2017

So What Are You Going to Do with That? The Promises and Pitfalls of Massive Data Sets

Sigrid Anderson Cordell

Melissa Gomis

Published in *College & Undergraduate Libraries* 24:2–4 (2017), pp. 350–363; doi: 10.1080/10691316.2017.1338979

Submitted February 10, 2017; revised June 2, 2017; accepted June 2, 2017.

So What Are You Going to Do with That? The Promises and Pitfalls of Massive Data Sets

Sigrid Anderson Cordell[1] and Melissa Gomis[2]

1. Hatcher Graduate Library, University of Michigan, Ann Arbor, Michigan, USA
2. Perkins Library, Doane University, Crete, Nebraska, USA

Corresponding author – Melissa Gomis, Perkins Library, Doane University, 1014 Boswell Ave., Crete, NE 68333, email msgomis@gmail.com

ORCID
Sigrid Anderson Cordell http://orcid.org/0000-0003-3956-0606
Melissa Gomis http://orcid.org/0000-0002-5622-8560

Abstract

This article takes as its case study the challenge of data sets for text mining, sources that offer tremendous promise for digital humanities (DH) methodology but present specific challenges for humanities scholars. These text sets raise a range of issues: What skills do you train humanists to have? What is the library's role in enabling and supporting use of those materials? How do you allocate staff? Who oversees sustainability and data management? By addressing these questions through a specific use case scenario, this article shows how these questions are central to mapping out future directions for a range of library services.

Keywords: data mining, library services, supporting digital humanities (DH) across the institution, teaching digital humanities (DH)

Introduction

When the first set of texts from the Early English Books Online Text Creation Partnership (EEBO-TCP) was released on January 1, 2015 (Text Creation Partnership [TCP] 2014), there

was understandable excitement about the release of 25,000 openly available texts from the Early Modern period (Levelt n.d.). In addition to making these texts available to read, this release also opened up possibilities for text mining the EEBO-TCP data set. However, while there is clear potential for digital humanities research in making a relatively clean data set of texts from the early modern period available, the structure of the data set itself poses considerable challenges for scholars without a background in programming. Most humanities scholars cannot take advantage of a data set like this one—or similar data sets, such as the historical newspapers that ProQuest has recently made available to institutions that have purchased perpetual access—without considerable training and support. The question becomes, who is best positioned to provide that support? For many, the obvious answer to this question is the library because of its position as provider of resources and expertise in navigating them. If the library is to provide this support, however, how can it do so most effectively? The gap between the promise and usability of massive humanities data sets like the EEBO-TCP project presents an opportunity to consider a host of questions facing libraries today as they develop service models and expertise to support traditional and emerging forms of scholarship.

This article takes as its case study the challenge of massive data sets for text mining, sources that have been lauded as offering tremendous promise for DH methodology but present very specific challenges for humanities scholars with minimal programming skills. The data management and use issues with which we are concerned in this article engage the question of whether humanists should learn to code; however, they go beyond that in scale and scope. The text sets under discussion in this article raise a broad range of issues if they are to be used by researchers:What skills do you train humanists to have? While the library in most cases helped to create and provides access to these data sets, what is the library's evolving role in enabling and supporting use of those materials? How do you allocate staff in this situation? Who is going to oversee sustainability and data management? By addressing these questions through the lens of a specific use case scenario, this article shows how these questions are central to mapping out future directions for a range of library services.

Background

New digital methodologies and sources for humanistic scholarship raise new questions for training humanities scholars, as well as for the roles that libraries can play in supporting emerging scholarly approaches. As many have noted, emerging digital methodologies in humanities scholarship have opened up new ways to analyze texts at scale. As Heuser, Le-Khac, and Moretti (2001) observe, digital methodologies open up the possibility of asking broader questions of larger corpora to understand texts and underlying social and cultural phenomena at scale. Traditional scholarly methods, in particular the close reading of texts, necessarily limit the scale of analysis, leaving open the question of how authoritative any analysis based on reading a necessarily limited corpus can be. As Heuser, Le-Khac, and Moretti point out, machine reading methods hold promise for allowing us to answer new questions based on a larger, more inclusive corpus: "These emerging methods promise ways to pursue big questions we have always wanted to ask with evidence not from a

selection of texts, but from something approaching the entire literary or cultural record. Moreover, the answers produced could have the authoritative backing of empirical data" (79).

