Asking Questions and Building a Research Agenda for Digital Scholarship

Amy Friedlander

At a critical moment in the movie *The Unbearable Lightness of Being*,[1] the audience sees a glass tremble on a table. Because the story is set in Prague Spring of 1968, we know the rattling glass and then the faint rumble signal the arrival of Soviet tanks. But the tremors go unnoticed by the characters, Tomas, played by Daniel Day-Lewis, and Tereza, played by Juliette Binoche, who are arguing over their disintegrating relationship. She decides to leave him, steps out into the street, and realizes that there has been an invasion.

In this scene, the director Philip Kaufman has engaged in an elegant bit of storytelling that takes advantage both of the audience's knowledge, which exceeds that of the characters, and the attributes of the medium. He uses film, photography, and sound to fill in the story around the narrative conveyed by the script to evoke apprehensive, emotional responses from the audience precisely because they are more knowledgeable than the characters. It is similar to the poignancy that accompanies a good production of *Romeo and Juliet*. Yes, we know it will not end well, but somehow every time, we root for the lovers. In the case of this film, the audience's foreknowledge is triggered by adroit use of the camera and the medium rather than by familiarity with the story.

So it is with computation and humanities scholarship. We have inherited a cyberinfrastructure of systems, data, and services that arose from and is optimized for research in science and engineering. As a result, humanists have access to technology but are in search of questions: What scholarship becomes possible when, from their desktops, scholars can access vast stores of admittedly highly heterogeneous data together with powerful capabilities for analysis and presentation? In the terms set by this scene, how do we use comput-

[1] The film is based on a novel of the same name by Milan Kundera. The novel was written in 1982 and published in Paris in 1984.

ers as adroitly as the director used the camera to enable research that takes advantage of the capabilities of the technology to tell "the story"—to conduct research and convey findings—in new and important ways? To get beyond, as one participant in a September 15, 2008, symposium on promoting digital scholarship sponsored by CLIR and the National Endowment for the Humanities (NEH) said, treating the computer like "a black box"?

Cyberinfrastructure and Scholarship

We find ourselves at a tipping point. Several decades of research that combines humanities scholarship with computational resources are accumulating into a transition from a field characterized by a series of interesting projects to one that is more cohesive, collaborative, and less confined to the interests of a relatively small number of scholars. Organization of NEH's Office of Digital Humanities signals coalescence of support behind applications of information technology to topics in the humanities. The American Council of Learned Societies' 2007 report, *Our Cultural Commonwealth,* reflected a broad interest in the cyberinfrastructure of research and articulated a sense that the nature of computationally intensive research transcends traditional boundaries. And increased awareness of the value of collaboration is evidenced by the organization of centerNet, Project Bamboo, and a workshop jointly sponsored by the National Science Foundation (NSF), the Institute of Museum and Library Services (IMLS), and NEH in October 2008 on Tools for Data-Driven Scholarship. The National Digital Information Infrastructure and Preservation Program (NDIIPP) at the Library of Congress, its network of partners, and mass digitization projects together with the proliferation of more traditional text conversion and markup projects have created collections of information in digital form in a quantity, diversity, and scale hitherto unknown as well as a community of scholars, librarians, and archivists with a common interest in long-term preservation of digital content. Finally, a generation of young scholars who are comfortable with computational techniques has begun to change the intellectual complexion of traditional faculties, although access to facilities and resources are still unevenly distributed.

These young scholars can feel ghettoized and even disadvantaged when seeking grants and when promotion and tenure review committees evaluate their computationally intensive work.[2] Indeed, many digital humanities centers studied by Diane Zorich in her report for the Scholarly Communication Institute in July 2008, which is summarized in this anthology, were founded in part to provide a sense of community for these scholars. Somewhat paradoxically, these centers now risk becoming silos and may constitute barriers

[2] One participant in the September 15, 2008, meeting argued against the distinction between "humanities" and "digital humanities," noting "Aren't all scholars digital in some ways, even if they simply use the Internet to search?" We agree that this is an important point but have retained the phrase, "digital humanities," since it is now commonly used to identify a specific kind of scholarship.

to the evolving trans-institutional cyberinfrastructure, collaboration, and resource management necessary to achieve efficient allocation of expensive resources and to enable research at a scale that takes into account the wealth of heterogeneous digital source material as well as computational and analytical power.

