Working with “Big Data” and creating sustainable projects are current concerns in the field of Digital Humanities. Five diverse projects are targeted as cases studies to explore these issues. Projects include The Old Bailey Proceedings Online, The Perseus Digital Library Project, Boston Streets Mapping Directory, Digital Harlem, and Going to the Show. These projects are highlighted as they showcase a variety of models and techniques for promoting documentation and sustainability of big data work in the digital humanities. From these case studies a story emerges. Sustainable, interoperable projects in the digital humanities are deeply tied to acknowledging and integrating the work of digital humanists into wider, already established realms of research and scholarship. This is reflected in three unique areas can be used to evaluate the proficiency of these projects. These areas are technical documentation, citation recommendations, and bibliographies.

Headings:

Digital libraries

Data libraries

Digital humanities
CITING, COMPILING, AND DOCUMENTING:
SUSTAINING USE OF BIG DATA RESEARCH IN THE DIGITAL HUMANITIES

by
Willa M Armstrong

A Master’s paper submitted to the faculty of the School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Library Science.

Chapel Hill, North Carolina
July 2012

Approved by

_______________________________________
Richard Marciano
# Table of Contents

**Introduction** .......................................................................................................................... 2

**Literature Review** ................................................................................................................... 5

**Case Studies** .......................................................................................................................... 4
  - *Old Bailey Online* .................................................................................................................. 10
  - *Perseus Digital Library* ........................................................................................................ 14
  - *Boston Streets Mapping Directory* ...................................................................................... 17
  - *Digital Harlem* ..................................................................................................................... 18
  - *Going to the Show* ................................................................................................................ 19

**Summary** .................................................................................................................................. 22
  - Technical Documentation ........................................................................................................ 22
  - Citation Recommendations ..................................................................................................... 22
  - Bibliographies ........................................................................................................................ 23

**Discussion** ............................................................................................................................... 23
  - Future Research ....................................................................................................................... 27

**Conclusion** ................................................................................................................................ 29

**Works Cited** ............................................................................................................................. 31
Introduction

The term “Digital Humanities” encompasses a broad range of subjects, programs and projects performed by people grounded in a wide array of disciplines. Much has been said about the inherently interdisciplinary nature of such pursuits. These interdisciplinary scholars are united in excitement over the asking of new questions and the creation of new knowledge. This is knowledge that can be used by a yet unknown audience. With the rapid changes in technology and scholarship in the last few years, it is important to step back and evaluate strategies and norms of sustainability found in digital humanities projects.

A deep vein of knowledge management runs throughout the digital humanities. The culture of the field encourages creating and sustaining a connected body of knowledge generated by interdisciplinary workers. “Boutique” projects are instances when all the energy and dedication of collaborators end at the creation of one insular output. This is an outdated and unrealistic model. Those in the field are working to establish modular project skeletons: workflows, technical underpinnings, and user interfaces which can be picked up and adapted in other projects. There is an emphasis in creating data sets upon which expansion of scholarship on which is easily built. Strong, sustainable projects are those which enable original and new researchers to draw on them in new ways. This potential helps create buzz around projects: “People will use this data in ways we can’t even imagine yet…that is one of the most exciting developments in the
humanities” (P. Cohen). The culture of digital humanities is a rich, supportive environment for these endeavors.

In the digital humanities, a great deal of work is required in the process of cleaning up and manipulating data. This work results in the creation of unique, value-added data sets. It is important to consider the ease with which unaffiliated users and scholars can engage with these data sets. The development of curated, communal knowledge requires standards and practices which encourage correct citations. In order to build up a bank of knowledge, scholars must build on one another’s efforts and not recreate work unnecessarily. The five case study projects are evaluated in how well production work is publically documented and the extent to which team members’ contributions are recorded. These projects range from a closely curated collection of literature on the picture show to a broad, ongoing digital library experiment.

The importance of achieving these goals increases as the scale of data grows. Big data is a buzzword conjuring up the challenges of managing terabytes of data. However, work done in big data requires researchers to not only lasso and corral large quantities of data, but also to perform work with the datum which composes the corpus. Several case studies demonstrate how various projects handled making individual units of data made accessible to scholars. After an analysis driven by these questions, several patterns emerge which mirror the concerns and trends in digital scholarship on the whole. These trends include ideas about open source fundamentals and access issues; perhaps most influentially, there exists underlying concerns about attribution and credit in the changing landscape of scholarly publications and academic output. These anxieties were matched with solutions and improvements present in digital humanities projects. Gathering
together case studies importantly creates a space for discussion of successes and shortcomings. Most importantly, this paper suggests some concrete examples of well-formed solution incorporated into a variety of projects.

Other obstacles stand in the way of embracing big data research. A hurdle in the adoption of digital humanities methods is frequently rooted in the fact that the codified schema for advancement in various scholastic fields is only slowly taking into account new formats for scholarship. Guidelines put out by organizations such as the Modern Language Association respond to diverse new forms of putting forth scholarly work; these methods are finding a place in the wider context of academia. There needs to be clear communication to future scholars documenting the contributions of respective team members and recommending citation format. Polished interfaces with nifty displays and unique data sets are the result of a lot of hard work by, to name a few, primary investigators, computer scientists, undergraduates, and seasoned professors. Their work at the back end as well as ongoing and future work deserves to be clearly acknowledged. Both for the benefit of their professional reputation and for future work to be built upon it.