Alongside the "authoritative backing" that "empirical data" promises, these approaches raise concerns among humanists, especially for disciplines that have long defined themselves in opposition to the sciences. As Heuser, Le-Khac, and Moretti (2011) observe,

> By offering an entirely different model of humanities scholarship, the digital humanities raise many questions . . . . Can we leverage quantitative methods in ways that respect the nuance and complexity we value in the humanities? . . . Under the flag of interdisciplinarity, are the digital humanities no more than the colonization of the humanities by the sciences? (79)

In conjunction with this lively debate over whether the core values of the humanities are lost by drawing on computational approaches is the question of how best to train humanists to undertake these approaches, as well as a necessary discussion about what might get lost in the process. Some of the resistance to computational training by humanists, Kirschenbaum argues, stems from a misunderstanding of what computer science is about, as well as its relevance to critical thinking:

> Many of us in the humanities think our colleagues across the campus in the computer science department spend most of their time debugging software. This is no more true than the notion that English professors spend most of their time correcting people's grammar and spelling. More significantly, many of us in the humanities miss the extent to which programming is a creative and generative activity. (2009, B10)

Scholars like Kirschenbaum(2009) have argued forcefully for rethinking humanities training so as to incorporate programming skills. One way to make space, Kirschenbaum suggests, is to replace the foreign language requirement in PhD programs with programming. These skills are crucial, he argues, because

> Computers should not be black boxes but rather understood as engines for creating powerful and persuasive models of the world around us. The world around us (and inside us) is something we in the humanities have been interested in for a very long time. I believe that, increasingly, an appreciation of how complex ideas can be imagined and expressed as a set of formal procedures—rules, models, algorithms—in the virtual space of a computer will be an essential element of a humanities education.

As Kirschenbaum argues, humanities scholars cannot explore the "complex ideas" that humanities computing generates without an understanding of the underlying computational systems.

Likewise, scholars connected to the Humanities, Arts, Science, and Technology Alliance and Collaboratory (HASTAC) have devoted considerable energy to advocating for humanists to learn coding. Hunter (2016) describes an anecdote that her advisor told her when she wanted to do DH work but resisted taking a programming class: "'I'll never forget this young scholar who put himself forward as an expert on Chekhov,' he mused. 'I asked if he spoke Russian, and he proudly said he'd never even taken a class. He lost all credibility in that moment. Don't be the Chekhov scholar who didn't take Russian 101.'" As Hunter suggests, scholars need to understand code to design digital projects.

While there is some consensus in the scholarship that it is valuable for humanists to learn programming skills, there has been less detailed attention paid to what the best process is for teaching those skills. Antonijevic's (2015) ethnographic study of digital humanists reveals an informal, unstructured mode of learning that is focused on point-of-need,

> where learning is linked to immediate scholars' needs, arising from specific research problems, which generally makes this way of learning preferred over organized efforts, such as library workshops, where learning is decontextualized from scholarly practice. This method also successfully makes use of one of the scholars' most scarce resources: their time. (80–81)

As Antonijevic (2015) points out, this method has the disadvantage of "depend[ing] on a scholar's social network and its knowledge capacity" (81).

The idea of a "social network" as the basis for acquiring programming skills is linked to another solution to the training dilemma offered by the literature on digital scholarship: collaboration. Gibson, Ladd, and Presnell (2015) argue that,

> "Unlike traditional humanities research, digital humanities scholarship is not a solitary affair. Generally, no single person has all the skills, materials, and knowledge to create a research project. By nature, the digital humanities project, big or small, requires a collaborative team approach with roles for scholars, 'technologists,' and librarians." (4)

Liu echoes this sentiment, arguing that DH work

> requires a full team of researchers with diverse skills in programming, database design, visualization, text analysis and encoding, statistics, discourse analysis, website design, ethics (including complex 'human subjects' research rules), and so on, to pursue ambitious digital projects at a grant competitive level premised on making a difference in today's world. (2009, 27)

Collaboration, however, requires considerable support and advocacy in a disciplinary landscape where it is not the norm. Reid points out that,

> Unlike a laboratory, which requires a team of people to operate, the default mode for humanities academic labor has been for a professor to work independently. . . .

> It is unusual for humanities scholarship to appear with more than two authors, let alone the long list of authors that will accompany work in the sciences. . . . While there are certainly examples of notable, long-standing collaborations in the humanities, they are exceptions to the rule. (2012, 356)

Although collaboration can be fruitful for scholars in the humanities, it requires both a cultural shift and a rethinking of the workflow for scholarly projects. At this point, collaboration has not been fully embraced by scholars across the disciplines.