Zorich's report is part of an extended, distributed conversation that CLIR has sustained over the last 18 months. This conversation ranged broadly over the confluence of cyberinfrastructure, scholarship, and collections, in particular the preservation of those digital collections to enable access, verification of results, and reuse and repurposing of materials. CLIR sponsored two major events, in addition to its contribution to the annual Scholarly Communication Institute. The first was a one-day workshop in November 2007, Promoting Digital Scholarship: Building the Environment, which resulted in a report, *Many More than a Million: Building the Digital Environment for the Age of Abundance* (Crane and Friedlander 2008). The second, mentioned earlier, was the CLIR-NEH symposium, Promoting Digital Scholarship: Formulating Research Challenges in the Humanities, Social Sciences, and Computation, held September 15, 2008. The administrative report, which includes an account of the day's discussion, has been posted to the symposium Web site, where the prospectus, agenda, and list of participants and their brief biographies are also located.[3]

Discussions at the November 2007 symposium had focused on issues that arise as a result of mass digitization projects. Among the recommendations was a call for the articulation of "marquee" research questions, analogous to the grand challenge questions in the sciences, which provide large-scale intellectual coherence without constraining individual or unique projects. This call led directly to the September 2008 symposium, which invited about 30 scholars across the humanities, social sciences, and computer science to look squarely at the role of research questions in promoting new scholarship. The white papers commissioned to frame the discussions appear in this volume and, together with themes in the discussions themselves, form most of the content of the research program described in the remainder of this chapter.

There exists an important but often-ignored distinction between the research programs that rely on an infrastructure and the research infrastructure itself. The term "cyberinfrastructure" originated in a report by the NSF, where it is defined as the comprehensive infrastructure required to capitalize on advances in information technology, which "integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools" (NSF 2007, 6). The ACLS subsequently adopted the term in *Our Cultural Commonwealth*, and the word has crept into routine discourse in higher education and advanced research. While there is an intimate connection between the instrumentation, software tools and

[3] See http://www.clir.org/activities/digitalscholar2/.

platforms, resources and facilities, on the one hand, and the research programs on the other, they are, nonetheless, distinct. Yes, research is conducted on the infrastructure—how to make it better, faster, more reliable, and, in a sense, smarter. But that work is distinct from the research that the infrastructure has been invented and optimized to enable and support. So if the infrastructure answers the question, *how*?, the research program answers the questions *what?* and *why?*

Undoubtedly tools are important. They are common features of digital humanities centers, can do many things researchers want to do, and are concrete. They can be evaluated and compared according to agreed-upon protocols, like the "evaluation-guided research paradigm" that Douglas Oard describes in his essay, which consists of the challenge problem (perhaps a set of texts to be classified), the answer key (the correct answers), and the evaluation measure (the fraction of the system's assignments that are considered "right").[4] Humanists have developed a plethora of tools of varying quality, few of them apparently used by more than a relative handful of scholars.[5] In response, one of the symposium participants recommended that digital humanists generally had to become more disciplined about evaluating the utility of their tools. Indeed, Project Bamboo and the October 2008 Tools for Data Driven Scholarship workshop are steps in precisely that direction.

But tools can also deflect attention. "Are we letting our anxieties about tools and protocols and methodologies obscure bigger questions?" a scholar of medieval literature asked in the CLIR-NEH September 2008 symposium, before observing that methods, protocols, and disciplines gradually evolve only after the need for a function or capability has been perceived. Historically, research has driven the development of cyberinfrastructure, whose roots trace back to the development of computer networking in the 1960s and advances in high performance computing in the 1980s. These technologies enabled organization of distributed research teams and access to data and other resources as well as computationally intensive analysis in a range of fields in the life sciences, social sciences, and physical sciences. After listening to some of the discussion, one of the computer scientists at the September 2008 symposium suggested that humanities scholars need to "get to the next level of problem definition, perhaps talking about the tasks they need solved (such as finding something particular in text) rather than the system they need built." This comment resonated with a recommendation from another computer scientist, a specialist in human-computer interfaces and design, who advised humanists to be able to answer the question, "What is it you

---

[4] It is used for example in the well-known TREC competitions, run annually by the National Institute of Standards and Technology (NIST), which seek to support research within the information retrieval community by enabling large-scale evaluation of text retrieval methodologies; see Text REtrieval Conference, http://trec.nist.gov/.

