The Hitchhiker's Guide to Data in the History of Science

Julia Damerow, *Arizona State University*
Dirk Wintergrün, *Max Planck Institute for the History of Science*

Abstract: Every project in digital and computational history of science starts with the collection of data. Depending on the research project, the subject of study, and other factors, data can comprise a variety of different types, including full texts, images, audio, video, and bibliographic metadata. Publications and project reports generally describe their results and the methods and algorithms employed, but few discuss the challenges of the initial data collection process or how it fits into the overall research data life cycle. This essay discusses a concrete research data life cycle and takes a look at the difficulties it involves. It also explores the strategies and challenges of data collection and the question of the comparability of datasets.

> ***Don't panic.***
> **—Douglas Adams, *The Hitchhiker's Guide to the Galaxy***

Every project in the history of science relies on data. Collecting data by visiting archives, deciphering manuscripts, or describing artifacts was always essential in the daily work of the historian; digital methods are changing the amount of data that can be processed, and, more important, complex algorithms can be applied to analyze these data. In order for these methods to be applied, data has to be made available digitally. The focus of this essay is on how to work with data in this sense, although we are fully aware that digitizing sources is still one of the main challenges in the life cycle we will describe.[1] For a detailed discussion of this topic see, for example, Stéphane Nicolas, Thierry Paquet, and Laurent Heutte's essay "Digitizing Cultural Heritage

Julia Damerow is a scientific software engineer at Arizona State University with a degree in computer science and a Ph.D. in computational history and philosophy of science. Her interests are the application of computation to the field of history and philosophy of science, software development in the field of digital humanities, and how these can be combined with digital humanities education. She is Head of Development and Cofounder of the Digital Innovation Group at Arizona State University. Global Biosocial Complexity Initiative at ASU, Engineering Center—A Building (ECA), 1031 South Palm Walk, Tempe, Arizona 85281-2701, USA; jdamerow@asu.edu.

Dirk Wintergrün is a research scholar in Jürgen Renn's department at the Max Planck Institute for the History of Science. Since 2000 he has been developing infrastructures and tools for applying digital methods in the history of science. His current work focuses on the applications of network theory and semantic modeling to analyze and describe historical developments. Max Planck Institute for the History of Science, Boltzmannstraße 22, 14195 Berlin, Germany; dwinter@mpiwg-berlin.mpg.de.

[1] In this sense, this essay should also be understood as a guideline for judging whether digitizing is worthwhile for research in a specific field.

Isis, volume 110, number 3. 0021-1753/2019/0110-0006$10.00.

Manuscripts: The Bovary Project," which, though it was published sixteen years ago, describes considerations that are at least in part still valid today.[2]

Depending on the specifics of a research project, data can comprise a variety of different types: full texts, n-grams, annotations, images, audio, video, and bibliographic metadata, to name just a few. Two of the essays in this Focus section, "Triangulation of History Using Textual Data" and "Network Analysis for the Digital Humanities: Principles, Problems, Extensions," discuss the use and analysis of two specific types of data (textual data and network data). This variety creates significant challenges for storing, sharing, and reusing data, as Toby Burrows has discussed.[3] It also means that no workflow for gathering, storing, and analyzing data fits all projects; to some extent, for every project the researchers will have to develop their own methods and processes.

Hand in hand with data collection and storage goes the issue of data sharing, as discussed, for example, in Christine Borgman's "The Conundrum of Sharing Research Data."[4] While data sharing practices are an important topic that needs to be addressed with respect to the history of science, it is not the focus of this essay. Instead, we will discuss the challenges of creating a data corpus by developing a concrete data life cycle. We will use the term "data corpus" to refer to any collection of source documents to be analyzed. Documents in a data corpus can be of any kind, depending on the research project. We will examine what constitutes a source in the computational age. Finally, we will explore the strategies and challenges of data collection and the question of the comparability of datasets.

