for monetary gain. This is particularly true of data associated with individuals or identifiable groups. Data are never neutral, but always constructed from some point of view, with a specific set of values and beliefs.

Exercise

Exercise #1: creating a scenario

What kinds of data can a scholar working with Rare Book auction catalogs use to produce their research project? What materials will they need and how will they structure data from them? Create a scenario in which this occurs. www.rarebookhub.com/catalogues/

Recommended readings

Owens, Trevor. 2011. "Defining Data for Humanists: Text, Artifact, Information or Evidence? Journal of Digital Humanities." *Journal of Digital Humanities* 1 (1). http://journalofdigitalhumanities.org/1-1/defining-data-for-humanists-by-trevor-owens/.

Taylor, Jack. 2016. "Data Cleansing Pitfalls." *Invensis*. www.invensis.net/blog/14-key-data-cleansing-pitfalls/.

References cited

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

Elgabry, Omar. 2019. "The Ultimate Guide to Data Cleaning." *Towards Data Science*. https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4.

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

Onuoha, Mimi. 2018. "On Missing Data Sets." https://github.com/MimiOnuoha/missing-datasets.