[5] Humanities' tools have not been systematically studied; this occasioned the workshop on Tools for Data-Driven Scholarship in October 2008. One example of evaluation of two aspects of tools—their findability and usability—is Nguyen and Shilton 2008. CLIR has commissioned a follow-up study on tools and infrastructure, which is scheduled for release in the summer of 2009.

are trying to do?" and to explain the kinds of evidence that would be necessary to adduce to answer a given question rather than focusing on the available technologies or the technologies they believe are available. Embedded in these remarks are different notions of what constitutes a question. Indeed, questions exist at many scales, and famous scientific grand challenge questions (for example, the relationship between electricity and magnetism) in practice resolved into a series of questions that converged on an answer over the centuries.

Humanists do not lack for questions. For example, Gregory Crane wants to understand how the contemporary Islamic Republic of Iran arose from the Persian Empire of antiquity, a question that requires an enormous array of disparate sources in many languages spanning centuries. Anthropologists and archaeologists want to delineate the prehistoric migrations to the Americas. Medieval scholars want to plumb the surviving manuscripts and compare them in ways not possible in analog and thus reinterpret the texts themselves. In so doing, Stephen Nichols has argued, the modern reader confronts the original texts the way the original readers did—without the intermediary of the nineteenth century standard editions (Nichols 2008).

Posing questions at the right level of abstraction, as suggested by one of the participants, is non-trivial. Answering "big" or "marquee" questions that provide high-level coherence and allow individual scholars to find common ground with others engaged in related research requires experimentation as well as consensus building. The next step of parsing these marquee questions into operational questions is its own intellectual exercise that may involve exploration to see what exists or happens when a technique is tried before a formal research project is posed. Moreover, the term *humanities* is misleading in the sense that it imparts high-level unity where in fact, humanities scholarship subsumes an array of disciplines from archaeology and art history to literary criticism to history of science, each with its own literatures, methods, and traditions. Yet there is a sense that there is sufficient common ground to articulate a shared infrastructure of tools, services, and collections that would reduce unnecessary redundancy, allocate human and information resources efficiently, and, most interestingly, enable a different kind of scholarship.

Caroline Levander makes the latter point when she argues that the deep significance of Our Americas Archive Project (OAAP) is its ability to restructure the categories of knowledge precisely by restructuring collections related to the Americas and hence access to materials, so that the structure of the collections helps scholars "pry [their research] loose" from the self-limiting assumptions of the nation state. Some at the symposium suggested that boundaries be cast differently, perhaps, for example, to see the Atlantic world as a historically coherent framework of population and economic interchange rather than defining the scope as continental landmasses. Nevertheless, no one quarreled with her fundamental insight: that the organization of collections is inherent in the way that research is framed, that such organization of knowledge bounds the way

that research is then undertaken, and that challenges to conceptual boundaries can sometimes begin with the organization of source material.

Questions and Collaborations

If this new scholarship is to be more than a series of boutique projects that use computers, one component must be a set of organizational topics and questions that do not bind research into legacy categories and do invite interesting collaborations that will allow for creative cross-fertilization of ideas and techniques and then spur new questions to be pursued by colleagues and students. Collaboration across traditional boundaries is particularly important for ambitious projects that require years of research and cannot be summed up in a single dissertation or monograph. However, collaboration is a social as well as an intellectual process and can be difficult for many reasons, some of them having to do with institutional and disciplinary cultures, language and terminology, mental models about the research process, trust, appropriate credit, and a sensible allocation of tasks.[6] For example, Andreas Paepcke points to the "agenda mismatch" between the requirements of the domain scholar and the trajectory of computer science research, typically done by a doctoral student. The student's product is usually a prototype; it works "well enough." "Well enough" is probably not sufficient for most humanities scholars, but the time required to create the robust tool is not justified in terms of the student's career path (Paepcke 2008).[7] So one metric that this computer science researcher uses to determine a suitable collaborative project is the project's ability to yield publishable research in peer-reviewed journals for *both* lead investigators.

The key is the appropriate level of abstraction, that is to say, questions and topics that represent major areas of research, are broad enough to embrace a number of related topics, and allow individual researchers to find an intellectual home. They are not so narrow as to constrain the research nor so expansive as to be meaningless. In the discussions that have taken place, we have observed four themes that transcend traditional disciplinary boundaries and resonate with major research topics in computer science: scale, language and communication, space and time, and social networking. The boundaries between them are indistinct, and techniques that are developed in one may apply to problems in another.

