Thomas Cipro - HIST 501 Term Paper

Over the past few decades, the humanities witnessed several different "turns," parallel developments in theory, methodology, and practice that have changed how scholars approach and think about their work. While it is virtually impossible at present to distinguish which of these turns will end up being the biggest break from previous approaches to history, it is clear that the so-called digital turn is one of the most significant developments within the humanities of the twentieth century. Humanists harness the power of technological, methodological, and institutional innovations to process mind-boggling amounts of data at an unprecedented speed. However, with great power comes great responsibility. Being able to process large data sets raises old questions about the proper treatment of data. Despite the changes brought about by utilizing digital tools, humanist scholars must still confront several questions about methodology and practice. Regardless of how a scholar implements digital methods in their work, data analysis remains one of the driving forces of humanist scholarship. Therefore, it is imperative to consider how the advent of digital scholarship shapes the guiding principles that govern how data is analyzed. In short, digital approaches to the humanities have not erased the big questions asked by scholars, but digital media may lend a new perspective to these issues. This paper will ask why the digital turn brings questions about data management to the forefront and offer a vision of ethical data management. Due to the author's background, the present analysis will primarily consider things from a historian's point of view. This being said, most, if not all, of the fields within the humanities utilize different forms of data and analysis within their work. Thus, the question of how to properly manage and manipulate data is at the heart of all humanist disciplines, not just history.

Many digital tools exist to aid historians in analyzing their sources, including programs that help historians analyze texts, mark up documents, perform computational work, create visuals, and manage or sort sources. Digital tools facilitate data management and analysis because so many different programs exist and can be applied across many circumstances. These developments, which occurred over the last several decades, gave individuals and institutions the means to collect astounding amounts of data, with most data having been generated more recently (Graham, Milligan & Weingart 2016, 27). As acknowledged in the first chapter of *Exploring Big Historical Data: The Historian's Macroscope*, the pace of development within the digital world means that historians are increasingly required to acquire the skills and knowledge to work with these digital tools (Graham, Milligan & Weingart 2016, 2 & 27). Others have stated that a working knowledge of how databases work and how they might be useful for historical analysis will become an essential skill for historians (Harvey & Press 1996, xi). The explosion in data acquisition capabilities raises new questions about storage, sustainability, and privacy-related ethical considerations (Graham, Milligan & Weingart 2016, 29). However, at least in the first chapter, they seem to overlook questions regarding data management.

One of the most important and commonly used digital tools is the database. Charles Harvey and Jon Press defined a database as "a collection of interrelated data organized in a predetermined manner according to a set of logical rules, and is structured to reflect the natural relationships of the data and the uses to which they will be put, rather than reflecting the demands of the hardware and software" (Harvey & Press 1996, 22). Thus, databases exist as a means of storing data in an accessible and somewhat isolated format that suits the needs of the organizer more than anything else. The purpose of this organization is to facilitate analysis. Data is much less useful without being analyzed, and without management or structuring, analysis becomes virtually impossible. Therefore, those who work with data dedicate countless hours to organizing their data and giving their databases structure to make the data more useful.

In his article "Tidy Data," Hadley Wickham makes several prescriptions for what makes a dataset "tidy." Most importantly, he defines what the goal of data tidying is, saying that the purpose of data tidying is to "structure datasets to facilitate manipulation, visualization, and modeling" (Wickham, 20). It is essential to organize and place individual data points into a structure that makes sense to the individual or group of individuals looking to analyze a data set. Thus, people often make a series of decisions to "tidy" their data so that it is easier to manage and analyze. However, these decisions are, ultimately, arbitrary. This is not to suggest that those who tidy their data make their decisions with abandon. Similarly, Harvey and Press describe a comparable means of data organization called 'normalization,' which can be defined as a way of organizing data by establishing relationships between data points and reducing redundancies or duplicates. Regardless of the term used, these processes require plenty of forethought and consideration so that analysts can make the most of the dataset they are working with. Rather, it is to suggest that there is no objective reason to organize a given dataset in a certain manner, even if it is done to adhere to broadly accepted standards and practices. The decisions made when tidying data are made to best suit the needs of the analyst and their objectives. There are countless ways to organize a single dataset depending on the analyst's mission and motivations. Even so, it is almost essential to tidy one's data. Therefore, tension exists between the need to organize, structure, and tidy data and the understanding that the decisions made throughout this process are ultimately arbitrary. Scholars have attempted to alleviate this tension by making their prescriptions regarding data tidying, and the author of this paper will be no exception. However, the contribution to the conversation made in the present paper will emphasize ethical considerations when making claims regarding the best data management practices.