This paper presents a widespread survey on the intersection of big data and the digital humanities with a special focus on the manipulation and handling of the individual units of data which create large data sets. Examining in unison the challenges to both large scale management and individual pieces of data can yield valuable insight in how projects are working and how they should work. Through exploration of big data in the digital humanities a clearer picture emerges. Digital humanists are working to see the acceptance of their work in larger academic circles. These efforts are intrinsically
connected to the continued embrace of the fantastic potential in new technology and large data sets.

**Literature Review**

The digital humanities is a field which thrives on collaboration. Conversations about what digital humanists are doing and how and why they are doing it abound. Within this space big data generates stimulating, reflective talk. This is unsurprising given what many see as the game-changing nature of big data:

To grasp the potential impact of Big Data, look to the microscope, says Erik Brynjolfsson, an economist at Massachusetts Institute of Technology’s Sloan School of Management. The microscope, invented four centuries ago, allowed people to see and measure things as never before — at the cellular level. It was a revolution in measurement. (Lohr)

This comparison to the microscope is woven throughout discussions of big data. Balanced use of techniques to examine big data is one of much debate. Some scholars are hesitant to embrace the new manipulation of humanities’ big data. There is a concern that the old, deep analytic techniques will be tossed out in favor of superficial statistics standing in for humanistic research. These fears are most constructively re-articulated as a clear decision to embrace all methods. “The digital humanities do fantastic things,” said the eminent Princeton historian Anthony Grafton. ‘I’m a believer in quantification. But I don’t believe quantification can do everything. So much of humanistic scholarship is about interpretation…It’s easy to forget the digital media are means and not ends,’ he added” (P. Cohen). It is useful to extend this analogy; the microscope may have revolutionized scientific measurement, but it did not entirely eclipse the methods of measurement then established. When scientists study how diseases spread, the research occurs at the cellular, individual, and aggregate level of the population. When new
technology emerges allowing inquiry at different levels, these methods are integrated into extant tools and technologies. This is true also with digital tools and the humanities. To this end, projects in the digital humanities must balance the access tools to big data and datum so that scholars may grapple with these data at multiple levels.

Brett Bobley is the director of the National Endowment for the Humanities’ Office of Digital Humanities. When describing works in the digital humanities, Bobley includes “open access to materials, intellectual property rights…[and] digital libraries” to the “digital humanities rubric” (Bobley 61). The digital humanities is a term for a wide array of materials, scholarship, and institutions. Methodologically, focusing on several case studies helps ground the subsequent discussion of issues within this digital humanities rubric. This discussion is an important one as the rate and scope of change is growing. In speaking to this change, Bobley remarks:

On the consumption side, people get their materials in all kinds of new ways. Reading has changed with the web. It has changed from a technology perspective, of course – thinking of e-readers and laptops and mobile devices (and some of the now-starting-to-get-obsolete tech products like microfiche machines). But the changes are more profound than that. The way we read is changing – bits and pieces of varied content from so many places and perspectives. (Bobley 63)

This varied, multi-origin content is being studied and manipulated by scholars across disciplines. However, the origin of and entry points to content are not the only metric changing; Bobley continues, “if I had to predict some interesting things for the future in the area of access, I’d sum it up in one word: scale. Big, massive, scale. That’s what digitization brings – access to far, far more cultural heritage materials than you could ever access before” (Bobley 63). Big data means, on an increasing scale, items from multiple sources layered together against one another. The burgeoning scale of scholarship underscores the need for well-formed citations. Standards are needed more
than ever. This is especially true as doing so will promote an increase in the understanding and acknowledgement of people’s contributions.

Discussions about correct citations and the acknowledgment of all collaborative contributors are lively and well in the digital humanities community. Workshops, publications, and models have risen from the debate. Bethany Nowviskie directs Digital Research & Scholarship at the University of Virginia Library. In a discussion on citation practices Nowviskie remarks “if you look only for finished products and independent lines of responsibility, you will meet with frustration in examining the more interesting sorts of digital constructions…precisely the sort of innovative work you want to be presented with” (Nowviskie). The expectation that scholarly work will result a single journal article or monograph fails to allow for the type of work accomplished in the digital humanities. Debates over citation tie into larger issues of changing scholarly publications and alternative careers in academia. The awareness and rigor these conversations instill on citations strengthens the standards of interoperability in digital humanities projects. In the past, work in the humanities was often communicated through monographs and articles. These forms are easily cited and built upon. The work and intellectual creativity behind digital humanities projects requires acknowledgement. Currently citations for value-added data sets are less established than those for a book. How should we cite the useful visualization tool adapted for a new use? How do we promote adoption of these citations? Several digital humanities projects are reviewed to investigate how well the data is presented to ease future scholarship. Techniques and qualities used to promote citation are also examined.
"One Culture," a report put out by the Council on Library and Information Resources (CLIR) overviews the projects from their first Digging into Data Challenge.