---

[6] Collaboration in science has been extensively studied. For example, see Hackett 2005. On the specific issues cited here, see Olson et al. n.d.

[7] Others at the September 15 symposium concurred with Paepcke's observations about mismatches in expectations between computer scientists and domain scientists. One researcher said that tools existed that would be of interest to humanists yet using them would be arduous because the interfaces were "abhorrent" and not intuitive to relatively naïve users. She said, "It is not the algorithms but how people can make use of and interact with them that is still so far behind."

Scale

Issues of scale resonate across many disciplines and conversations.[8] The most obvious evidence of scale for humanists is access to heterogeneous digital information of varying quality and in quantities that were unimaginable in prior generations, creating what Crane has dubbed "the million book problem." That is, "even if we could marshal the resources to do so, the human life contains only about 30,000 days—reading a book a day we would only finish a million books after thirty lifetimes of reading. Only machines can process or 'read,' much less analyze, the written record of humanity." So analyzing material at scale requires computation. Scale also means diversity. Collections will increasingly include images, video, and audio, as well as multiple languages, many of them using different scripts requiring transliteration and cross-language capabilities. Some of this information will have been formally ingested into well-managed archives; some will be captured on the fly and deposited into repositories with minimal attention. Making sense of this welter of material implies authenticating the sources through new, automated methods and combining them in creative ways to answer important questions and employing increasingly powerful machines and creative strategies[9] to do so.

Computationally intensive research allows for both very expansive and very detailed investigations. For example, nineteenth-century railroading in the United States has been extensively studied in part because the history can be read as a proxy for the importance of technology and transportation in promoting economic growth, both core questions in economics with clear implications for public policies. Scale allows for both international and subregional comparisons, as Will Thomas, an historian at the University of Nebraska, has suggested (Crane and Friedlander 2008). Scale also allows for greater detail. Dan Cohen, an historian at George Mason University and director of the Center of Technology and the New Media, has pointed out that tracking references to the Bible and/or to specific religious terminology across thousands of text references allows rigorous examination of the secularization thesis, which states that the role of religion declined in general discourse during the nineteenth century. Other topics might include analysis of the poetry cited in popular literature such as magazines and newspapers, or the changing role of Shakespeare as seen by the plays mentioned and passages quoted (Crane and Friedlander 2008).

---

[8] Scientists face a "tsunami" of data, one participant said, and in 2007, the volume of information created is estimated to have exceeded available storage capacity (see Gantz et al. 2008). Not all of that data should be archived. Nevertheless, current capacity to store, manage, access, retrieve, and repurpose information is reaching its limits. Even IT professionals who focus directly on storage systems acknowledge, "The data center process and archive system is technologically broken. It doesn't scale" (Peterson et al. 2007, 7-8).

[9] For a concise discussion of some of the technologies required to create, store, manage, and analyze large data sets, see Purdue University 2003.

Language and Communication

Developing evidence to the questions posed by Thomas and Cohen relies on linguistic and geospatial techniques, the second and third themes. Language is central to much of humanities scholarship, and many of the early digital humanities projects revolved around mark-up of text converted from analog to ASCII.[10] In addition to the traditional projects that typically combine scanned images with marked-up text, the mass digitization projects are yielding extremely large digital corpora that are both problematic from the perspective of quality (Duguid 2007) and fascinating from the perspective of their content. As Oard explains, human language actually exists in several forms: spoken, written, and character-encoded—that is, the digital representation of language—as well as sign. His paper provides a context for understanding some of the research computer scientists and linguists conduct. It is complemented by the paper by Crane and his colleagues, who examine the role of several of these techniques in the context of classics and philology with the twin goals of increasing scholars' access to more materials while expanding the potential audiences for their work.

The research potential is obvious in several dimensions. The scale, complexity, and heterogeneity of the material challenge researchers to make sense of the data, to find patterns at multiple levels (book, page, paragraph, sentence), detect anomalies, and derive meaning. Such corpora represent a rich source for cross-language studies[11] and create an opportunity for language and text-intensive disciplines in the humanities to become partners in the research process, as Oard argues, because their research materials can also offer challenging problem sets that are central to the way language systems are built and evaluated. Advances in capture technologies and broadened participation in the research process imply that different kinds of content, notably speech, can be taken into multiformat research collections and made discoverable through unified search not only to ethnographers and linguists but also to literary scholars, art historians, archaeologists, and students and researchers who might not otherwise think of these kinds of sources as relevant to their studies.[12]

---

[10] For example, see the rich set of articles in Siemens and Schreibman 2008.