In addition to differing disciplinary attitudes that engender resistance to collaboration in the humanities, collaboration can have its own drawbacks, especially when the collaboration is not seen as fully equitable. As Edmond points out, "In the worst cases, teamwork based on an ethos of knowledge sharing can degenerate into the negotiation of uncomfortable tacit hierarchies, where some contributors (regardless of their expertise or seniority) feel like service providers working in the shadow of otherwise autonomous project leaders" (2015, 57). Further, Edmond observes that collaboration doesn't just require bringing people together but also reimagining projects so that all people involved have an intellectual stake. According to Edmond, successful digital humanities collaborations "ensure from the outset that the project objectives propose interesting research questions or otherwise substantive contributions for each discipline or specialty involved" (56). As Reid (2012) explains, "Given that the assemblage operates effectively with a single author, one essentially has to invent new roles for additional participants" (356).

Because of their well-established role supporting research, librarians have taken up the question of how to enable fruitful collaborations and how best they can train humanists seeking to create DH projects or learn programming skills. Green asks how libraries can facilitate "scholars' initial skills acquisition in text encoding" (2014, 222). Green recommends a workshop model that does "not simply inculcate scholars with the latest software; rather librarians and scholars work together to facilitate scholars' entry into the communities of practice that make up digital humanities" (222). Pointing to the TEI (Text Encoding Initiative) consortium as a model, she argues that it "presents a strong case study of the role of librarians in building learning environments that enable scholars to become members of its community of practice" (223).

One key question is whether it is the role of libraries to offer technical support for digital projects, train researchers in attaining new skills (through workshops, for example), or enable collaboration. Lewis et al. assert that "Organizations most successful at building expertise among faculty, students, and staff tended to share characteristics such as *an open and collaborative interdisciplinary culture* in which each team member contributes expertise and is respected for it" (2015, 2).

Discussions of the library's role in supporting scholars in emerging digital scholarship skills necessarily invites a conversation about staffing in libraries. Should the library provide support staff for digital projects, or should that support staff come from the ranks of graduate students? If graduate students are used as labor for these projects, how can it be organically integrated into graduate training? Lewis et al. (2015) point to both the advantages and disadvantages of this model for graduate students:

> Often, digital scholarship projects rely on graduate student assistants. The experience gives students opportunities to build their knowledge and provides inexpensive labor. But such projects must contend with frequent turnover; as one faculty member put it, "I get these MA students, I train them, they graduate." One university that offers degree programs in digital scholarship tries to recruit its own students as staff, but there aren't necessarily enough students to meet the demand, especially with competition from other organizations. Most of their graduates go to industry, since "they can offer more money. The only people we have are here because of idealism." (2015, 27)

Likewise, sustainability can be an issue when the support model is based on labor by students who necessarily stay only a short period of time. In describing the community of practice support model that has been used by various projects such as TEI, Documenting the American South, and the Victorian Women Writers Project, Green points out, "The labor and craft taught for encoding texts generates a 'shared repertoire' of skills that is continually disseminated and refined through the training of new and established scholars. This shared repertoire is a critical element to the ability of a communityof practice to sustain and expand itself"(2014, 228). The community of practice model constantly requires new participants, especially because many graduate students in library and information science programs or schools of information are only pursuing master's degrees and graduate after two years.

At the center of the question of library staffing, training, and support for digital scholarship is the debate over whether libraries should establish digital humanities centers. Ithaka's report on supporting DH outlines three "campus models for support": the service model, the lab model, and the network model. In the network model, "there are multiple units whose services have developed over time, in the library and IT departments, but also visualization labs, centers in museums, and instructional technology groups, each of which was formed to meet a specific need" (Maron and Pickle 2014, 34).

Maron follows up on the Ithaka report on DH centers by arguing that the service model has been controversial in libraries because of the debate over "the degree to which librarians should envision themselves in a 'service role'" (2015, 33). Nevertheless, this is the most common model, and it is driven by the fact that it

> meet[s] faculty and students where they are—to offer courses, training, and some programming support for members of the campus community. This often takes the form of developing a full range of programming, from workshops to courses, and bringing in guest speakers. The library or center following this model seeks to identify and respond to faculty needs rather than "independently identifying a path of innovation" (33), Maron identifies the "path of innovation model" as closer to the lab model.