THE RESEARCH DATA LIFE CYCLE

Juliane Stiller and Dirk Wintergrün, as well as Oona Leganovic and her colleagues, have described the research life cycle of humanities projects.[5] We will use a version of this life cycle, depicted in Figure 1. Planning a digital corpus first requires an assessment of what sources are available—or can be made available—for digital processing. Sources could be a bibliographic database, a collection of images, scans, or photographs, spreadsheets containing measurements, or annotations.[6] In many cases a researcher might start out with a box of letters, manuscripts, and images. In this context we don't distinguish between data and metadata. Metadata is just a specific kind of data: data about (digital) data or any physical or conceptual object.

Once the data sources have been identified, relevant data must be acquired. This can be as simple as downloading a dataset to a researcher's local computer or a server so as to make it available for analysis; or it might involve creating "new" data as part of a project. If this creation process includes the scanning of physical objects, issues pertaining to copyright or ownership have to be

---

[2] Stéphane Nicolas, Thierry Paquet, and Laurent Heutte, "Digitizing Cultural Heritage Manuscripts: The Bovary Project," in *DocEng '03: Proceedings of the 2003 ACM Symposium on Document Engineering* (New York: ACM, 2003), pp. 55–57.

[3] See Toby Burrows, "Sharing Humanities Data for E-Research: Conceptual and Technical Issues," in *Sustainable Data from Digital Research: Humanities Perspectives on Digital Scholarship*, ed. Nick Thieberger (Melbourne: PARADISEC, 2011), pp. 177–192.

[4] See Christine L. Borgman, "The Conundrum of Sharing Research Data," *Journal of the American Society for Information Science and Technology*, 2012, 63:1059–1078.

[5] See Juliane Stiller and Dirk Wintergrün, "Digital Reconstruction in Historical Research and Its Implications for Virtual Research Environments," in *Research Challenges in Cultural Heritage II*, ed. Sander Münster *et al.* (Cham: Springer, 2016), pp. 47–61; and Oona Leganovic *et al.*, "Anforderungen und Bedürfnisse von Geisteswissenschaftlern an einen digital gestützten Forschungsprozess," http://gams.uni-graz.at/o:dhd2015.abstracts-poster.

[6] We use the term "database" to refer not just to the technical implementation (e.g., a specific relational database) but also to the service providing access to it, such as Web of Science, Scopus, or HathiTrust.

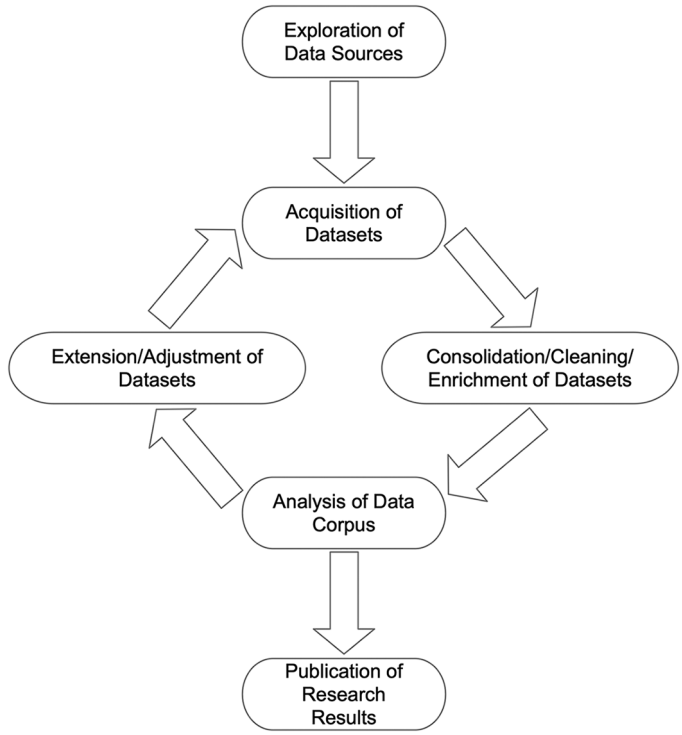

Figure 1. Research data life cycle.

resolved.[7] Before the collected data can be analyzed, and especially if several data sources are involved, it needs to be consolidated. Often different data sources provide their data exports in different formats or highlight different aspects of the data. It might also be necessary to clean the datasets or enrich them with additional information.