[11] Note that China, Japan, and South Korea combined now account for 27 percent of world research and development (R&D), and China is second in the number of scientists and engineers engaged in research activities. Substantial contributions to the global scientific literature may not be published in English. In this context, machine translation systems as well as other forms of document analysis, recognition, summarization, and categorization take on practical urgency; see http://www.aaas.org/spp/rd/guiintl.htm. According to the analysis by the American Association for the Advancement of Science based on data from the Organization for Economic Cooperation and Development in 2007, the United States still led the world in its investment in R&D with 36 percent of projected world R&D performance.

[12] The Oyez project is a multimedia archive, combining audio, images, and text, devoted to the Supreme Court of the United States and its work has demonstrated the potential of such integration of sources. It is both a source for all audio recorded in the Court since the installation of a recording system in October 1955 and has been a testbed for experiments in audio capture. See http://www.oyez.org/about/.

Searching[13] across large, heterogeneous collections is obviously important. But the technologies create other opportunities for analysis and presentation. For example, visualization is one way that investigators can identify patterns and detect anomalies in large corpora as well as display results. Moreover, there is substantial evidence that the next generation will be graphical learners and communicators (Fisch 2007), implying that visualization will become increasingly important as a means of analysis as well as a mode of presentation and communication. Maureen Stone explores the topic of visualization, emphasizing the need to educate consumers as well as users of graphical media. The Web, which is an inherently graphical and interactive medium, increases the likelihood of confusion and misinformation; it requires an expanded notion of literacy, she argues. She cites a number of examples in which an image was either based on inaccurate information or was constructed in a way that conveyed confusing or inaccurate information, offering the hypothetical example of pricing information over time that fails to control for inflation (or price indexing).

There is some historical precedent for such concerns. During the 1884 presidential elections in the United States, a map of the western states and territories was published in which the proposed route of the transcontinental railroads through public lands was indicated by a thick black line. It occasioned an outcry over an apparent land grab by the railroad barons. In fact, the line had been drawn without regard for scale or for the rather convoluted terms of the grants, which had made shares in the companies that held these grants all but impossible to sell (Henry 1966).

Space and Time

Maps are a form of visualization, and visualization is closely linked to geographical information systems (GIS) and simulations. Both are intrinsic to the third theme, time and space. Geographers, one of the participants observed, have made considerable headway with space but time is still a problem. Space and time have been manifested in different ways in humanities scholarship. One obvious way is the organization of a collection of materials, reference tools, and analytical services by geography and period, like the OAAP or the Persepolis Fortification Archive Project. Space and time may encompass the detailed work of establishing provenance, authenticity, and versioning of source material, which becomes difficult and therefore interesting in the messy and heterogeneous output of mass digitization projects. Or, scholars may seek to understand the use of terms and phrases over time, as Cohen has suggested. Jonathan Bengston outlined work at the John M. Kelly Library at the St. Michael's College in the University of Toronto to coordinate an effort to digitize the works of John Henry Newman, feed the digitized output into a

[13] We are aware that the term "searching" in this context is actually a shorthand that embodies a larger array of behaviors (e.g., browsing and discovery) and technologies, including information retrieval, human computer interface design, database and repository systems.

document analysis system, and identify subtle changes in language and meaning. Longer term, he speculates, it will be even more interesting to see if relationships can be traced between the evolution of Newman's thought and the wider intellectual milieu by comparing this database of materials with larger corpora at a far more granular level than has been achieved by traditional scholarly methods (Crane and Friedlander 2008).

The notion of "space" can mean also physical or social spaces and their historical changes, where visualization and simulation can be very powerful. Archaeologists have taken advantage of the digital medium to render their information in three-dimensional modes, allowing virtual reconstructions of their sites that provide views that cannot be obtained even on the physical site itself.[14] Stephen Murray, an historian of French gothic cathedrals, uses a mix of capture and display technologies to re-create or simulate the three-dimensional spaces so that his students can also re-experience the soaring interiors at an otherwise inaccessible level of detail and to demonstrate relationships among resources that are geographically separate. He argues that this pedagogical technique removes the cathedral from its status as a fully formed and static object represented by a slide in a darkened lecture hall and allows students to understand that these were works in progress over a period of decades, embodying countless choices and decisions. For the symposium, he demonstrated a simulation that employed engineering algorithms to simulate the stresses on a Romanesque arch as it was made larger to show that the transition from the rounded Romanesque form to the pointed Gothic form was an aesthetic and a structural choice.