(These projects include two which grew out of larger projects highlighted as case studies in this paper: the Old Bailey Online and the Perseus Digital Library Project.) The authors of One Culture touch on the change in research time-spans; scholars can quickly achieve results from analysis that used to take far longer. They also discuss the impact of dealing with large scale data. The authors declare:

By choosing to work with very large quantities of digital data and to use the assistance of machines, the Digging into Data Challenge investigators have demarcated a new era—one with the promise of revelatory explorations of our cultural heritage that will lead us to new insights and knowledge, and to a more nuanced and expansive understanding of the human condition. (Williford and Henry 1)

This joyful statement embraces the new directions research now follows; furthermore it assumes that in a changing landscape, the new scholar will adopt skills of manipulating digital technology as part and parcel of their expertise. Scholars have always manipulated technology to pursue humanistic research. Turning the pages of a book is a manipulation of technology! However, these established technologies are readily embraced and so prevalent as to be unseen and not cited. The newness of some of this technology encourages articulated sourcing. Some look to social scientists and their treatment of methodology:

Social scientists are generally comfortable foregrounding explanations of methodology in discussions of their research; humanists, by contrast, tend to foreground the argument or interpretation resulting from scholarly investigation rather than the research methods. Asserting the value of one’s approach to research as a model for others is a more comfortable position for social scientists than for humanists. Humanists often see greater understanding of the subject matter with which they are concerned as their primary contribution to their fields, or at least a more important contribution than the preparatory work necessary to describe new findings and support new claims. (Williford and Henry 13)
The authors go on to claim that interdisciplinary researchers adopt conventions of one another. These researchers are creating, from the many extant research discipline cultures, one culture. As many information professionals work in the digital humanities it makes good sense to investigate how to keep track of all these newly combined research sources and techniques.

*One Culture* mentions a “data hierarchy” suggested by Richard Healey in response to his experience working on the initiative. This model parses raw data into five levels which describe the degree to which they currently support various digital humanities activities (Williford and Henry 14). These tiers range from 0-4 and cover extremely error filled data, raw data which has been corrected for obvious errors, and “value-added data sets” (Williford and Henry 14). The penultimate data tier contains value-added data sets which have been linked together. Healey reserves the highest classification for data which “facilitates extensive additional types of digging activity to be undertaken on substantive products beyond those of the investigators who created them, i.e., they become ‘authority files’ for the wider research community” (Williford and Henry 14). This is a good framework for the evaluation of projects both before and after they have been completed. The creation of authority files is an important goal and hints at an established, growing canon of resources. This model also underlines the importance of creating a data set which is reusable both to outside scholars and to future projects; data in this category occupies the paramount tier in Healey’s hierarchy.
Case Studies

Old Bailey Online

*The Proceedings of the Old Bailey, 1674-1913* is a “fully searchable edition of the largest body of texts detailing the lives of non-elite people ever published, containing 197,745 criminal trials held at London's central criminal court” (*The Proceedings of the Old Bailey*). The project presents the text of “all surviving editions of the *Old Bailey Proceedings* from 1674 to 1913, and of the *Ordinary of Newgate's Accounts* between 1676 and 1772” (Emsley, Hitchcock and Shoemaker, Old Bailey Online - About This Project). The text is fully searchable and digital images of the original pages are also provided. A number of associated resources can also be found in the site, as discussed later.

Recent updates to the *Old Bailey Online* demonstrate manifestations of the concern those in the digital humanities have over topics such as correct attribution and citation. To encourage correct citation the project now includes a page detailing the preferred format. Then the project goes beyond simply having a generic suggested format; there are links on every page and record which provide a complete suggested citation for that object. The citation page explains how to cite the project and website as a whole as well as trials, sessions, and ordinary’s accounts - different levels in the organization of records. This page also contains information on citing associated resources including web pages containing historical background and research and study guides put together by the team. Uniquely, the project’s team also supplies formats for citing searches and “statistics searches” (Emsley, Hitchcock and Shoemaker, Old Bailey Online - Copyright Information and Citation Guide). The creators want people to use this material in new, unexpected ways. They are enabling use in publications and public
venues which require formal attribution of sources. While citations affiliated with each
data piece may be outside the realm of some projects, citation guides are commonplace.
The resources of associated institutions can be leveraged to this end; this technique is
seen in case studies to follow. The team also writes that they are maintaining more
closely lists of staff and their respective contributions (Emsley, Hitchcock and
Shoemaker).

