International Journal on Digital Libraries manuscript No.
(will be inserted by the editor)

Gregory Crane · Alison Babeu · David Bamman

eScience and the Humanities

Preprint of paper published in the International Journal on Digital Libraries, 7 (1-2), October 2007, available at http://www.springerlink.com/content/x1v71k512433027m/.

Abstract Humanists face problems that are comparable to their colleagues in the sciences. Like scientists, humanists have electronic sources and datasets that are too large for traditional labor intensive analysis. They also need to work with materials that presuppose more background knowledge than any one researcher can master: no one can, for example, know all the languages needed for subjects that cross multiple disciplines. Unlike their colleagues in the sciences, however, humanists have relatively few resources with which to develop this new infrastructure. They must therefore systematically cultivate alliances with better funded disciplines, learning how to build on emerging infrastructure from other disciplines and, where possible, contributing to the design of a cyberinfrastructure that serves all of academia, including the humanities.

Keywords Cyberinfrastructure · eScience

> To the great Variety of Readers. From the most able, to him that can but spell: There you are number'd. We had rather you were weighd. Especially, when the fate of all Bookes depends upon your capacities : and not of your heads alone, but of your purses. Well ! It is now publique, & you wil stand for your priviledges wee know : to read, and censure. Do so, but buy it first. –*Epistle To The Great Variety of Readers from the First Folio, 1623*

Gregory Crane
The Perseus Project
Tufts University, Medford, MA, USA
E-mail: gregory.crane@tufts.edu

Alison Babeu
The Perseus Project
Tufts University, Medford, MA, USA
E-mail: alison.jones@tufts.edu

David Bamman
The Perseus Project
Tufts University, Medford, MA, USA
E-mail: david.bamman@tufts.edu

Scientists have already begun building a new "cyberinfrastructure" to manage data-driven science. Petabytes of data stream from proliferating networks of increasingly perceptive sensors. These sensors track the depths of the oceans, the furthest reaches of space, and even the earliest moments of creation. Watchful machines note the formation of galaxies and the flight of birds alike, recording in every second far more than any human observer could see in a lifetime. No one research team or even nation can collect and assemble all the pieces in these many, ultimately interrelated scientific puzzles. However weary we may be with neologisms such as cyberinfrastructure, portentously simple labels such as "the grid," or the ungrammatical prefixes in e-commerce and eScience, we need radically new technology and social conventions if we are to build on the galaxies of data now taking shape [12,10]. The papers in this issue of the *International Journal on Digital Libraries* bring home rapidly growing needs, unevenly changing practices and recently emergent approaches from the scientific community. Every scientist may not feel the same pressures, but human civilization probably depends upon the ability of our colleagues in environmental sciences to process these streams of data with no precedent in the history of human intellectual life. The needs are very real and the stakes could not be bigger.

Cyberinfrastructure addresses at least two complementary needs. The first and most obvious is scale: we need a higher order of infrastructure if we are to manage and analyze the staggering bodies of data that we are now collecting. We need to combine the decentralized services of desktop computing with seamless access to high performance computing applied to distributed collections [20]. Humanists have, of course, long had more data than they can analyze by hand – no one has been able to read and analyze all scholarship about Shakespeare, for example, in decades, if not generations. Datasets from archaeology, linguistics, and other sources have begun to present problems and opportunities similar to those faced by our colleagues in the sciences [11,19].

A second function of cyberinfrastructure is even more important for the humanities. Scientific problems have become so complex that no one, however highly trained, can master all the disciplinary knowledge needed to understand the problem as a whole: there is always more math or another domain of specialized knowledge. We need systems that can make specialized content intellectually as well as physically accessible.

In the humanities, language provides the foundational challenge of intellectual access. "It's all Greek to me," says Shakespeare's Caesar after overhearing Cicero speaking in the international language of the time: we can have access to documents in Greek, Latin, classical Arabic, Chinese, Sanskrit, Old Norse, Syriac, Akkadian, Sumerian, Middle High German, and any other human language, but that physical access has traditionally meant little unless we had extended training in those languages [14]. Language is, however, one example of the general problem of intellectual access: when a user confronts an object, a mature cyberinfrastructure analyzes the user's background and immediate purposes (e.g., browsing vs. deep analysis, focus on one topic vs. another), the content of the object and its intellectual presuppositions, and then provides intellectual scaffolding with which that user can make the best possible use of that object. This can include not only full machine translation but customized translation and reading support that points out idioms or expressions that the reader has never encountered before or highlights particular topics. The state of the art in machine translation, cross language information retrieval, recommender systems and other well-published areas would, if implemented for cultural heritage materials, radically lower, if not remove, the oldest and most intransigent barrier in the humanities [15,6].