Only after a first round of consolidating, cleaning, and enriching the collected datasets is completed can the actual analysis of the data begin. Often the analysis step takes much less time than the preparation steps, and it frequently leads to the realization that additional or different datasets are required—in which case the researcher has to start again with data acquisition. If the analysis step produces answers to a project's research question, results can be published or used as input for further research activities. As in nondigital source analysis, it is critical at each step to document the sources used and the methods employed for analysis.

The research data life cycle shown in Figure 1 and discussed here plugs into the research life cycle outlined by Leganovic *et alia* at the "*Zusammenfassung existierender Quellen*"/"*Korpusbildung*" ("summary of existing sources"/"building of a corpus") steps.[8] However, other parts of the research life cycle—especially the "*Annotieren*" ("annotate") step—might themselves produce data and can therefore be part of the research data life cycle as well. In what follows we will discuss the different parts of the research data life cycle in more detail.

---

[7] See Maggie Dickson, "Due Diligence, Futile Effort: Copyright and the Digitization of the Thomas E. Watson Papers," *American Archivist*, 2010, 73:626–636.

[8] See Leganovic *et al.*, "Anforderungen und Bedürfnisse von Geisteswissenschaftlern an einen digital gestützten Forschungsprozess" (cit. n. 5).

EXPLORATION OF DATA SOURCES: TOOLS AND METHODS

This fundamental step might seem to be a straightforward matter, but as is so often the case the devil is in the details. First a scholar has to identify what type of document is needed for the desired analysis—for example, full texts of documents, bibliographical metadata, or scans of images in books. The exploration step must also include a first assessment of what tools and methods could be used to analyze a dataset. There might be different options and variations, each with its benefits and disadvantages.[9] Scholars need to ensure that the chosen data sources can provide the necessary data or data format. This step also includes the evaluation and selection of appropriate tools for collecting and managing the desired datasets. Such tools can range from widely used software applications (e.g., citation managers such as Zotero or Mendeley), to custom-built applications created by a lab or institution, to project-specific scripts (e.g., specifically developed Python scripts that facilitate the downloading of data from a database).[10]

Next, data providers have to be chosen. Data providers might be archives or libraries that hold physical copies of the documents needed in a research project. In this case, the exploration step includes identifying the organizations that own the required documents and perhaps acquiring permission for their digitization. If the data provider is a citation database, it has to be determined whether the journals or authors relevant to the project at hand are indexed and, if so, whether the information of interest about the citations is indexed as well. For instance, although the Thomson Reuters Web of Science (WoS) indexes funding acknowledgment information, datasets built to analyze such information using WoS might not be reliable due to incomplete data. Arezoo Aghaei Chadegani and her coauthors provide a good comparison of Web of Science and Scopus with respect to some factors a project might consider when choosing which citation databases to use.[11]

Scholars also need to be able to assess the stability of a dataset and to track changes that have been made to it over time. In software development, for example, there is an established practice of describing the differences between different releases, and older versions of software applications are typically available. Databases, however, often do not provide this kind of documentation and don't offer the possibility to roll back to older versions. Thinking about ways to describe the state of a source when accessed is therefore a crucial part of exploration. Although this is also part of the nondigital assessment process, the hurdles are significantly higher in the absence of established memory institutions that take care of digital data—a function archives, museums, and libraries fulfill in the nondigital world. The formation of new digital memory institutions is still in process.

ACQUISITION OF DATASETS

When the initial step of data exploration has been completed, data has to be collected. Collecting can mean downloading datasets to project servers or local machines or identifying ways to access the data for processing using machine readable interfaces (APIs). Using an API often allows the

---

[9] See, e.g., David Westergaard *et al.*, "A Comprehensive and Quantitative Comparison of Text-Mining in Fifteen Million Full-Text Articles versus Their Corresponding Abstracts," *PLOS Computational Biology*, 2018, *14*(2):1–16, for a discussion about text-mining full texts rather than abstracts.