The Old Bailey Online also contains a number of valuable bibliographies. Twelve
of these bibliographies are each “intended as a comprehensive bibliography of
scholarship on the topics covered in the Proceedings” (Hitchcock, Howard and
Shoemaker, Research and Study Guides - Bibliography). This collection of sources was
first published in 2008. However, the creators also established a Zotero Public Group to
allow users to add suggested publications in order that the bibliography might be
regularly updated (Hitchcock, Howard and Shoemaker, Research and Study Guides -
Bibliography). The collaborative spirit of the digital humanities shines through in this
gesture; the people behind this project are asking for input as they create a web of
knowledge associated with their work. Additionally, the project presents a noteworthy
bibliography of works which cite the Old Bailey. At the time of writing this bibliography
included 158 books and articles. This does not include the projects which sprang out of
The Old Bailey Online, such as Locating London’s Past (Hitchcock, Howard and
Shoemaker). This section also has a public forum for recommending new publications.
This is a fantastic example of leveraging the use of digital environment. The project not
only allows users to explore a very large, important set of digitized records, it actively is
promoting bank of knowledge on the subjects covered. Users may ground their work and
simultaneously explore secondary materials along with primary sources. Users are actively becoming part of the big data initiative. In the least they are being linked to it. (Concerns over who makes the cut to be included in these bibliographies could still do with analysis, but the very fact that they exist is heartening.)

The team at the Old Bailey Online worked on one of the eight inaugural Digging into Data challenges. They worked to enhance users’ ability to query, manage, and analyze the records in the Old Bailey Online. The result was Data Mining with Criminal Intent (http://criminalintent.org/getting-started/). The team’s report on their process gives insight into their method of considering the user. This report offers a methodological and theoretical backstory. Their model of project development enhances sensitivity to future data uses. It is particularly interesting that in the report on their work, they cited the “Ordinary Working Historian” (D. Cohen 2) as the guiding principle in the project design. The team sought out this input aiming “to discover what tools and infrastructure would enable the ‘ordinary working historian’ (OWH) to integrate text mining into her or his day-to-day work” (D. Cohen 2). This input was not gathered only once. The team states that: “informally and iteratively, at each stage our development, we asked them to provide feedback” (D. Cohen 20). The team also created personas and “explored a range of interdisciplinary use cases” (D. Cohen 21). This technique is an established one in library reference and user services and user experience design. It makes sense to employ techniques like this in an environment with a heterogeneous user base. The nature of the World Wide Web, where many of these projects are hosted, can help promote accessibility. As discussed, the potential that projects will be used by unexpected populations and is heralded as an exciting part of the digital humanities. The team at the
Old Bailey Online demonstrates that even though an audience may be unknown, they can still influence project development throughout all stages.

The team recognized that in order to sustain this big data project they needed to encourage use through continued creation and improvement of tools for manipulating the data. The team for the Digging into Data challenge decided that simply preparing the project to be accessible to users was not enough to encourage sustainable use. “We concluded that we need to publish serious research that uses these tools to answer historical questions. We need to mash up traditional forms of article writing that engage with historical debates, with evidence drawn from what would have been overwhelming sums of text objects. We need to lead by example, or no one will follow” (D. Cohen 21).

Team investigators used the digital tools they developed and the record corpus they curated to “trace the rise of greater latitude in female behavior in the late Victorian period” (D. Cohen 24, 7) and to discover the most common way of administering poison. (It was often coffee!) These findings serve as templates for future researchers. This teases out the desire to encourage use and weaves the pedagogy of using big data digital humanities into research. The team performed unique research tied to promoting and, through wider use, sustaining the project. Their model is to lead by example and to demonstrate clear, easily recognizable “scholarly mash-ups”.

Perseus Digital Library

The Perseus Digital Library is a longstanding and expansive library coming out of Tufts University. Their “flagship collection, under development since 1987, covers the history, literature and culture of the Greco-Roman world” (Crane “About”). The Perseus Digital Library homepage presents entry points to a sampling of the diverse uses and
resources to be found in the library including curated online exhibits, popular texts, and featured sites. Clearly presented are a number of the supporting institutions associated with the project along with links to contact and support information. However, the central point of the entire page is the long lists of “Announcements.” These posts cover areas such as grants received, new contributions and innovations driven by the project, and details on technical overhauls and updates. The Perseus project’s mission is to create a digital library. The project was conceived as “a practical experiment in which we explore possibilities and challenges of digital collections in a networked world” (Crane “About”). The frequent updates and embrace of project and sub-project documentation is vigorous. The About section of the website contains a page with associated project publications dating from 1987 to the present. This bibliography is similar to the bibliographies found in the Old Bailey Online. However, the scope is much more limited. Cultivating this open documentation is a way of tackling some of the challenges articulated in the introduction.

While the information is open to the public, Tufts University does maintain copyright on materials held in the digital collection. Listed exceptions to this include open source code, public domain texts, and some search and analysis results (Perseus Copyrights & Warranty). Citation guidance and examples are provided for the site as a whole and for individual artifacts. This is a contrast with the Old Bailey Online’s suggested citations tailored for every respective record. The Perseus project promotes use of their value-added big data corpus; however, this can present a challenge as the materials included are under greater restrictions than projects which deal exclusively with records in the public domain. The project succinctly handles issues of attribution by recommending linking to the site whenever possible. They clearly state their reasons:
We welcome all links to the Perseus site and to any page therein provided that you properly attribute the source of the data and add a pointer, whenever appropriate. Links to Perseus are always preferable to any copying of materials. Links provide a foundation for scholarly conversation across the WWW, linking avoids confusion of source attribution and gives credit for the work of the project, and some materials in Perseus are governed by special copyrights which prohibit reproduction outside of the Perseus environment, and thus should never be copied onto another site without permission. (Perseus Copyrights & Warranty)

This is a strong way of encouraging correct citation. It is similar to other case studies in spirit though not in technique. The use of linking is a strong practice which uses the power of digital projects to good advantage. One criticism of is that the availability of these resources is not guaranteed by the project. This could pose problems for researchers who come to depend on cited links. Nevertheless, encouraging use of a big data set tempered by an understanding of this data’s confines does not unduly limit sustainability.