Likewise, digital humanities centers can create a central space for networking and collaboration. As Freistat explains,

> Digital humanities centers are key sites for bridging the daunting gap between new technology and humanities scholars, serving as the crosswalks between cyberinfrastructure and users, where scholars learn how to introduce into their research computational methods, encoding practices, and tools and where users of digital resources can be transformed into producers. (2012, 281)

While there is much support for the development of digital humanities centers, there are also detractors. Schaffner and Erway argue that "There are many ways to respond to the needs of digital humanists, and a digital humanities (DH) center is appropriate in relatively few circumstances" (2014, 5). Instead, libraries can draw on a host of other approaches to support DH on their campuses. In this case, Shaffner and Erway assert, "[i]n most settings, the best decision is to observe what the DH academics are already doing and then set out to address gaps" (5).

Whether or not libraries build digital humanities centers, there is widespread consensus that libraries are natural partners in supporting digital scholarship. At the same time, there has been much less discussion of the specific challenges raised by complex data sets that are not inherently user-friendly. Libraries offer varying models of support, and there is a robust conversation in the scholarly literature about whether training, direct technical support, or enabling collaboration—or a combination of all three—is the best approach to supporting digital scholarship. As we argue in the next section, the potential and challenges of large data sets provide an opportunity to think through approaches to training, as well as the library's role in supporting teaching and research using these data sets.

Case study: The EEBO-TCP data set

As new digital methodologies emerge, along with new data sets that enable textual analysis at scale, many scholars have sought help from librarians, other researchers (both in and beyond their disciplines), and technology experts as they begin navigating resources and methodologies far outside their traditional training. While there are expected challenges to learning the basic methods of digital scholarship and analysis, a significant additional barrier exists in formatting and preparing the data sets themselves, even beyond the programming skills that are necessary for analysis. For example, while many researchers can operate basic web-based text visualization tools such as Voyant with relative ease, finding and then preparing a corpus for analysis with these tools is often far more daunting. The challenge in this case comes from the complex nature of raw data sets, as well as other factors that work against usability. Creating data sets for analysis often involves individual downloads of plain text files (in the relatively limited cases in which platforms allow that functionality), using R or Python to isolate subsets of larger corpora, or being limited to corpora that are larger than the researcher may need. While it would be unrealistic to suggest that it is possible to eliminate all challenges to creating corpora, putting resources toward facilitating the creation of corpora from raw data sets would offer significant advances in scholars' involvement with digital scholarship. Even data sets that have been produced by libraries pose challenges in usability for researchers.

Without a significant infusion of resources aimed at increasing the usability of these data sets by researchers at all levels of technical abilities, the question becomes, who is best positioned to offer researchers and instructors support in using these data sets? Likewise,who is best positioned to communicate the research possibilities, as well as how to determine a fruitful research question, for using these data sets? Preparing a corpus takes time, and there is no guarantee that text analysis will yield usable results. This article takes the EEBO-TCP data set as a case study to discuss the challenges and potential approaches for libraries to support digital humanities work using these corpora. We draw on the EEBO-TCP data set both because its potential and challenges are representative of other data sets being made available for humanities research and because it is openly available.

EEBO-TCP offers considerable potential because it makes transcriptions of early modern texts available for scholars as well as because it is a clean data set. EEBO-TCP is based on the Early English Books microfilm collection that includes over 130,000 titles from Pollard and Redgrave's *Short Title Catalogue (1475–1640)*, *Wing's Short-Title Catalogue (1641–1700)*, and the *Thomason Tracts (1640–1661)* (Early English BookOnline [EEBO] n.d.).When the microfilm set was originally digitized, the scans appeared as images, and only the metadata was searchable. To make the texts themselves searchable, and because optical character recognition (OCR) software has not yet advanced to handle early modern fonts with any degree of accuracy, the Text Creation Project made the ambitious decision to re-key (i.e., transcribe) the texts as well as to mark them up using XML/SGML encoding. Although the original goal was to make the texts full-text searchable, emerging text mining methodologies have made the existence of clean data sets particularly desirable for researchers. Because the texts have been rekeyed, there are fewer errors in the texts than in those that have been OCR'd. As part of its agreement with ProQuest, which makes the EEBO database commercially available, Phase I of the EEBO-TCP texts, which includes the first 25,000 rekeyed texts, was made publicly available in December 2014.