As these examples demonstrate, phenomena have been reinterpreted over different times and at different scales, and materials associated with an individual, group of individuals, theme, or with geographical spaces have been assembled to create collections characterized by richly marked-up text, concordances, and other reference tools. Scale, as Bengston's example demonstrates, allows this kind of focused work to become expanded. Scale also allows for conceptualizing more complex projects incorporating other types of data—in particular, scientific datasets that might allow for reconstructions and simulations of early landscapes, climate, and habitat. As one participant commented, interesting work is possible in simulating development of cities or agrarian societies, providing opportunities for multidisciplinary synthesis that is difficult to achieve without involving data on geography, weather, construction, social history, and so on. Certainly the demographic data assembled by the Minnesota Population Center or curated by the Inter-university Consortium for Political and Social Research are obvious candidates for such integrative research, as are the environmental collections

[14] A simple search of the Web using the terms "archaeology" and "simulation" yielded 554,000 hits. The 20 most highly ranked covered (1) journal articles that used simulation techniques to do site reconstructions and artifact distributions, (2) references to a textbook on use of simulation in archaeology that is in press, (3) conferences and seminars, and (4) use of site simulation software to teach archaeological methods.

managed by the University Corporation for Atmospheric Research and others. At the same time, the potential of historic travelers and explorers' accounts to add temporal depth to ecological and environmental studies is substantial but difficult to use.

Extracting the relevant information from texts, manuscripts, and drawings is a challenging technical problem, as Oard's essay suggests. Still, the nineteenth and early twentieth-century collections of specimens together with the field notes and laboratory descriptions represent a potential wealth of biological information that could enable reconstructions of historic landscapes that might inform research in literature and art history as well as ecology, environmental studies, and climate studies.[15] Layering such information onto the already complex problem of normalizing heterogeneous sources in the social sciences (Berman and Brady 2005) increases the complexity. But it remains a topic where analysis of text, language, history, and science may intersect and where GIS, visualization, simulation, and linguistic and statistical tools all have roles.

Social Networking

Bengston's example of questions that might be posed of Newman's papers calls attention to the relationships in the information as well as to discerning patterns in the use of language. Social networking, described by Bernardo Huberman, is simultaneously a technique (or set of techniques) and an object of study. This paper excited substantial discussion during the symposium, and in it he argues that the web of information represents a network of social relationships as well as a technological network. The information can be read to expose relationships that might not be otherwise evident and to illustrate how the specific technologies affect the allocation of human attention. There have been similar findings, as Huberman acknowledges, and the significance of this work lies in its scale, rigor, and level of abstraction; the algorithms can be applied in any body of work where the links can be established.

Social network analysis, one participant noted, has been successfully used in national security analyses.[16] Like GIS or visualization, these social networking algorithms represent a set of analytics that could be used to characterize text corpora, enabling researchers to identify patterns and detect anomalies more generally. For example, the scholar of Old Norse suggested that these analytics could be used to "map the social network in [Icelandic] sagas over time and then

---

[15] As a step in this direction, CLIR recently funded the cataloging of botanical collections at University and Jepson Herbaria, University of California, Berkeley, as part of the Hidden Collections program. For more information on this program, see http://www.clir.org/hiddencollections/index.html.

[16] The Visualizing Patterns in Databases of Cultural Images and Video project proposes to identify such patterns in heterogeneous data. Led by Lev Manovich, director of the Software Studies Initiative at the University of California, San Diego, the project was among those recently funded under the NEH High Performance Computing Program; see Cultural Analytics, http://lab.softwarestudies.com/2008/09/cultural-analytics.html, and Humanities and High Performance Computers Connect at NERSC, December 22, 2008; http://newscenter.lbl.gov/feature-stories/2008/12/22/humanitiesnersc/.

perhaps integrate with GIS and use this to try to draw actual historic and geographic interpretations." Equally importantly, Huberman's essay calls attention to the importance of studying the Web as an object. It ceases to be a neutral technology but instead affects the outcomes by amplifying and instantiating certain behaviors. In short, the Web is the new "text" for humanities scholarship.