Manipulation of large datasets can often be made difficult by the degree of dirtiness found in the data. The Perseus Library allows insight on the interoperability promoted in this paper. Old Bailey Online’s Digging into Data project showcased an example of a project which built upon a previous data set. In this case they created a tool. What happens when a dataset is found insufficient for the planned research? The Perseus team performed an evaluation of their digitized materials’ metadata. They discovered that there were a number of mislabeled items: “As others have pointed out…problems plague these massive collections in their use for scholarly research, not only in the quality of the image scans and the resulting OCR but also in the metadata itself that describes the texts” (Bamman and Smith 2.2). The team was particularly focused on ensuring that the language of each text was correctly reflected in the metadata. In a Digging into Data project, the team chose to tackle this metadata issue and the associated problem of misattributed text language. A bit on their technique follows:
In order to solve both problems—that of incorrect or incomplete metadata and of mixed works—we trained an n-gram language classifier … on the complete texts of Wikipedia in 24 different languages and the Ancient Greek and Latin collection of the Perseus Digital Library … and used that trained model to create a language fingerprint for each text, allowing us to pinpoint exactly where the Latin showed up in each volume…While much of this research operates on the textual data itself, the ability to chart such movement in both space and time requires information on the place and date of a work’s composition. (Bamman and Smith 2.2-2.3)

The team trained using data from their own project and the resources of others. This work was necessarily done to augment existing digitized texts so that users could accomplish further research goals. As this project was being applied to a limited corpus, the end goals and accomplishments were clearly laid out. Publications show that the team debated manual versus digitally processing these texts. With every grouping of information, unique challenges in the presentation and manipulation of data present themselves. In the case of this venture, these challenges were leveraged as research questions. The benefits of using well-formed pieces of individual data not only allow for strong manipulation in the intended manner, but set the stage for the easy transition of applications in future. Future work will benefit from this enriched data which approaches the highest tier in Healey’s hierarchy. This is an example of a project working to continue its scholarly work and expand rather than remain a boutique project. The vitality of digital humanities work depends on the sustained lifecycles of projects.

The work done for the Digging into Data challenge necessarily built on previous digital humanities work. The team’s research is an example of how projects can grow and evolve. They targeted a common problem impeding the use of “massive collections” for scholarly research. The team improved the smaller units within the realm of big data and maintained a balanced approach. Big data is composed of little data and metadata. This
research not only fortifies the integrity of the Perseus collection, it creates a method available and is applicable to other research.

**Boston Streets Mapping Directory**

The *Boston Streets Mapping Directory*, also coming out of Tufts, has stitched together modern and historic maps. Onto this landscape, the team geolocates historic images and information mined from city directories (About Boston Streets). This project excels at publishing and making very clear their data set. Their work deals, as they state, with “A Variety of Materials, Many Uses” (About Boston Streets). The project’s website contains detailed technical documentation on various procedures: “These procedures were used by the staff to perform tasks associated with the project. Some of the procedures are specific to particular software or conditions at Tufts or the Bostonian Society. Others are more general in nature. All can be used to understand the workflow and process considerations required for a project of this type” (About Boston Streets). These procedures include documentation of “XML Directory to ESRI Shapefile Conversion” and “Calculating Data Entry Estimates” (About Boston Streets). This documentation allows other researchers not only to use this project’s data, but also its tools. Others can emulate *Boston Streets*’ creation with new sets of raw data. In the digital humanities technical documentation can be as important as a publication detailing the historical research accomplished through digital techniques. Technical documentation makes visible the work performed behind the wall of the computer screen. The technical resources in this site date from 2004. While these resources may be outdated they are still important. Even when this technical documentation become obsolete for the purposes of creating similar projects, it remains a critical account of the project and a testament to the work done by those involved.
**Digital Harlem**

The *Digital Harlem* project “presents information, drawn from legal records, newspapers and other archival and published sources, about everyday life in New York City's Harlem neighborhood in the years 1915-1930” (*Digital Harlem*). The user interface allows users to plot records of people, events and place on a historic map which has been georeferenced to a modern one. These instances can be filtered in a number of ways and premade map layers, such as one mapping churches, are available.