While the data set offers considerable potential for researchers and also makes the texts themselves available, the data set itself is not easy for researchers to use for a variety of reasons. The texts are available either as a full data set on Box and GitHub, or as individual, HTML, ePUB, and TEI P5 XML files through the Oxford Text Archive. The files on Box and GitHub are referenced by TCP number, a number that is not available on the ProQuest platform, meaning that researchers who are not interested in working with the corpus as a whole—who, for example, are interested only in texts from a specific time frame or author—have to do considerable extra work to identify the relevant files before they can begin downloading and formatting them for analysis.

While researchers who are fluent in programming languages such as R or Python have little trouble accessing these texts, in our experience many researchers in the humanities are understandably daunted when faced with zip files containing 25,000 files, each of which contains XML or SGML markup that they must decide whether (and how) to scrub or retain. There is little documentation on strategies for accessing and cleaning up the text in preparation for mining or information on analysis tools once you have the data.

Likewise, ProQuest has recently made their historical newspaper collections available (for a fee) to libraries that have already purchased perpetual access to specific titles. When libraries license the full-text data sets of historical papers, they are given access to the

marked-up files. The *Los Angeles Times*, for example, is a collection of 4.5 million files, presented in no particular order and with no metadata in the file names. As in the case of the EEBO-TCP data set, to make use of these files, researchers must begin by pulling down slices of the corpus (such as by year or article type) using R or Python. Unlike the EEBO-TCP files, most *LA Times* articles are not available one by one as plain text files on a platform for researchers to cobble together a corpus through the search interface (and license agreements generally limit bulk downloads in any case). Once researchers have pulled down a subset of the corpus, they must decide how much of the markup to keep or strip out before they can run it through a text visualization tool (unless they decide to use the text mining package in R or a similar programming language). Leaving aside the technical skills needed to do this, researchers must also decide how to approach the dirty OCR problem because the texts themselves are riddled with errors caused by the conversion process from microfilm. While data sets like this offer tremendous potential, it is not feasible for humanities scholars to make use of them without considerable support.

Another example outside of the humanities is the United States Census Bureau, which provides access to data sets through a variety of different websites and formats. Determining the type of data that is needed and locating that data can be challenging to researchers new to working with census data. The Census Bureau offers a list of recommended software and provides workshops, webinars, and classroom trainings to help people get what they need. They also provide phone and e-mail support for researchers and people using census data in their work. Libraries are just beginning to offer a range of data sets to their users either through their subscription databases or through their own digital projects. Usually this type of information is provided without creating a service model. Faculty and students often have to figure out how to use these data sets themselves. Once users have the data set, the library doesn't play a strong role in helping them use it. The U.S. Census Bureau could serve as a service model for supporting text mining in the digital humanities.

When an institution or a company provides access to a data set, do they have a responsibility to assist researchers in using the data set? The following section presents different support models that allow us to examine the ways libraries are supporting digital scholarship projects with large data sets for research and learning. Gaining access to the texts and analysis tools is not always the barrier to digital scholarship, especially for content out of copyright. Researchers often need help locating resources, including money for staff, storage space, and software and technological expertise to execute their projects.

Potential support models for digital scholarship using unwieldy data sets

Although there are certainly scholars out there who are capable of making use of raw data sets, the majority are not. We as librarians and scholars need to advocate for the ways in which our scholars want to use these materials.At the moment, we are operating in a bifurcated context: On the one hand, there exist graphical interface tools that do not give you much flexibility or control to manipulate or build the corpus you are analyzing but that meet the needs of some researchers, such as the Google N-Gram tool, or on the other hand, a move by publishers to dump the raw data. As in the case of the ProQuest Historical Newspapers data sets, publishers have responded to requests from researchers by making

data sets available; these data sets are usually delivered in large raw text file dumps that are not manageable to the average humanist scholar.

Advocacy

As a first step in enabling research with these data sets, libraries, as the purchasers and as the supporters of researchers, need to advocate for tools that create bridges between easy-to-use digital tools (like Voyant and AntConc) and the data sets. For example, rather than having either the entire raw data set for EEBO-TCP or the Oxford cut-and-paste formatted version, why not create tools that make it easy to use the platform to designate a corpus (i.e., by doing a search using the parameters on the platform) and then extract plain text files from the search results? In the case of the ProQuest Historical Newspapers example mentioned, it is not consistently possible across the PQHN platform to download plain text files of individual files, although this would make text mining custom corpora much more manageable for researchers without a background in programming or the resources to hire an assistant to manage the technical aspects.

