Public Witness Testimony

Submitted to the Senate Appropriations Committee

Digital Humanities Innovation from NEH Funding

Matthew G. Kirschenbaum, Associate Professor of English and Associate Director, Maryland Institute for Technology in the Humanities (MITH), University of Maryland

As a scholar and researcher in the new and emerging field of Digital Humanities, it is my pleasure to offer remarks in support of the role federal funding plays in this work at the University of Maryland College Park, specifically the Office of Digital Humanities in the National Endowment for the Humanities.

Digital Humanities involves the application of computational tools and techniques to large bodies of digital material (so-called "Big Data") drawn from the cultural heritage sector. Examples include textual corpora containing literally millions of books, large-scale image collections digitized by museums, and collections of audio recordings. Digital humanities also involves innovation in new forms of scholarly communication, including social media; and new forms of publication and analysis, such as visualization, GIS mapping, and 3-D animation or reconstruction. Crucially, Digital Humanities emphasizes public and outward facing access, with open Web resources being the most typical form of disseminating this work. Training in Digital Humanities tools and techniques, meanwhile, is helping revitalize undergraduate and graduate education in the humanities at moment when the public at large increasingly sees the humanities as at best secondary pursuits alongside supposedly more pragmatic STEM fields.

At the University of Maryland, we have enjoyed the benefit of funding from several federal agencies in this work, including the Institute for Museum and Library Services, the Library of Congress, and the National Science Foundation. However the National Endowment for the Humanities is unquestionably the single most important federal funding source for Digital Humanities research. Since 2009 alone, MITH at Maryland has been the beneficiary of over a dozen NEH grants, the majority from the Office of Digital Humanities, ranging in value from roughly $25,000 to $400,000 and totaling approximately $1.26 million. These have supported projects from (at the low end) "Enhancing Music Notation Accessibility" and "Data, Biomedicine, and the Digital Humanities" to (at the high end) "Building an Accessible Future for the Humanities" and "The Text-Image Linking Environment," as well as the "Shelley-Godwin Archive." This last project is representative of Digital Humanities work: on a free and publicly accessible Web site (http://shelleygodwinarchive.org/) it makes available all of the notebook and manuscript materials involved in the composition of the classic novel *Frankenstein* by Mary Shelley. Members of the public can follow the composition of the novel line by line as Mary worked with her equally famous husband, the poet Percy Bysshe Shelley, to edit and revise the manuscript; the site also allows the public to participate in transcribing the manuscript materials themselves, subject to the oversight of the expert editorial team. A formerly cloistered scholarly activity, restricted to those with privileged access to reading rooms and special collections, has been opened to anyone with the interest and an internet connection.

The remainder of this testimony will describe how an extremely modest initial investment on the part of the NEH (under $12,000, one of the smallest grants the agency has ever awarded) has led to substantial ongoing innovation. The project was entitled "Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use."

Today nearly all published poetry, fiction, and drama is *born-digital* in the sense that the text is composed with a word processor, saved on a hard drive (or other computer storage media), and accessed as part of a computer operating system. True, some writers will still employ longhand or even mechanical typewriters as a step in their composition process, but sooner or later the text will be keyed into a computer, almost always to be further revised. Often the text is emailed to an editor, along with other correspondence. Editors edit electronically, inserting suggestions and revisions and emailing the file back to the author for approval. Publishers use electronic typesetting and layout tools, and only at the very end of this process is the electronic text of the manuscript (by now the object of countless transmissions and transformations) produced as the static material artifact that is a printed book. This new technological fact about writing is already having an impact, from office work to government and the academy to literature and the creative arts. President Obama's use of a Blackberry and the implications for the Presidential Records Act is a high-profile example of how the public is coming to terms with the consequences of born-digital authorship. In the particular realm of literature and literary scholarship, this means that writers working today will not and cannot be studied in the future in the same way as writers of the past, since the basic material evidence of their authorial activity—manuscripts and drafts, working notes, correspondence, journals—is, like all textual production, increasingly migrating to the electronic realm.

Meanwhile, librarians and archivists have not failed to take note as computer storage media, as well as entire computers, have begun arriving on their doorstep as an increasingly routine part of the acquisition of an author's "papers." Notable authors represented with at least some born-digital material include John Updike, David Foster Wallace, Russell Banks, Samuel Beckett, Lee Blessing, John Crowley, Robert De Niro, Michael Joyce, Thomas Kinsella, Bernard Kops, Norman Mailer, Terrence McNally, Tim O'Brien, Salman Rushdie, Ronald Sukenick, Leon Uris, Alice Walker, and Arnold Wesker.

To date, however, the activity associated with processing such born-digital material has been localized and idiosyncratic, and, at least in the US, without much cross-communication among the different archives and repositories involved; moreover, the archives and repositories, for their part, have not yet addressed these challenges with the scholars who will seek to access born-digital literary material in the years to come. Literary scholars are going to need to play a role in decisions about what kind of data survives and in what form, much as bibliographers and editors have long been advocates in traditional libraries settings, where they have opposed policies that tamper with bindings, dust jackets, and other important kinds of material evidence. This grant therefore brought together scholars, archivists, digital curators, and technical personnel associated with three significant born-digital collections for a series of targeted site visits and planning meetings at each of their respective institutions, with the goal of working towards a larger project proposal designed to address the needs of both archivists and scholars in this new milieu.