[10] https://zotero.org/; and https://www.mendeley.com/.

[11] Adèle Paul-Hus, Nadine Desrochers, and Rodrigo Costas, "Characterization, Description, and Considerations for the Use of Funding Acknowledgement Data in Web of Science," *Scientometrics*, 2016, *108*:167–182 (on evaluating whether the data provided is suitable for answering a given research question); and Arezoo Aghaei Chadegani *et al.*, "A Comparison between Two Main Academic Literature Collections: Web of Science and Scopus Databases," *Asian Social Science*, 2013, 9(5):18–26.

researcher to download only those parts of a dataset that are needed for a specific step of data processing.[12]

Data acquisition is nearly always a multiphase endeavor. Often projects cannot be based on the study of digitally born materials and already digitized sources alone. Thus the objects of study may have to be brought into digital form before further transformation into the type of data required for the desired analysis is possible. For example, a project with the goal of running topic model algorithms on seventeenth-century books might first have to scan all the books of interest or—if such scans already exist—download them before running optical character recognition (OCR) on them. That process itself might take several iterations of training an OCR model and checking its results before a satisfactory level of precision is reached. Similarly, if the objects of study are oral interviews that exist solely on tape, they might have to be transcribed before they can be annotated or coded so as to address the project's research questions.

From such examples emerge questions that deserve some attention: What is a historical source in the digital age? How detailed must the digital representations of physical objects be to make them useful for historical research? We believe that there is no general answer to either of these questions. The amount of data that can be collected to describe a physical object is not limited; it may include 3D data and also detailed data about the materiality of the object—stemming from spectral analysis, for example—or other physical properties. As discussed by Stiller and Wintergrün in considering guidelines for assessment of the desired quality of digital sources, sources can often be conceptualized as reconstructions of physical objects in different levels of detail.[13] Such reconstructions can range from scans and photographs to networks representing the historical context of an object. For instance, a project might analyze networks of how certain physical objects have been moved over time (from one owner to another), with no need for detailed information about the objects moved. Another project might be researching how the intellectual content of those same objects developed over time and therefore would require scans, transcripts, and chronological data.

In contrast to such reconstructions, some digital sources are truly born digital, such as messages on a mailing list or email metadata (e.g., who emailed whom and when). Additionally, research projects might create new datasets based on digital representations of physical objects (e.g., a project might run named entity recognition [NER] algorithms on full texts). Those new datasets can then be used as sources in subsequent projects. Scholars need to recognize that the kind of data and the degree of precision best suited to answer a particular research question might not be readily available or might be difficult to acquire.

While there exist well-developed tools for some tasks (such as citation managers for building citation corpora), this step often still involves manual labor or the programming of custom scripts to build the desired data corpus. A typical example is the creation of a full-text corpus with metadata. Often the articles of interest for a project (usually PDF files) have to be downloaded into a local citation manager—either manually or in batches, depending on the selected databases. Then the full text of each downloaded article has to be extracted by using extraction libraries in custom programmed scripts or by utilizing an extraction service such as the Giles Ecosystem.[14] Depending on the size of the desired data corpus and the specifics of the data

---

[12] An API (Application Programming Interface) is a way for one program to "talk" to another program. It is a set of functions that, e.g., allow a Python script to call a remote service to retrieve data. Obviously accessing data only via APIs during processing carries the risk that an analysis might not be repeatable if the data or the APIs change.

[13] See Stiller and Wintergrün, "Digital Reconstruction in Historical Research and Its Implications for Virtual Research Environments" (cit. n. 5).