Whereas staunch proponents of data tidying, such as Wickham, argue that data cleaning or tidying (used interchangeably here) is an essential part of data analysis, others embrace a different approach. In their article "Against Cleaning," Katie Rawson and Trevor Muñoz argue that data cleaning can serve as a barrier to data analysis. Essentially, Rawson and Muñoz believe that the term data cleaning itself is problematic because it is simply used as a catch-all term to describe numerous different processes (Rawson & Muñoz 2019, 279). Furthermore, they take issue with an existing notion that working with data is in some fashion mutually exclusive with broader criticism, claiming that it "has too often been offered as a binary choice in which scholars may choose to work in the tradition of cultural criticism or they may choose to work with data" (Rawson & Muñoz 2019, 280). Rawson and Muñoz take issue with how the term data cleaning is used to encapsulate the whole multilayered process and glosses over and inadequately describes the series of subjective decisions made therein. According to the authors, messy data reflects the messiness of the world outside of a researcher's spreadsheet. (Rawson & Muñoz 2019, 291).

There is a great deal of truth in what Rawson and Muñoz claim. Researchers often offer shallow explanations of how they arrived at the decisions made when describing their data-cleaning methods. Furthermore, the insistence that there exists some natural underlying order behind raw data is false. As mentioned by Wickham, all data is essentially a collection of either variables or observations (Wickham, 3). These categories might refer to natural occurrences, such as a measure of temperature, but are, in the end, ways of describing objects or occurrences for the benefit of the researcher. Researchers impose their structures, or rather their vision of what structure is most suitable to their goals, on these descriptions and use data cleaning to improve data representation within these imposed structures. Since there is no objectively correct way of structuring or cleaning a dataset, it is reasonable to assume that different organizational methods offer different analytical insights (Leonelli 2016). What Rawson and Muñoz advocate for is more transparency in how researchers describe their data-cleaning processes so that their decisions may be subject to the same level of scrutiny and accountability as other research process stages. Even Harvey and Press say that "The importance of thorough documentation at every stage of the project must also be emphasized" (Harvey & Press 1996, 97). This way, there is no longer a hardline choice between cultural criticism and working with data because, when working with human products as historians often do, the data generated and cleaned is subject to the same cultural criticism applied elsewhere in methodological reviews.