What Comes Next?

Infrastructure is both social and engineered and is built from both the bottom up and the top down. It has historically been successful when local needs align with regional and national goals and when local activities take place within a sometimes loosely organized, yet coherent framework. The current landscape in digital scholarship is replete with examples of bottom-up enterprise; the open question is whether and how to stimulate large-scale coherence without stymieing individual enterprise, frustrating existing self-organization, or threatening the individualism that traditionally characterizes humanities research. The infrastructure itself is so costly and the potential gains from collaborative research are so appealing that some form of loose coordination seems appropriate.

We believe that research should drive the large-scale coherence to enable scholars of diverse backgrounds and interests to devise rich new projects and work creatively across disciplines, including computer science, while avoiding the continued proliferation of stovepipes. One participant observed, "We need to think holistically about the integration of all of these services and tools in terms of the user experience—we don't want to create multiple fragmented environments." Many participants called for various kinds of demonstration projects that would, as one scholar noted, show "people that computational tools will help them." Such projects, she continued, let "people explore new methodologies" and see how results can be transferred from one project to another. The four themes or topics that have been proposed as an initial umbrella—scale, language and communication, space and time, and social networking—tap into well-established communities of researchers. Projects conceived in this framework are likely to be robust enough to accommodate both team-based and single-investigator approaches as well as avoid the pitfall Paepcke has called "agenda mismatch," where the results of the collaboration are sufficient for the computer science student but sadly wanting for the humanities researcher.

In addition to agreeing on the importance of research as a long-term driver and the importance of demonstration projects, symposium participants offered some concrete next steps. Several proposed formulating ontologies as one avenue for future collaborative research. The term "ontologies" as used by computer and information scientists can be confusing to some humanities scholars who may have first encountered the word in an introductory course on the history of philosophy where it meant studying the nature of reality. A computational ontology is a hierarchical organization of

a domain of knowledge that a machine can process with the most general categories at the top and the most specific categories at the bottom. In a forthcoming article for the journal *Synthese* (anticipated late 2009), Cameron Buckner, Mathias Niepert, and Colin Allen offer the example, "*Wine* ➔ *Red Wine* ➔ *Beaujolais*;" everything that "is a" instance of *Beaujolas* "is a" instance of *Red Wine*, and everything that "is a" instance of *Red Wine* "is a" instance of *Wine."*[17] Although some work has been done, no large teams have formed, despite the fact that there is substantial interdisciplinary potential in such collaborations between domain specialists and computer scientists. Ontologies can be used to capture the formalization of basic concepts and can then inform more sophisticated tools and systems that are directly relevant to coping with both scale and language.

Another practical recommendation, echoed by several participants, was to create test sets that can afford investigators opportunities to experiment and learn. The most ambitious version of this idea consisted of putting existing large text corpora on powerful computer systems where researchers could explore some of the possibilities. On the basis of that experimentation, innovative questions that several people called for might emerge, thus addressing the intellectual problems inherent in asking the "right questions." At the same time, the shared resource becomes central to the structure of a discipline or set of disciplines whose research depends on it. One participant asked rhetorically, "What is the Protein Data Bank for the humanities?" And by extension, where is the motivation to support long-term preservation of these resources?

One answer to her question is: all the libraries, archives, museums, and collections of the world. So in a sense, there is no analogy, digital or otherwise, in humanities scholarship to the role of some of the key scientific datasets. But there are shared, enduring values and protocols about methods and evidence, about what constitutes an acceptable argument, and about the importance of the integrity of the source material and the research on which it is based, thus putting primacy on the importance of continuing to build sustainable and reliable collections. The challenges associated with technology-intensive management of digital collections over time are substantial, but the goals of these collections are clear: They must allow digital collections to be explored, expanded, and repurposed as the research questions evolve, and users must trust the data repositories both to safeguard their contents and to serve up reliable and trustworthy data sets upon request. Building and managing digital collections remains a fundamental condition for any research agenda.

---

[17] Colin Allen, personal communication by e-mail, January 16, 2009. Professor Allen graciously explained the concept of ontologies and provided additional background from the cited forthcoming article in the journal *Synthese*, anticipated late 2009. A version of the article, jointly authored by Buckner, Niepert, and Allen, can be found at http://inpho.cogs.indiana.edu/Papers/TaxonomizingIdeas.pdf.