Language provides only one example of a more general problem. We are always moving through spaces, physical and digital alike, teeming with meaning and significance that print tools cannot deliver to us in the form and at the point that we need. Consider, for example, a three mile path from Somerville and Medford – two largely working class, often grey cities outside of Boston, neither a tourist destination.

The three mile walk begins at Powderhouse Square in Somerville, where a British force marched out from Boston to confiscate munitions which American colonists had stored in an old mill that still stands on that site. The walk crosses the location of the first bridge across the Mystic river, built in 1638 just a few years after Boston was founded, on the traditional land route north from Boston, the Middlesex Canal, the first railroad in Massachusetts (which made the canal obsolete), a merchant family's mansion and the still-standing outbuildings built for his slaves in the eighteenth century, a bell cast for the Bell and Everett ticket in the 1860 presidential campaign that Lincoln won and that led to the American Civil War, the home of the American writer and anti-slavery activist Lydia Maria Child, the site where George Luther Stearns maintained in his mansion a stop on the underground railroad and from which he sent funds as one of the "secret six" to support John Brown, nineteenth century shipyards on the Mystic, the old red light district of the puritan city, etc. Rows of one, two and three family houses, tightly placed, sprawl over the farmlands that had covered this area until the railroad restructured space in the latter nineteenth century. Many might long for the fields which these houses replace but these were often the first homes to which their owners, a working middle class from the tenements of Boston, or any of their ancestors had ever laid claim since property rights had been established.

Few, if any, of those who have passed through these streets for years and decades have much understanding of what they pass. A few historical markers with a few dozen words and a picture or two are a beginning – but also an end, for few have found their way to the specialized collections that would carry them further.

The cell phone in my pocket, however, contains a GPS and can display my location on a map. Such a device, primitive as it may soon seem, can connect the library to the world. Spatial queries can scour vast collections for information relevant to the places through which we pass [13]. Information extraction and named entity identification distinguish one Elm Street from another, matching addresses on picture captions to call up historical views of the street before us, while articles from historical newspapers, clustered and analyzed, identify famous events that happened beneath our feet and for which no historical marker remains [1].

Humanists have already made the case that they need cyberinfrastructure as well. The Commission on "Cyberinfrastructure for the Humanities and Social Sciences," sponsored by the Mellon Foundation and the American Council of Learned Societies and inspired by the National Science Foundation Cyberinfastructure report, recently held hearings around the United States. On July 18, 2006, after two and a half years of work, the commission produced a report arguing that cyberinfrastructure was a necessary component to advance both scholarly communication and the humanities in general. Where print infrastructure emerged over centuries, "Cyberinfrastructure is being built much more quickly, and so it is especially important that humanists and social scientists actively engage with it, articulate what they require of it, and contribute their expertise to its development." [1]

Humanists still, however, lag behind their scientific colleagues. No humanists are publishing articles in this issue of *IJDL*. The social sciences appear only insofar as they contribute to research about scientific practice (see the articles by Zimmerman and Borgman et al. on scientists' use of data by others, both in the context of ecology, and Barkstrom on the interaction between the producers and users of scientific data). Humanists may believe that they need a cyberinfrastructure too,