[14] See Julia Damerow, B. R. Erick Peirson, and Manfred D. Laubichler, "The Giles Ecosystem—Storage, Text Extraction, and OCR of Documents," *Journal of Open Research Software*, 2017, 5(1), https://doi:10.5334/jors.164.

providers, this process can be fairly time consuming. Even for this relatively common kind of primary data, APIs for accessing it are still not standardized, despite the long-term efforts of different initiatives like the OAI and the Europeana or projects like CiteSeerX.[15] This also means that often a scholar needs at least some programming skills to automate certain tasks. The legal issues pertaining to the automated harvesting of full texts cannot be discussed here in detail, but they remain a major obstacle for digital scholarship.[16]

CONSOLIDATING, CLEANING, AND ENRICHING DATASETS

This step is often executed in parallel with the data acquisition step. The tasks in this phase are highly dependent on the type of data in the corpus and the corpus provider. The data of some providers is cleaner than that of others, and some data providers offer more information than others. For example, if a data corpus consists of bibliographic metadata, a scholar might have to clean it by correcting some of the entries by hand. Or he or she might enrich it by adding information (such as author affiliations) to each entry, either by hand or by merging a different set of data into it. A scholar who uses two different data providers will probably need to delete duplicates to consolidate the datasets. If a data corpus consists of full texts that were the result of OCR processes, then cleaning the dataset might mean correcting errors in the full texts. To enrich a full-text data corpus, one could add formatting markup to identify titles, headings, paragraphs, and sentences.

As in the data acquisition step, there are tools for some scenarios that support consolidation, cleaning, and enrichment. A widely used tool for working with data in spreadsheet format is OpenRefine. When working with XML files to, for example, mark up the structure of texts using TEI, one can use XML editors such as Oxygen.[17] In some cases, however, this requires scholars to be able to program scripts to automate certain processes or, alternatively, to put in a lot of manual effort.

Not every project requires its data corpus to be extensively cleaned, enriched, or consolidated. Many big data projects base their algorithms on what Christof Schöch calls "messy" data.[18] The underlying assumption in this case is that the sheer amount of data outweighs errors in some individual data points. Still, we would argue that in the digital/computational history of science, the cleaning/consolidating/enriching step can be skipped in only a very few cases, whether because of the limited size of even a "big" dataset (it might be bigger than a "regular" dataset but not big enough to compensate for errors) or because the dataset becomes big only by consolidating different sources or enriching a specific one.

In the sciences and social sciences dealing with "messy" data can never be completely avoided, since all measurement includes uncertainties and systematic errors due to the limited precision of the measuring device used. Therefore, deciding how to deal with errors in measurement is part of every study in the sciences, work systematized as error theory. An equivalent framework for the humanities is still lacking. Only fragments of such a theory exist in the traditional approaches—for example, as part of the problem of dating sources and historical events.

---

[15] https://www.openarchives.org/; https://www.europeana.eu/; and http://citeseerx.ist.psu.edu/.

[16] The most extreme and famous case is probably that of Aaron Swartz (https://en.wikipedia.org/wiki/Aaron_Swartz). But simply reading the "Terms and Conditions" of, say, Scopus (Elsevier) shows the issues that arise: "You may not use any . . . automated downloading programs, algorithms or devices, or any similar or equivalent manual process, to: (i) continuously and automatically search, scrape, extract, deep link or index any Content. . . ." See Elsevier, "Terms and Conditions," *Elsevier Website*, 2019, https://www.elsevier.com/legal/elsevier-website-terms-and-conditions.

[17] http://openrefine.org/; and https://www.oxygenxml.com/. TEI (Text Encoding Initiative) is a set of guidelines describing how to encode texts to make them machine readable (https://tei-c.org/).

[18] Christof Schöch, "Big? Smart? Clean? Messy? Data in the Humanities," *Journal of Digital Humanities*, 2013, 2(3), http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/.

ANALYSIS OF THE DATA CORPUS AND EXTENSION AND ADJUSTMENT OF DATASETS

Full control of a corpus is the prerequisite for any analysis. Although the techniques and methods may be different in a digital framework, historical research relies on trust in its sources. In our contexts this means that only after a data corpus has been acquired and processed (cleaned, consolidated, enriched) can it be analyzed. Depending on the type of analysis and the amount of effort required to build the data corpus, this step might take only a fraction of the time of the previous steps. Yet analyzing a data corpus often results in the realization that additional data is needed or that the data must be filtered or enriched before meaningful results can be obtained. In this case, scholars must identify how the data corpus has to be adjusted (e.g., if new data providers have to be added or if additional data points for the existing data are required) and start over with the data acquisition step. In the digital data life cycle there is no objective cutoff point; the decision as to when results are reliable and sufficient to answer a given research question has to be made by scholars on the basis of their knowledge.