Creating new tools

Leonard recommends that libraries create tools or adopt open source tools to make analysis easier. At the Yale University Library, they adopted the HathiTrust Bookworm tool to analyze a small digital corpus of the *Vogue* collection. By creating tools that researchers can use to search text in other ways, they also help patrons analyze their large digital collections (2014).

To facilitate work on the EEBO-TCP data set,Washington University in St. Louis created the Early Modern Print (n.d.) project, which is supported by the Humanities Digital Workshop at Washington University. The Early Modern Print project provides exploration tools tailored to the EEBO-TCP data. They describe the tools as

> an aggregate view of the corpus that enables us to probe English lexical and orthographic history in ways that usefully complement the search capabilities of EEBO-TCP and the Oxford English Dictionary; they also help us to see early modern book culture in a new way, as a structured flow of words. (Early Modern Print n.d.)

The developers have created graphical interface tools, such as an EEBO N-GRAM Browser, to facilitate use of the collection by researchers, but users necessarily have less ability to manipulate the corpus when they are using this tool.

Until there are more robust tools available to make working with a broad range of data sets easier for scholars, libraries can play a role in supporting emerging research by teaching scholars basic skills.

The workshop model: Creating stages for learning

In designing workshops to teach skills in digital scholarship, librarians need to be attentive to felt needs in their community and to carefully stage those workshops to make sure that instructors are not spending too much time on technical minutiae, such as constructing a

corpus or setting up frustration with tools. To do this, workshop facilitators need to draw on the principles of backward design by asking what is the intellectual outcome that they want to have in the session? Wiggins and McTighe explain backward design as a methodology that conceives of curricular design by thinking at the outset in terms of outcomes rather than lessons: "Given a task to be accomplished, how do we get there? . . . What kinds of lessons and practices are needed to master key performances?" (1998, 8). In just the same way that you might design a classroom exercise to focus narrowly on imparting a specific skill or research strategy, it is useful to isolate the specific technical skill, as well as the possibilities for further exploration, that you hope to impart. This is likely to require more setup in advance by the workshop leaders—for example, creating a specific corpus to work with or downloading example files to practice on—but it will allow the session to focus on that specific skill rather than the frustrations of getting ready to learn that skill. A scenario to avoid is workshop participants trying to download software and winding up spending most of the time troubleshooting the download and relatively little time using the tool.

Designing workshops in ways that focus narrowly on outcomes may also require participants to use the same operating system and computers that have all been set up the same in advance. Creating an equal computing environment is a big challenge, especially when people have different skill levels and different technology vocabularies. As the scholarship on how researchers learn technical skills suggests, if you can give an opening to the possibilities, and offer a framework for follow-up support, interested researchers will take the time to teach themselves or request consultations on how to do the technical minutiae. A key goal for a workshop can often be illustrating the possibilities. How can you illustrate the possibilities in the approach so that scholars are motivated to learn the details of downloading and constructing their own corpus? Can you create a session that focuses on a piece of the process—that is, looking at a predetermined corpus in AntConc? One approach is to make the entry easy so scholars can decide if they want to do more and then offer resources for them to take the next steps. A significant goal for workshops can be illustrating why researchers would want to learn these approaches.

Workshops can also be augmented by working sessions, such as the Hackfest sponsored by the Bodleian libraries in 2015 (Oxford University n.d.). This full-day session included researchers as well as robust technical support, as participants had a chance to "pitch ideas and find collaborators, firm up projects and groups, and request (or indeed recruit) technical help as necessary" (Willcox 2015). Key to the success of this model, practiced also by Software Carpentry, whose goal is "teaching basic lab skills for research computing" (Software Carpentry n.d.), is the availability of support from multiple people, rather than one or two workshop leaders trying to troubleshoot and lead the session.

Classroom approach

In addition to workshops aimed at researchers at all levels, librarians can offer considerable support for digital scholarship through course-integrated instruction at the undergraduate or graduate level. If integrated thoughtfully into a course's learning goals and assignments, course-integrated instruction can be, arguably, at least as effective as workshops because the individual skills to be taught are bound up with the questions raised by a specific course theme. By working with the faculty member leading the course, and by

being attentive to the specific learning goals and questions for the course, librarians can design exercises that are targeted toward specific research questions. Just as in workshops, it is essential that librarians front-load the planning for these instruction sessions to isolate the specific learning goal for the course. While it is not possible, nor is it realistic (or, really, desirable), to eliminate all possible frustration in working with complex data sets, librarians can anticipate and minimize potential pain points so the session can focus on the learning goals.