[1] http://www.acls.org/cyberinfrastructure/

but we lag far behind our colleagues in the sciences in understanding what intellectual activity an emergent cyberinfrastructure might support [5]. The term eScholarship is too narrow, at least insofar as it implies the production of finished scholarship. From a practical perspective, traditional humanities scholarship alone does not have at its disposal the resources needed to build an advanced infrastructure. In the United States, the situation is particularly striking. The National Endowment for the Humanities (NEH) has sought level funding ($141 million) for its 2007 budget request.[2] The National Science Foundation (NSF), by contrast, boasts a yearly budget of $5.5 billion[3] – 39 times larger than that of the NEH. Even when we consider traditional supporters of the humanities such as the Mellon Foundation, the amount available to humanists remains small. Perhaps more ominously, the relative federal budgets reflect not only current spending power but also, more significantly, a settled societal judgment as to the relative importance of humanities scholarship and scientific inquiry. Within Europe, the situation is somewhat better: the UK's Arts and Humanities Research Council (AHRC), for example, "provides approximately £90 million to support research and postgraduate study in the arts and humanities, from archaeology and English literature to design and dance."[4] With a population roughly one fifth that of the United States, the UK AHRC invests more than the US NEH. National funding sources from other European nations (especially Germany) and from the European Union as a whole provide substantial investments, but the scale of humanities investment as a whole remains far below that in the sciences.

The lack of resources are an obvious challenge but also point the way to two larger movements that can help the humanities redefine themselves, not only acquiring new material capabilities but also rethinking, from the ground up, the questions which they pursue and the relationship which those questions bear to the world as a whole.

First, humanists need to collaborate with their better funded colleagues, building wherever possible on work funded for other areas.[2] This includes not only the NSF but other federal agencies such as the National Institutes of Health ($28 billion/year[5]) and the Defense Advanced Research Projects Agency (DARPA) ($3 billion)[6] with which humanists have few traditional ties. A century of settled practice has tended to separate scholars from librarians, but both groups need to work with each other if they are to flourish [16]. On the one hand, the Institute for Museum and Library Services (IMLS) ($247 million in fiscal year 2006[7]) and NEH have already agreed to a formal partnership.[8] The largest and most stable source of support for the infrastructure of humanistic inquiry resides with libraries – the 123 members of the Association of Research Libraries invest "more than one billion dollars every year on library materials."[9] More recently, internet giants such as Google[10] and Microsoft[11] have begun investing in content and infrastructure to support vast digital libraries which will, if fully realized, be not only more accessible but more comprehensive than any academic libraries in human history. Entrepreneurial enterprises such as the Open Content Alliance provide a common infrastructure in which communities can create purposeful, open source libraries of content.[12]

Collaboration needs to be international as well as interdisciplinary. National funders in the United States place particular emphasis on American cultural heritage: the Library of Congress is now building a World Digital Library but only after years of work on its American Memory project;[13] the NEH received a major addition to its budget to support an initiative, "We the People," focused primarily on American subjects; the IMLS has primarily (though not exclusively) funded digitization and leadership projects that make American content more accessible. The ACLS/Cyberinfrastructure report was written in, and by scholars who worked primarily with, English. We need sustained collaboration among institutions of higher education and funders from multiple countries [18]. But we also need to recognize that, in the humanities, language – and especially the variety of languages – is at the core of human cultural heritage. While experts in American history and English literature can live in a world of English, that monolingual intellectual space is atypical and cannot serve as the model for any serious humanities cyberinfrastructure. Language is not an afterthought. Language must be the starting point and grand challenge as we consider the services that we need to understand and to disseminate our understanding of the shared culture of humanity.

In order to collaborate, humanists will, however, themselves need to do a better job identifying the technological interests which they share with their colleagues in the sciences, engineering, medicine, defense, libraries and industry. Scholarly primitives[14] and digital tools[15] envisioned by humanists overlap substantially with the needs that scientists (such as the authors in this special issue) have begun to identify. Humanists need to identify what

[2] http://www.neh.gov/news/archive/20060206.html
[3] http://www.nsf.gov/about/
[4] http://www.ahrb.ac.uk/news/news_pr/2006/pre-budget_report.asp
[5] http://www.nih.gov/about/
[6] http://www.darpa.mil/body/pdf/FY07_Final.pdf
[7] http://www.imls.gov/about/apprfy2006.shtm
[8] http://www.neh.gov/news/archive/20060928.html
[9] http://www.arl.org/arl/arlfacts.html
[10] http://books.google.com/intl/en/googlebooks/about.html
[11] http://publisher.live.com/
[12] http://www.opencontentalliance.org/
[13] http://www.worlddigitallibrary.org/project/english/index.html
[14] http://www.iath.virginia.edu/~jmu2m/Kings.5-00/primitives.html
[15] http://www.iath.virginia.edu/dtsummit/SummitText.pdf

emerging technologies they can use, what better funded groups can develop in their own self-interest, and what humanists themselves must develop, document and support over time.