It is also not uncommon for a project to collect a first version of a data corpus and then to realize that additional questions could be answered by collecting additional information. In such cases, several versions of a data corpus or entirely different data corpora might be created. For example, a first round of analysis might categorize documents in a text corpus using topic modeling. Looking at the results, the scholar running the analysis might then decide to employ NER tools to see if there is a correlation between topics and named entities in the text. After a second round of data acquisition (using NER to generate lists of entities in the analyzed documents), the scholar can then analyze how the results of the topic modeling align with the occurrences of named entities.

There are many different tools that can be used during the analysis step. Network data can be analyzed with tools such as Cytoscape or Gephi. Texts can be analyzed with, for example, Antconc, GATE, Mallet, or Voyant Tools.[19] Often, however, as in the other steps described here, some programming skills are required to run algorithms that have not yet been added to a software package.

PUBLICATION OF RESEARCH RESULTS

There are different types of results in computational projects, as we have briefly described. First, there is the data corpus built during the iterations of the data cycle we have introduced here. This corpus might consist of multiple layers that correspond to the multiple phases of the data acquisition process. Second, we have data that stand as the outcome of the data analysis part of a project. Last but not least, there are the results from interpretative work based on the data analysis. To guarantee traceability and to allow reconstruction of research results, all data obtained in the different phases of a research project need to be made available.[20] For example, a project might have created an image collection of high-resolution scans of seventeenth-century physics books, as well as a full-text corpus that contains text files resulting from running OCR on those images. Another project might have compiled a well-curated bibliography of publications about Marie Curie that includes author affiliation and acknowledgment information, as well as a network file that connects all the authors in the bibliography with their coauthors and the institutions they worked at.

While the results created during the analysis step are usually published in the "traditional" way, through publications or conference talks, analytical data is often not published. While digital text (and image) corpora are becoming more and more available and ways to publish these

---

[19] https://cytoscape.org/; https://gephi.org/; http://www.laurenceanthony.net/software/antconc/; https://gate.ac.uk/; http://mallet.cs.umass.edu/; and https://voyant-tools.org/.

[20] We recognize, however, that this is not always possible owing to legal and copyright restrictions.

primary data have become increasingly standardized (e.g., texts made available by HathiTrust or by the Europeana), publication of computer- (and human-) accessible datasets in the humanities is still a desideratum. In the first case, the traditional memory institutions—museums, libraries, and, recently, archives—have accepted that it is now their responsibility to preserve digital heritage as well.[21] The second type of data has not yet found its place in the traditional publication landscape in the humanities. This is contrary to developments in other fields of research—such as meteorology, the earth sciences, empirical social research, or high energy physics and astronomy—where infrastructures for data storage have been developed. Even though there exist repositories and solutions to host data corpora and make them available to other scholars, as discussed by Burrows, several conceptual and technical challenges are involved.[22] In addition, copyright practices often prohibit sharing acquired data, which significantly hampers attempts to reproduce or build on the results of a project.

DISCUSSION

To lift computational history of science to the next level, data corpora need to be treated as first-class citizens with regard to research and project results. It should become common practice to publish data corpora along with descriptions of their provenance and context of creation. It needs to be possible to signal versions of and compare different data corpora in order to integrate them if desired. For the most part the necessary technology exists; it is more a question of infrastructure building and adoption than of feasibility.

An example of how these different technologies could be integrated is being developed as part of the Edition Open Access and Edition Open Sources initiated by the Max Planck Institute for the History of Science. This infrastructure is based on integrating technologies like Jupyter Notebooks, storage systems like Dataverse, and repositories for full-text documents.[23] These development efforts aim to ease re-creation of the research results of computational history of science projects and facilitate documentation of the tools, methods, and data employed.