The site does not offer clear recommendations on citing the project. However, the team on *Data Mining with Criminal Intent*, which commented on the need to lead by example, would find some strong examples here. This site is part of the broader project “Black Metropolis: Harlem, 1915-1930”. This project demonstrates how research into humanities topics using digital means fits into other types of research. The team states: “Our analysis of this material pays particular attention to the spatial dimensions of black urban life, breaking new ground in the use of Geographic Information Systems (GIS) to map cultural life. Those maps form the basis of a web site, ‘Digital Harlem,’ derived from our research” (Robertson, White and White). The website *Digital Harlem* grew out of broader research as the investigators began to use digital techniques. This site contains a bibliography of associated publications of this wider research. All associated publication originates from the team of researchers responsible for the site. These publications include a book, articles, chapters in books, presentations, and the website itself. This method of placing the website in the context of more traditional research publications shows how strong digital humanities outputs can be integrated into the landscape of scholarly publishing (Robertson, White and White).
Digital Harlem encourages play and the diversity of mapped information is interesting. However, the large number of digitized records is not immediately accessible, except through their mapping interface. A lot of information is presented, but access is more closely mediated. Though well crafted, elements of boutique design obstruct the easy integration of the team’s work on the backend of the project into the wider community.

**Going to the Show**

*Going to the Show* is a project coming out of The University of North Carolina at Chapel Hill. This GIS-based endeavor aims to “documents and illuminates the experience of movies and moviegoing in North Carolina from the introduction of projected motion pictures (1896) to the end of the silent film era (circa 1930)” (Moviegoing in North Carolina). Historic Sanborn Fire Insurance Maps have been knitted onto current maps in Google Earth. This has been done with maps from several years for a number of North Carolina cities. Venues and other items, such as postcards, are then plotted onto these historic maps.

*Digital Harlem* integrates itself in broader landscape by placing the project under the umbrella of existing research. *Going to the Show*, on the other hand, draws that research into the project. The web site includes GIS mapping and also serves as a platform for delivering background research on pertinent subjects. *Going to the Show* contains a thorough, well-linked bibliography of scholarly sources draw on in the creation of this content (Going to the Show Sources). Unfortunately, this resource is only accessible when a hyperlinked citation is clicked. This limited access is problematic. Like many digital humanities projects with a pedagogical air, this site encourages the exploration of linked scholarship. The project’s blend of sources initiates the creation of
new connections by its users. Aiding this is a great deal of linkage between resources within the project. This very linked project shows on a micro level how deep thinking into detailed areas can be drawn into a larger context.

*Going to the Show* is a part of a larger digital initiative: Documenting the American South (Doc South). As such, it links to generic Doc South information pages concerning the copyright, usage, and citation of digital resources. Citation recommendations are offered in both the Modern Language Association and Chicago Manual of Style formats (The University Library, The University of North Carolina at Chapel Hill). These are critical attributes for a project to have. Although the original source material is clearly documented, it is less clear how to cite the project’s value added management of this data found in geolocated maps. This citation guidance is not as immediate as that of the Old Bailey, but is, perhaps, a more realistic model for many digital projects hosted within a larger institution. Big data resources are growing. Increasingly, projects can point outward to existing resources in other institutions. As trends of documentation continue, these resources will only increase. Sanborn fire insurance maps are incredibly rich resources. These resources were created for cities across the United States. By focusing on North Carolina, these researchers are using only a subset of the source material. Similar projects could be initiated with a different subset or the entire body of records. However, if the team at *Going to the Show* documents their technique, future projects handling similar materials will not have to recreate established work.

It is noteworthy that each of the site’s pages link to an established persistent Uniform Resource Locator (URL). This contrasts with the Perseus Project which does not
guarantee that resources will remain accessible. Furthermore, a link designates the option to open in Google Earth. This means that any user can, after installing free Google Earth software, fairly directly access and manipulate the value-added maps without the mediating layer of a search interface. In fact, although this feature is labeled “Open in Google Earth,” clicking on the “Open in Google Earth” link prompts users download a Keyhole Markup Language (KML) file. KML is a standard Extensible Markup Language (XML) format for describing geographic information. This means that these georeferenced maps and associated data points are also able to be opened on platforms other than Google Earth. This feature is significant because it acknowledges the work done to georeference this content. By granting users this data to download, the projects creators are allowing future scholarship built on their own work. This openness is perhaps one of the benefits of projects evolving as pedagogical institutions. Some projects are created with the understanding that their stewardship may be passed on in the future; that understanding encourages well-formed documentation of past contributors and sources since firsthand witnesses may move on.

In the Going to the Show model, the project was placed into larger contexts: the Documenting the South collection, the University of North Carolina at Chapel Hill Library system, and beyond. This allows the project creators to leverage certain inherited resources which aid in making a strong project. The success and future of digital humanities projects depend on continued support. It is important to tie projects to existing resources in their environment. This moves them towards the status of a permanent fixture.
Summary

Three main areas emerge which, in different ways, allow one to gauge how well various digital humanities projects are leveraging the potential of big data and creating projects which are scalable, accessible, and able to evolve. These three parameters are technical documentation, citation recommendations, and bibliographies.

Technical Documentation:

Internal documentation should always be kept to ensure that a project can be properly maintained. Making these records public is a good way of establishing the work done by team collaborators and creating resources for future projects. This aids in promoting the work of all team members throughout the process. Digital humanities projects are collaborative in nature and there is a two-fold benefit of documenting team member’s work. The work they establish can be reused and recycled into future projects. This will eventually banish expending energy and resources on purely boutique projects. The other benefit comes from the fact that as people’s hard work is acknowledged, their careers flourish and their personal dedication to the projects increases (Nowviskie).