Digital Humanities Level 1 Start-Up funding ($11,708) was received in support of a series of site visits and planning meetings for personnel working with the born-digital components of three significant collections of literary material: the Salman Rushdie papers at Emory University's Manuscripts, Archives, and Rare Books Library (MARBL), the Michael Joyce Papers (and other collections) at the Harry Ransom Humanities Research Center at The University of Texas at Austin, and the Deena Larsen Collection at the Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland. The meetings and site visits were undertaken with the two-fold objective of exchanging knowledge amongst the still relatively small community of practitioners engaged in such efforts, and facilitating the preparation of a larger collaborative project proposal aimed at preserving and accessing the born-digital documents and records of contemporary authorship.

While there have been numerous studies of the impact of computerization on composition, these often present the computer as simply another tool or instrument. In fact, a computer functions much more like an environment—or a writing space, to use a term popularized by Jay David Bolter. Access to an entire computer is not unlike having a key to an author's study or workroom. On the one hand, this creates an awesome burden of responsibility for the archivist, since all manner of sensitive personal information can be inappropriately exposed. Sometimes donors will make their wishes explicit in this regard; but what if they don't? Should a researcher be allowed to see an author's choice of desktop wallpaper? After all, scholars have traditionally been interested in the physical setting in which an author worked, even to the level of such details as furnishings, decorations, and, yes, wallpaper. Other examples begin to enter into the realm of forensic information recovery. A computer's registry, for example, stores information related to all of the device drivers and application software in the operating system. Access to the registry is among the outsider could undertake; but its value as a record of the digital environment of the computer is enormous. At the level of individual works, scholars will surely want to examine a file's properties, which contain records of when it was last opened and closed and how many hours and minutes was spent accessing it. This kind of metadata, while hardly infallible—it could be spoofed by something as simple as an incorrect system clock—could, with care, be used to establish chronologies that could date the composition of a work—or specific passages within a work—to the hour, minute, and second.

Likewise, various word processing packages incorporate "track changes" features which preserve a record of a document's internal edits, as well as marginal commentary. Track Changes is already a widely used editorial tool, and can systematically capture the kind of revision history for a document that had heretofore been available only incidentally. Given the ease with which multiple versions and drafts can be saved—often this occurs automatically as a function of the software—it is not hard to imagine a scenario in which a scholar may potentially have access to hundreds or even thousands of versions of the same work, and be faced with the prospect of discovering what significant differences between them actually exist. Here we may see textual scholarship begin to draw heavily on text mining and visualization, methods which are specifically aimed at sorting and sifting large volumes of data. For example, a scholar might use a combination of data mining and visualization to discover "hot spots" in the evolution of a work, points at which especially significant revision activity took place.

The meetings and site visits supported by NEH led to the production of a extended white paper which has been downloaded many thousands times and is routinely cited in the emerging professional literature on the archival processing of born-digital literary and cultural materials. This extremely modest investment by the NEH has also led to additional funded research, notably an effort to develop digital forensics tools for archivists so as to begin actually addressing the concerns articulated in the white paper.

Digital forensics is an applied field originating in law enforcement, computer security, and national defense. It is concerned with discovering, authenticating, and analyzing data in digital formats to the standard of admissibility in a legal setting. While its purview was once narrow and specialized (catching black hat hackers or white collar "cybercriminals"), the increasing ubiquity of computers and electronic devices means digital forensics is now employed in a wide variety of cases and circumstances. The floppy disk used to pinpoint the identity of the "BTK Killer" and the GPS device carried by the Washington DC sniper duo—both of which yielded critical trial evidence—are two high-profile examples. Digital forensics is also now routinely employed in counter-terrorism and military intelligence.

But the same forensics software that indexes a criminal suspect's hard drive allows the archivist to prepare a comprehensive manifest of the electronic files a donor has turned over for accession; the same hardware that allows the forensics investigator to create an algorithmically authenticated "image" of a file system allows the archivist to ensure the integrity of digital content once captured from its source media; the same data recovery procedures that allow the specialist to discover, recover, and present as trial evidence an "erased" file may allow a scholar to reconstruct a lost or inadvertently deleted version of an electronic manuscript—and do so with enough confidence to stake reputation and career.

Based in part on the successful work done in the NEH project, the Andrew W. Mellon Foundation has provided extensive funding for the University of Maryland's participation in the BitCurator project. The BitCurator project began on October 1, 2011. It is a joint effort with the School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS) and Maryland Institute for Technology in the Humanities to develop a system for that incorporates the functionality of many existing open source digital forensics tools. The BitCurator project is thus an effort to build, test, and analyze systems and software for incorporating digital forensics methods into the workflows of a variety of libraries, museums, and archival institutions. BitCurator currently enjoys extensive uptake in the practitioner community, and is allowing both archivists and scholars to take active steps toward the preservation of born-digital cultural heritage. The early and prescient support from the Office of Digital Humanities at the National Endowment for the Humanities significantly contributed to the articulation of community goals that this research is now answering.

Reference URLS

Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use
https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-50346-08

The BitCurator Project
http://www.bitcurator.net/

The Maryland Institute for Technology in the Humanities
http://mith.umd.edu/