For example, in one undergraduate class session at the University of Michigan, the librarian and technology specialist worked closely with the faculty member to design an instruction session that drew on the EEBO-TCP data set in a 300-level course. Because the point of the assignment was not necessarily to teach students how to compile corpora for analysis but rather to allow students to perform text analysis on a set of relevant texts, they set the session up so that students were creating a limited corpus of only ten texts, based on search criteria that students determined (and determining the search words was part of the goal for the exercise). To minimize frustration with the data set as a whole, they first showed students how to use the EEBO platform so as to explore texts related to their topics and identify ten potential texts. Once they had identified the ten texts, it was relatively easy for students to find those texts on the Oxford platform and cut and paste the text into plain text files. Although this approach may have glossed over some of the intricacies of the data set and corpus creation, it allowed students to create a minicorpus relatively easily to import into Voyant, where the bulk of the learning was meant to happen.

The lab approach: ScholarSpace at the University of Michigan Library

ScholarSpace at the Graduate Library at the University of Michigan provides access to technologies for small-scale experimentation and technologies for formal project support with the understanding that anyone can access them. ScholarSpace supports humanists working on text mining projects by providing access and expertise for digitization, storage, text cleanup, and analysis.We have purchased text mining software that is not available elsewhere on campus, thereby providing access to anyone affiliated with the university. This approach relies on humanists to be willing to experiment with librarians and to train each other. Text mining varies greatly by discipline; through creating a community of scholars, we can build a network of experts and draw on experiences and expertise related to text mining in Chinese studies, economics, history, English language and literature, and more.

Staffing models

Across these different models, the question remains as to how best apportion staffing to support digital scholarship. In a distributed model, where librarians are leading workshops for the campus community and for classes, subject specialists, technology librarians, and undergraduate learning librarians can provide considerable support, especially if they are provided training and if the workshops are a natural extension of their expertise and outreach areas. Depending on the demand on campus, this model can, however, lead to librarians being stretched too thin; thus, creative staffing, such as training students to lead or support workshops, is necessary.

Likewise, students can be brought into a project to work on a specific slice—such as OCR-ing pdf files and cleaning up the resulting OCR. In this case, however, it is important to bring the students into the conversation about the project at some level, so they understand how their work fits into the larger intellectual work of the project. Otherwise, libraries miss out on the opportunity to mentor students in emerging questions and methodologies of digital scholarship. The bulk of preparing texts for mining and analysis can also be tedious, and it requires careful attention to detail. Librarians or others overseeing students working on DH projects need to be vigilant in keeping the work moving forward and in checking the quality and consistency of the work.

Sustainability and scalability are challenges across all staffing models. Projects that have dedicated funding may not have enough funding to cover the entire project. Students cycle off projects either because they graduate or because they receive other opportunities such as internships or jobs.

Conclusion

As the preceding discussion of staffing illustrates, challenges remain in thinking through collaborative work in digital scholarship, especially in terms of the necessary—but not as obviously exciting—work of data preparation and cleanup. The need to develop and create digital scholarship projects will continue to grow in the humanities, and at some institutions it will be embedded into the curriculum. Learning project management, digitization, and analysis are skills humanists will need in the future, and they will learn them through the channels available. These skills can translate easily to a number of positions postgraduation and will be desired by employers. Having graduate students work on digital projects can provide them with perfect opportunities to obtain new skills.

Considering that resources are not currently in place to make data sets easier to use in the near future, librarians can advance digital scholarship by helping scholars in incremental ways targeted at the specific challenges and frustrations that data sets pose. Librarians can set the expectation that they will work with students and faculty to explore these new areas together and work to scaffold the learning experience so that humanists beginning text mining see the possibilities and not just the minutiae. Some challenges that still persist include developing relationships across campus, continually building skills, and finding partners with which to collaborate.

References

Antonijevic, Smiljana. 2015. *Amongst Digital Humanists: An Ethnographic Study of Digital Knowledge Production*. NewYork: Palgrave Macmillan.