Citation Recommendations:

The presence of citation recommendations helps serve as a metric of encouragement for greater use and incorporation of data. Several models emerged in the case studies. Interestingly, these models reflect some techniques for project sustainability on the whole. From the Old Bailey’s complete integration of record level citation to the Perseus Project’s advocating for linking it is clear that designers are aware of the need to facilitate proper acknowledgement. Solutions for promoting citation are available which are sensitive to both the project resources and material restrictions.
Bibliographies:
The promotion of associated materials results from feedback loops of scholarship by both internal and external users of project resources. There are three general categories of parallel research found in case study bibliographies. These categories are: background material on the contents of the project, ongoing research by project team members, and work generated by external users of the projects. These features encourage continued use; however, to be most effective, there must be mechanisms in place to ensure that they, like the project itself, are maintained and updated.

Technical documentation, citation recommendations, and bibliographies are explored in the five case studies. Several techniques for each were discovered. This paper explores the play between big data sets and the digital humanities. The evaluation of size or diversity of data sets is one way to approach analysis. However, the way these resources are presented to users is an equally important characteristic. The continued vitality of these projects depends on sustained use which can be encouraged or impeded through each of these three areas. Hopefully, this paper consolidates some of these examples and generates discussion on their respective strengths and weaknesses.

Discussion
The landscape of scholarly publications is changing. Reevaluation of the merit and analysis of scholarship includes taking into account the changing “outputs” of scholarly activity. The Modern Language Association’s “Minimal Guidelines for Authors of Web Pages,” last reviewed in 2002, includes recommendations for “the minimal
reference information that should be provided in Web pages intended for use by students, teachers, and scholars in the modern languages” in order to “help readers to use, evaluate, search for, and cite information found on the web” (Modern Language Association).

Many of these suggestions were met by the projects reviewed. When it came to citation and copyright features these projects often greatly excelled the minimum guidelines.

The Modern Language Association also saw a need for parameters of evaluation for work in this field. They have put out “Guidelines for Evaluation Work in Digital Humanities and Digital Media” which was most recently reviewed January, 2012 (Modern Language Association). Speaking to the change wrought by new technology, the MLA says: “Digital media are transforming literacy, scholarship, teaching, and service, as well as providing new venues for research, communication, and the creation of networked academic communities” (Modern Language Association). The concept of networked academic communities is clear in the case studies’ usage of links, bibliographies, and the foundations created for future projects; these are all features found in the projects surveyed. While the MLA’s guidelines are generic recommendations, they are directly related to the topics reviewed in these case studies.

Two of the three recommendations to candidates and faculty members who wish to demonstrate their work in the digital humanities encourage documentation. The first form is documentation of their role; in many of the case studies, this was accomplished through both publication credits and the listed descriptions of individuals’ contributions to projects. The MLA’s second encouraged form of documentation advises faculty members to explain their work. This includes describing “the process underlying creation of work in digital media (e.g., the creation of infrastructure as well as content) and their
particular contributions” (Modern Language Association). Scholars should also document “how work in digital media requires new collaborative relationships with clients, publics, other departments, colleagues, and students” (Modern Language Association). The projects reviewed contain a variety of documents which help these end goals. These range from dynamic bibliographies of related scholarship to easily downloaded, multilayered, value-added historical maps. Projects of from scholars in the humanities display the hyperlinked possibilities of the digital age. Examining cases studies such as these allows for those making decisions regarding promotion and tenure to gain insight into what types of documentation to look for in the projects of their own institution.

There is perhaps some discomfort with the idea of releasing a marked-up text or georeferenced map which has been much labored over, for fear of losing credit for this work. (Again, this is why offering information both on datasets and citation recommendations is crucial.) In light of this concern, it is easier to see the power in Going to the Show’s decision to make KML files for their maps available. We cannot always predict how existing research will be used; perhaps if we could we would be doing the work ourselves. The fact that data sets are often used in ways previously unimagined has certainly not gone unnoticed:

When the collected published works of Abraham Lincoln were posted online a few years ago, the director of the Papers of Abraham Lincoln, Daniel W. Stowell, said he expected historians to be the most frequent visitors to his project’s site. But he was surprised to discover that the heaviest users were connected to Oxford University Press; editors of the Oxford English Dictionary had been searching the papers to track down the first appearance of particular words. (P. Cohen)

Given this observation, allowing as flexible an approach as possible to data sets makes sense. While finely honed, faceted search interfaces and unique visualization abilities are important, so too is allowing fellow researchers and future users to get their hands dirty in
the raw data. For those projects envisioning used by teachers and students, the concept of access and ability to explore enhances the future learning experience. Scholarship should aim to teach, to reveal to others the knowledge gained and created during study. In the digital humanities, this can mean documenting and making accessible the work behind the curtain.