Rawson and Muñoz make a few different prescriptions about how they believe data should be ethically processed. They land on adopting practices from other disciplines, such as library studies, to better mark up and catalog data because they recognize that, as a field, library studies typically has more established standards for governing data cleaning and organization because of the field's demands(Rawson & Muñoz 2019, 289). Insisting on better markup or indexing does not address the ethical concerns presented by data cleaning. Marking up one's data still consists of a series of subjective choices made to better organize data. However, the value of Rawson and Muñoz's prescriptions lies not in the particular methods adopted but rather in the act of adopting itself. Bringing in data management standards and paradigms from other fields aids in establishing an intersubjective set of rules that researchers are expected to adhere to. "The transition to working in a linked data paradigm should be valued not principally for how it might make large-scale information systems operate more smoothly, but rather for how it can create localized communities of authority, within which people can take control of the construction of data and the contexts in which it lives" (Rawson & Muñoz 2019, 290). Exercising control over the contextualization or description of data and data management is a chief concern of Rawson and Muñoz because different means of knowledge production are subject to differing constraints and philosophical considerations. Thus, a new tension is created in pursuing a more intersubjective way of managing data that leaves room for context-dependent considerations. While this rings true, this tension does not present a direct threat to disciplinary practice that opaque descriptions of data cleaning do. One tension is the product of shallow or incomplete accounts of data-changing processes, while the other is born out of an attempt to respect the unique demands of different fields. Of course, new standards do not completely address the issues highlighted earlier. However, it may very well be the best option on the table scholars can exercise to reduce, but not clear, the philosophical and ethical gap presented by data cleaning. While creating an interdisciplinary set of data management standards is abstract, there are practical ways to work towards this goal. On the other hand, several different challenges and issues are left unaddressed. How should scholars across disparate disciplines arrive at a consensus over these so-called best practices? Do some fields within the humanities warrant greater inclusion or consideration in these conversations? Exactly how much should the humanities borrow from non-humanist fields such as those in the social sciences and natural sciences? These questions will not be answered here as they exist outside of the scope of this paper. Nevertheless, it is imperative to engage in these conversations when considering what is to be done to create a shared collection of data management expectations.

Construction of a set of shared standards for documenting the data cleaning process will inevitably have to answer questions related to storing and accessing said documentation. Authors Julie McLeod and Kate O'Connor began tackling similar questions in their article on data sharing and archiving ethics. The questions asked by the two authors closely mirror the types of questions that the architects of interdisciplinary data management standards would have to address: how to store data management records and how to ensure these records are accessible. As noted in their article, when most people think of data, they tend to think of quantitative and not qualitative data. (McLeod & O'Connor 2021, 528). Practically speaking, the prevailing conception of data tends to chiefly consider quantitative data. As such, many institutions and organizations that already maintain data archives constructed their practices around quantitative data, thus creating a need for changes to be made to suit the needs of those who work with qualitative data. Reasonably, a revolution in data archiving methods is not needed. Rather, slight tweaks would feasibly suffice to improve the abilities of existing data archiving institutions. As McLeod and O'Connor point out, revisions made to archival practice to improve accessibility must also be documented because these decisions are products of broader perceptions about the mission and purpose of archives. (McLeod & O'Connor 2021, 525). This bears direct relevance to the current conversation on data management ethics. No matter how people hammer out the best means of managing data and its dos and don'ts, it is essential to document the deliberative process for posterity. How people think of standard protocols reflects what they believe the overall purpose of said protocols is or ought to be. Subjecting these standards to regular review allows them to be updated to reflect changes in the needs or thinking of the professionals expected to adhere to them. "Revisiting records and data from earlier qualitative research projects not only offers comparative perspectives on particular social phenomena. It also offers a historical perspective on the methods of researchers and the history of social science practices..." (McLeod & O'Connor 2021, 530). Here, access to changes made to practice and convention is seen as a way for contemporary scholars to review and assess the benefits and shortcomings of methodological preferences within their discipline. Furthermore, McLeod and O'Connor argue that accessibility must also be considered from a political and economic perspective. They believe that the potential of creating an open-access research culture must not be spoiled by restricting access along monetary lines (McLeod & O'Connor 2021, 524). How meaningful is it to make research accessible if it is only available to those who can bypass a paywall? Rhetorical questions aside, the move toward creating an open-access research culture is stymied by prohibitive practices such as locking research behind paid access.