References

American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. 2007. *Our Cultural Commonwealth*. New York: ACLS.

Berman, Francine, and Henry Brady. May 12, 2005. *Final Report: NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Science,* p. 9, 18-21. Available at www.sdsc.edu/sbe/.

Crane, Gregory, and Amy Friedlander. 2008. Many More than a Million: Building the Digital Environment for the Age of Abundance. Report of a One-day Seminar on Promoting Digital Scholarship, November 28, 2007. Council on Library and Information Resources. Available at http://www.clir.org/activities/digitalscholar/Nov28final.pdf.

Duguid, Paul. 2007. Inheritance and Loss? A Brief Survey of Google Books. *First Monday* 12(8). Available at http://firstmonday.org/issues/issue12_8/duguid/index.html.

Fisch, Karl. 2007. The Fischbowl: Did you know? Shift happens, 2.0, June 22, 2007. Available at http://thefischbowl.blogspot.com/2007/06/did-you-know-20.html

Gantz, John F., Chrisopher Chute, Alex Manfrediz, Stephen Minton, David Reinsel, Wolfgang Schlichting, and Anna Toncheva. March 2008. The Diverse and Exploding Digital Universe: An Update Forecast of Worldwide Information Growth Through 2011, pp. 2, 3. Framingham, Mass.: IDC.

Hackett, Edward J., ed. 2005. Special Issue: Scientific Collaboration, *Social Studies of Science* 35(5).

Henry, Robert S. 1966. The Railroad Land Grant Legend in American History Texts. In Carl N. Degler, ed., *Pivotal Interpretations of American History*, Vol. II, pp. 36-66. New York: Harper & Row Publishers.

National Science Foundation Cyberinfrastructure Council. March 2007. *Cyberinfrastructure Vision for 21st Century Discovery*, 6.

Nguyen, Lilly, and Katie Shilton. 2008. Tools for Humanists. Appendix F in Diane M. Zorich, *A Survey of Digital Humanities Centers in the United States*. Washington, DC: Council on Library and Information Resources, 58-78. Available at http://www.clir.org/pubs/abstract/pub143abst.html.

Nichols, Stephen G. 2008. "Born Medieval": MSS. in the Digital Scriptorium. *Journal of Electronic Publishing* 11(1). Available at http://dx.doi.org/10.3998/3336451.0011.104.

Olson, Judith S., Garry M. Olson, and Erik C. Hofer. n.d. What makes for success in science and engineering collaboratories. Available at http://www-unix.mcs.anl.gov/fl/flevents/wace/wace2005/talks/olson-wace2005.pdf.

Paepcke, Andreas. 2008. An Often Ignored Collaboration Pitfall: Time Phase Agenda Mismatch. Blog posting to Stanford iLab November 8.

Peterson, Michael, Gary Zasman, Peter Mojica, Jeff Porter. 2007. 100 Year Archive Requirements Survey, pp. 7-8. Storage Network Industry Association. Available at http://www.snia.org/forums/dmf/programs/ltacsi/forums/dmf/programs/ltacsi/100_year/100YrATF_Archive-Requirements-Survey_20070619.pdf.

Purdue University. 2003. Massive Data: Management, Analysis, Visualization, and Security: A School of Science Focus Area, May 15, 2003. Available at http://www.science.purdue.edu/about_us/strategic_plan/COALESCEAreas/MassiveData03may.pdf. This is a chapter in the 2003-2008 Strategic Plan (http://www.science.purdue.edu/about_us/strategic_plan/) developed by the College of Science, Purdue University.

Siemens, Ray, and Susan Schreibman, eds. 2008. *A Companion to Digital Literary Studies*. Oxford: Blackwell. Available at http://www.digitalhumanities.org/companionDLS/.

Web sites

centerNet: An International Network of Digital Humanities Center; http://www.digitalhumanities.org/centernet/.

The Minnesota Population Center, 2003-2008; http://www.pop.umn.edu/.

National Digital Information Infrastructure Preservation Program. Digital Preservation, Library of Congress; http://www.digitalpreservation.gov.

Persepolis Fortification Archive; http://oi.uchicago.edu/research/projects/pfa/.

Project Bamboo; http://projectbamboo.uchicago.edu/.

Tools for Data-Driven Scholarship; http://mith.umd.edu/tools/.

University Corporation for Atmospheric Research—UCAR & NCAR; http://www.ucar.edu/.