Early English Books Online (EEBO). n.d. "What Is Early English Books Online?" http://eebo.chadwyck.com/about/about.htm#top

"Early Modern Print: Text Mining Early Printed English." n.d. http://earlyprint.wustl.edu

Edmond, Jennifer. 2015. "Collaboration and Infrastructure." In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 54–65. Chichester, UK: John Wiley & Sons.

Freistat, Neil. 2012. "The Function of Digital Humanities Centers at the Present Time." In *Debates in the Digital Humanities*, edited byMatthew Gold, 281–91. Minneapolis: University of Minnesota Press.

Gibson, Katie, Marcus Ladd, and Jenny Presnell. 2015. "Traversing the Gap: Subject Specialists Connecting Humanities Researchers and Digital Scholarship Centers." In *Digital Humanities in the Library: Challenges and Opportunities for Subject Specialists*, edited by Arianne Harsell-Gundy, Laura Braunstein, and Liorah Golomb, 3–17. Chicago: Association of College and Research Libraries.

Green, Harriett E. 2014. "Facilitating Communities of Practice in Digital Humanities: Librarian Collaborations for Research and Training in Text Encoding." *The Library Quarterly* 84(2): 219–34.

Heuser, Ryan, Long Le-Khac, and Franco Moretti. 2011. "Learning to Read Data: Bringing out the Humanistic in the Digital Humanities." *Victorian Studies: An Interdisciplinary Journal of Social, Political, and Cultural Studies* 54(1): 79–86.

Hunter, Elizabeth. 2016. "Must Humanists Learn to Code? Or: Should I Replace My Own Carburetor?" *HASTAC* (blog), December 7, https://www.hastac.org/blogs/shakespearegames/2016/12/07/must-humanists-learn-code-or-should-i-replace-my-own-carburetor

Kirschenbaum, Matthew. 2009. "Hello Worlds: Why Humanities Students Should Learn to Program." *The Chronicle Review* 55(20): B10.

Leonard, Peter. 2014. "Mining Large Datasets for the Humanities." *IFLA Library*. http://library.ifla.org/930/1/119-leonard-en.pdf

Levelt, Sjoerd. n.d. "#EEBOLiberationDay." https://storify.com/SjoerdLevelt/eeboliberationday

Lewis, Vivian, Lisa Spiro, Xuemao Wang, and Jon E. Cawthorne. 2015. *Building Expertise to Support Digital Scholarship: A Global Perspective*. Washington, DC: Council on Library and Information Resources.

Liu, Alan. 2009. "Digital Humanities and Academic Change." *English Language Notes* 47(1): 17–35.

Maron, Nancy. 2015. "The Digital Humanities Are Alive and Well and Blooming: Now What?" *Educause Review*. http://er.educause.edu/~/media/files/articles/2015/8/erm1552.pdf

Maron, Nancy, and Sarah Pickle. 2014. "Sustaining the Digital Humanities: Host Institution Support beyond the Start-Up Phase." *Ithaka S+R*. http://www.sr.ithaka.org/wp-content/mig/SR_Supporting_Digital_Humanities_20140618f.pdf

Oxford University. n.d. "Text Creation Partnership: EEBO, ECCO and Evans Texts." http://ota.ox.ac.uk/tcp/

Reid, Alexander. 2012. "Graduate Education and the Ethics of the Digital Humanities." In *Debates in the Digital Humanities*, edited byMatthew Gold, 350–67. Minneapolis: University of Minnesota Press.

Schaffner, J., and R. Erway. 2014. "Does Every Research Library Need a Digital Humanities Center?" *OCLC Research Report*. http://www.oclc.org/content/am/research/dpublications/library/2014/oclcresearch-digital-humanities-center-2014.pdf

Software Carpentry. n.d. "Software Carpentry: Teaching Basic Lab Skills for Research Computing." https://software-carpentry.org

Text Creation Partnership (TCP). 2014. "EEBO-TCP Phase I Public Release: What to Expect on January 1." http://www.textcreationpartnership.org/2014/12/24/eebo-tcp-phase-i-public-release-what-to-expect-on-january-1/

Wiggins, Grant P., and Jay McTighe. 1998. *Understanding by Design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Willcox, Pip. 2015. "Early English Books Hackfest." *Bodleian Libraries* (blog), April 22, http://blogs.bodleian.ox.ac.uk/digital/2015/04/22/early-english-books-hackfest/