By documenting the inner workings of big data collections, persons working on the project can more directly demonstrate their output. Some work may still need to be done adjusting the perception of digital humanities work. Many of the projects highlighted are quite large and well-established. The larger, established projects such as the Old Bailey Proceedings Online and the Perseus Digital Library are setting the stage for current and future projects. Projects of different scales may not have comparable support and access to resources. However, the examples described in this paper are just that: examples. The techniques and features highlighted are shown to be adapted to the unique needs of respective projects. The different models for recommending citation style are indicative of larger models for projects. Teams provide citations throughout the site or send users to already established resources on how to create citations themselves. Both these solutions are valid. They help safeguard against their project existing unmoored to future scholarship.

Every digital humanities project creates, to some degree, a digital library of information from which to pull on and manipulate. For scholars to work digitally, raw materials must be available digitally. Often this means transforming previously physical items into digital versions through photography, scanning, or transcription. It can mean gathering born-digital material or creating something altogether new. In any case, raw
data must usually be processed, cleaned, and augmented. It must to fall into a, hopefully high category in Healey’s hierarchy of data. The creation of this value-added big data can be as important as the journal article published about on a trend found through digital humanities’ techniques.

As scholars have built up this humanities research in the digital age, resources have grown. While recent publications like *Debates in the Digital Humanities* are still trying to put to bed discussions over what exactly constitutes digital humanities, a body of textbooks, criticism, funding organizations, graduate programs, and peer-reviewed journals has been established. So too have online resources for some of the common sources and problems in the Digital Humanities. For instance, several projects direct users to resources created and hosted by the University of Virginia on Sanborn© Maps. This is one of the benefits of handling big data. Flexible and innovative use of these data sets can gesture to the same resources leading to only one set of documentation.

**Future Research**

This study has targeted projects which cultivate a public user interface and contain to varying extents the spirit of public libraries and schools. This is because they are more likely to embrace the open documentation techniques explored in the paper. These case studies were evaluated from the perspective of a prospective user – casual, formal, student, and professional researchers of any level. Important issues outside the scope of this paper include evaluations of accessibility and ADA compliance. Another associated issue is the consideration of the digital formats of materials and digital preservation best suited to sustain big, digital humanities.
George H. Williams advocates for universal design and greater consideration of the “needs of people with disabilities” (Williams 206). To cap off a list of compelling reasons for accessibility, Williams writes:

It is the right thing to do. We recognize the importance of open access for scholarly materials…we would never use a proprietary format for preserving and sharing our work, in part because to do so would be to exclude those people who cannot afford or do not have access to the necessary software to use that format. However, few of us think twice about whether or not the format we have chosen and the design choices we have made exclude disabled people. (Williams 206)

Williams proceeds to detail the reasons why integrating universal design is necessary. A survey of projects elaborating on this topic could yield greater insight on the specific techniques and benefits of engaging in design which is truly accessible. Beneficially, this subject is closely linked with the design needs of mobile technology, another important consideration (Williams 206).

Storage and preservation are often invisible pressures on digital content. Well-formed data sets encourage the potential of new projects and ease of transfer between scholars. Many of these formats and conventions also make for data which is well-prepared for preservation through, for example, migration of content. The embrace of enveloping data in formats based in extensible mark-up language, such as the KML files of Going to the Show, is an embrace of a widespread and flexible convention in the digital world. Since XML is so prevalent, data described by it is more likely to be data which can be read and manipulated for a long time. A sight longer than data formatted in obscure or project specific ways. Furthermore, the vitality and continued use of collections in different ways encourages their continued upkeep. The Perseus project has been and ongoing effort for two decades now. It is dynamically growing their collection and scholarship while maintaining open access which has helped keep their project alive
Conclusion

The work of those in the digital humanities is to ask old questions with new tools. The purpose and importance of research and thinking in the humanities remains the same; it has been strengthened by an infusion of new tactics and perspectives fostered by digital technology. Ensuring that the initial work done on these projects is maintained can also be accomplished with digital technologies. The purpose of examining these case studies was to tease out concrete examples of sustainable digital humanities projects grappling with big data issues. Through the course of examining five different projects, three areas were found which yielded rich discussion of big data issues and the digital humanities.

The technical documentation found in many projects helps to dispel the creation of boutique projects; furthermore, this practice promotes an understanding and appreciation of work done in the digital humanities. Design standards and rigor in citation practices help successfully integrate big data scholarship into the culture and practice of the field. These features are part of creating a strong foundation for the future. Bibliographies and similar features encourage ongoing resources to infuse projects with new energy and aid prospective users. The networked nature of the World Wide Web is leveraged to capture collaborative work. This helps create and maintain a web of knowledge and infuses projects with vitality.

The hard work and creative endeavors of digital humanists need to be preserved. The unique preservation challenges need to be addressed. Scholars have always stood on the shoulders of giants and built up their own contributions on those of their successors.
Projects which encourage this to a greater extent are stronger, more long-lasting projects. The intertextuality of the World Wide Web aids in both accessing and attributing sources upon which to build. To maintain viability and use, it is essential that projects respond and adapt to changes in the environment. This includes being perceptive of changes in technology as well to the changing landscape of scholarly publications where many of these projects are judged. Making big data digital humanities projects sustainable means making the work of individuals accessible and acknowledging the efforts of all team members.
Works Cited