Conversations between data management and access to data are inextricably linked. The arguments laid out in this paper are not done simply to create new hordes and caches of data documentation. Instead, one of the goals of this paper is to make transparency a top priority in the research process. Descriptions of how a researcher or team manipulates their data to make it suitable for analysis are often not included or left seriously lacking. For those who are concerned with the rhetorical substance of narratives that utilize data, the absence of such conversations means that a significant portion of a researcher's methodological process is not liable to review or criticism. How to ensure that the argument holds water becomes a more difficult question to answer. Historians and humanists writ large are no strangers to working with data, qualitative or quantitative. Nor are historians novices to critiquing what data is used and how it is used. That is one of the main purviews of historiography. However, historians have yet failed to fully utilize the record-keeping and data management potential presented to them by digital tools. In theory, these digital tools give historians more to look at and put under the microscope of historiography. That being said, they simultaneously make the historiographical process easier and harder. While the historiographer has more to look at and analyze, it increases the level of work put before them. However, historians are rarely known to shy away from such a challenge. Transparency in data management is a fundamental component of data accessibility because it ensures that the whole analytical process is subject to the same scrutiny.

One of the biggest challenges facing the construction of interdisciplinary data management standards is the fact that different fields, which work with different types of data, have different data processing and management needs. Kristen Schuster and Vanessa Reyes believe that some of these gaps can be bridged if people develop a better understanding of the needs of other fields so that they are more equipped to participate in discussions about interdisciplinary structures. (Schuster & Reyes 2021, 130). Their end goal is to work towards a more collaborative set of communities that pull from different occupations to establish knowledge-sharing infrastructure (Schuster & Reyes 2021, 130). Their view addresses the issue at hand by improving broader information sharing and access so that experts in different fields have a more evolved understanding of what the data management needs of other fields are. The idea is that scholars would work collaboratively to arrive at the best data management and sharing practices through this discussion. Schuster and Reyes outlined a wonderful first step in the process of establishing a shared understanding of how to best manage data. Initiating conversations about the disparate data management needs of different fields is perhaps the only way to ensure that the ball gets rolling. Even within the field of history, scholars work with a whole array of different data types and process this information in different manners. This is why having conversations with other disciplines is paramount. Not only can interdisciplinary agreements be made, but it is quite possible to improve intradisciplinary practice, too. The data management needs of an economic or social historian, who is perhaps more likely to work with quantitative data, are going to be more closely aligned with scholars in other social sciences who do the same compared to the needs of a cultural or intellectual historian who is more likely to work with qualitative sources.

Furthermore, the humanities fields may be already moving toward more integration and collaboration. Paul Gooding points to "an influx of so-called 'non-credentialed' librarians from DH-related backgrounds – highly skilled researchers with relevant knowledge, but without formal accredited qualifications in librarianship – into libraries and archives" (Gooding 2021, 137). Gooding is pointing to the movement of many digital humanists into the field of library studies, likely because of economic pressures such as job availability. He is worried that those entering his field are not as skilled or knowledgeable at handling digital material curation and management in a library setting. A reasonable concern, no doubt. Gooding believes that the changes brought to humanities scholarship by the digital turn also change the skills required by scholars in other fields (Gooding 2021, 138). Where training on how to handle digital material and data was once more reserved for some fields, such as library studies, the increased use of digital tools by other fields resulted in more humanists being trained to utilize these digital tools. In turn, more people from backgrounds outside of library studies are equipped with digital management skills and can leverage this knowledge to get postings in libraries and archives. However, Gooding sees this process as rather disjointed and prone to causing problems because of the differences in needs and skills between different fields. Similar to Schuster and Reyes, he believes that one of the best ways to address these problems is to establish a shared understanding of how to utilize digital tools so that knowledge may be shared more easily (Gooding 2021, 140). Again, more and improved discussions are a great first step. However, they are merely a cornerstone that must be built upon. Once more scholars from different backgrounds are put into conversation with one another, it will become more feasible to develop a shared standard of data management and manipulation that prioritizes accessibility and transparency (Leonelli 2016).