Good practice in the humanities stipulates that the methods applied and the quality of the sources used are discussed in the publications resulting from a research project. In the long tradition of historical research, standards to describe methods and sources have been well established. For digital methods and sources, we believe, such standards are still lacking. Although it has become more common to mention the tools or algorithms employed for analyzing data, establishing publication practices and quality standards is still in its infancy. We regard software tools and the algorithms used in digital and computational projects as software, since they are either software applications or custom-developed scripts. In such cases widely adopted best practices in computer science ensure quality and reproducibility.[24] However, these still need to be translated into guidelines for humanities publications. Similarly, we need to develop best practices for data corpora to make it easier to start and carry out computational history of science projects. Efforts such as the Research Data Alliance aim to foster data sharing and to promote data-driven research.[25] With this essay, we hope to draw attention to the need to develop best practices specifically in computational history of science by identifying and describing the different steps of building a data corpus.

---

[21] See Brian Ogilvie, "Scientific Archives in the Age of Digitization," *Isis*, 2016, *107*:77–85, for a more detailed discussion.

[22] See Burrows, "Sharing Humanities Data for E-Research" (cit. n. 3).

[23] http://edition-open-access.de/; http://www.edition-open-sources.org/; https://jupyter.org/; and https://dataverse.org/.

[24] See, e.g., Morgan Taschuk and Greg Wilson, "Ten Simple Rules for Making Research Software More Robust," *PLOS Comput. Biol.*, 2017, *13*(4):1–10, for a discussion of how to develop robust research software.

[25] https://www.rd-alliance.org/.

We are aware that one of the main challenges in this context is overcoming disciplinary differences within the humanities. The different disciplines (and subdisciplines) within the humanities have developed their own standards for describing corpora and presenting research results over decades or even centuries. These specific codes are often not explicitly stated but are instead part of the implicit knowledge that is passed on over generations. These codes need to be made explicit in order to transfer them into data models.

In the digital age, we have to define new common grounds that can bring different disciplines together. The absence of such common grounds is one of the main reasons why large interdisciplinary corpora are still lacking.

CONCLUSION

In this essay we have presented a schematic research data life cycle and discussed the process of building a data corpus for computational history of science projects. We believe that the process of building a data corpus needs to be made much more explicit. We are not yet at the point where we have widely available tools that can be used to build a new data corpus with just a few clicks. (See "The History of Science and the Science of History: Computational Methods, Algorithms, and the Future of the Field," in this Focus section, for a description of a specific research project and the steps the researchers took to build a data corpus.) Often building a corpus takes up the greater part of a project's lifespan; at the same time, it is the part of a project with the shortest lifespan and the most limited distribution. There are many good reasons for this situation—but also many reasons that are not so good.

We believe that starting to eliminate those hurdles will move the whole field forward, to the benefit of any future computational history of science project. This is especially true for the elements of the cycle where standards and tools exist outside of the humanities and the process of implementing them in the humanities is mainly one of translation rather than invention. While a detailed discussion of this topic is beyond the scope of this essay, we are convinced that such a translation process would also provide benefits for computer science and the way it deals with data—a matter of ongoing importance in the age of the internet and social media. The expertise of researchers in the humanities—and particularly of scholars in the history of science—is urgently needed to contextualize data: contextualization is an intellectual process based on reflection that cannot be replaced by algorithms. How to describe contextualization formally—the results as well as the process—is still an open research question for digital humanities.[26] Which dataset, method, or technology is used has a great impact on the results of an analysis. Only if historians of science understand and embrace all aspects of computational methods can they use them to their fullest potential.

[26] An example of how to formalize argumentation processes is the CRM_inf Model (http://www.cidoc-crm.org/crminf/). This model might also be a first step toward formally describing how the contextualization of objects is discussed within the humanities research process. Formally, describing a context itself is a much larger challenge and touches the core of research in the humanities, where results are based on long scholarly traditions and implicit knowledge but belief systems are also central.