Readers will likely notice that library studies seem to play a key role in this conversation, and this is no mistake. If a shared data management standard is agreed upon and employed by numerous scholars from different fields, then the questions of how to store and manage the storage or access to the growing number of records called for by this hypothetical agreement will quickly be raised. The compilation of these data management records is likely to take on digital form, therefore creating new digital libraries and multitudes of digital information that need to be organized and curated. As a field, library studies already has theories and standards when it comes to managing large collections of digital material. This is not to say that the burden of addressing the needs of a shared data management standard is to be put squarely on the shoulders of librarians or archivists. Rather, it makes sense to borrow from a field that has already formed methods for answering the questions asked here. It should be noted that the wheel is not being reinvented for the digital turn has not changed the fundamental epistemological questions asked by historians on how to best use or take care of sources (Graham, Milligan & Weingart 2016, 33). As already mentioned, humanity's improved ability to create data shines a light on existing tensions regarding knowledge production and management. The utility of digital tools is that they engender scholars with improved management capabilities on top of production capabilities. The best way of managing our data may be in front of scholars already, with the most important question being how to adapt to a new paradigm created by new tools.

The digital turn has changed how much knowledge humans can produce and manage. To make the most of this explosion in knowledge, data must be cleaned and adapted so that it suits the analytical needs of those working with the data. However, the term data cleaning is rather vague, and scholars routinely gloss over how they manipulate their data. Since there is no objectively correct way to present or organize data, researchers make subjective decisions in the data management process. These decisions should be subject to the same criticism and review as other methodological processes, but to do this, they must be enumerated and explained by the researchers. It is for this reason that some scholars have come out "against data cleaning." That being said, the need to clean data remains. How to negotiate the ethics of data management? It is the opinion of the author that an intersubjective and shared set of data management rules and standards should be established. Such a set of standards would draw from different humanities disciplines to address the disparate needs of different fields that work with different types of data. In a world where academics work together more often and on bigger and bigger projects, collaboration is key. A shared understanding of how to treat data and approach data management is a key element of fostered collaboration in an increasingly collaborative academic world. Digital tools allow scholars to store and manage data on unprecedented scales. That being said, as the work of digital humanists continues to change, the core tenets of how to treat sources remain fixed. For historians, this means ensuring all aspects of the historical analysis process are subject to historiographical review. Here lies the historiographical impact of data management practices. Digital tools create opportunities to practically submit the steps of the data management process to historiographical scrutiny in the same way as working with analog tools might be put under the same microscope.

Bibliography

Graham, Shawn, Ian Milligan, and Scott Weingart. *Exploring Big Historical Data*. Singapore: World Scientific, 2015.

Gooding, Paul. "The Library in Digital Humanities." in *Routledge International Handbook of Research Methods in Digital Humanities*, edited by Kristen Schuster and Stuart Dunn, 137-51. New York: Routledge, 2021.

Harvey, Charles and Jon Press. *Databases in Historical Research: Theory, Methods, and Applications*. New York: St. Martin's Press, 1996.

Huhges, Geoff. "Sharing research data." *Emergency Medicine Australasia* 29, (2017): 4-5.

Kumar Roy, Bijar and Parthasarathi Mukhopadhyay. "Theoretical Backbone of Library and Information Science: A Quest." *The Jounral of the Association of European Research Libraries* 33, (2023): 1-57.

Leonelli, Sabina. "Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems." *Royal Society Publishing* December 28, 2016. https://doi-org.proxy.binghamton.edu/10.1098/rsta.2016.0122.

Mcleod, Julie and Kate O'Connor. "Ethics, Archives and Data Sharing in Qualitative Research." *Educational Philosophy and Theory* 53, no. 5 (2021): 523–35.

Rawson, Katie and Trevor Munoz. "Against Cleaning." *Debates in the Digital Humanities*, edited by Lauren F. Klein and Matthew K. Gold, 279-92. Baltimore: Project Muse, 2019.

Schuster, Kristen and Vanessa Reyes. "Manage Your Data." in *Routledge International Handbook of Research Methods in Digital Humanities*, edited by Kristen Schuster and Stuart Dunn, 125-36. New York: Routledge, 2021.

Wickham, Hadley. "Tidy Data." *Jounral of Statistical Software*: 1-24. https://vita.had.co.nz/papers/tidy-data.pdf.