Second, humanists need to rethink not only their relationship with better funded groups but also ways in which their output can better support intellectual life in society as a whole [8]. From a pragmatic perspective, we need to increase the perceived value of our work if we are to increase our resources, but in rethinking our audience we enhance expert and general user alike. We need to increase physical and intellectual access to every type of content and we need methods that are automated and can be applied to large bodies of content.

We cannot predict the full form that cyberinfrastructure will take over the coming years nor can we can anticipate every service that we will need. On the other hand, we should not use this uncertainty as a justification for hesitation. We may not know every service that we will need, but we can already identify services that are already in use and that should shift from isolated, project based applications to ubiquitous and often invisible elements of infrastructure. In this issue, the articles by Warner et al., Tsoi et al., Candela et al., Gahegan, and Witt and Brandt all address this need, from interoperability between services to extending them to better support eScience endeavors.

The best way to discover the services that we cannot predict is to build the services that we already need and then observe what new demands appear with real use. We need an extensible workflow that can provide at least the following services for all textual content in as many languages and formats as possible. The core services here echo the three core services identified in the DARPA GALE project (analogue to digital, one language to another, text to data) and they thus reflect, in our view, a convergence of interest that extends beyond the humanities.

1. Conversion of page images to text: In the simplest case, we are simply converting from one format to another: we have images of printed text and want to capture a transcription of that text in digital form. Extracting useful content from scans of printed historical sources for American history is relatively easy, but the general problem remains a non-trivial task. Modern business publications have highly evolved conventions, such as regularized spelling, clear print, and recurrent page layouts, that developed to support high volume processing long before digital technology. Even when we have cleanly printed cultural heritage materials, the fonts may differ from those of modern business exchange (e.g., classical Greek which resembles, but has very different accents from, modern Greek vs. fonts with no modern analogues such as Syriac). Even where a Roman font is used, the language models may be radically different (e.g., English OCR often converts Latin words such as t-u-m, 'then,' into English t-u-r-n). The grand challenge here, however, lies in processing the wealth of handwritten materials that have never been typeset: this includes millions of documents, on paper, papyrus, stone, clay tablets and other media, and in languages that include not only Latin and Greek but every cultural heritage language [4,7].
   In effect, OCR is a special case of the sensor networks described elsewhere in this issue. Multiple optical character recognition systems, automatic speech recognition, image classification, global positioning systems etc. scan the audio, visual and text information embedded in both static and time based data files producing transcriptions of spoken language (e.g., Russian lyrics), classification and/or recognition of objects (a church vs. a particular church), analyses of written language (e.g., burial inscriptions in video or carefully captured page images), records of where new data is collected or alerts about locations of interest (e.g., the path of the now invisible Middlesex Canal).
2. Raw text to structured data: We want to identify and provide background about all people, places, organizations, linguistic constructions, idioms, and any other well defined entities: if we are standing on Elm Street or viewing a statue of Charles Sumner, we want a system that can match references to our Elm Street or the Charles Sumner at whose statue we are gazing. Tasks include semantic classification (is Washington a person or a place?) and morphological analysis (especially important in highly inflected languages where a single dictionary form may have thousands of different spellings), as well as additional analyses such as identification (if Washington is a place, which Washington is it?), syntactic analysis (if a word is in the accusative case, does it depend upon a preposition, a verb or some other construction?) and propositional analysis (e.g., converting "Smith in Washington" to a proposition about the location of a particular Smith in a particular Washington).
   Humanists will need to provide the language and cultural specific services (e.g., morphological analysis of classical Greek) and knowledge sources (e.g., historical gazetteers documenting places as they existed in different periods of time, databases of morphological and syntactic analyses). For further consideration of these challenges in terms of eScience, see the article in this issue by Borgman et. al., which explores how ecology data has been traditionally shared and structured in the field and how technologies such as embedded sensor networks will allow aggregation and evaluation of raw data on a much larger scale. This issue is also addressed by Hunter and Cheung, who note the importance of data provenance for scientists seeking to share their datasets.
3. Multi-lingual support: This includes machine translation (generate a new translation), cross language

information retrieval (e.g., pose a query in English to search Russian), translation identification (apply cross language information retrieval to locate a pre-existing translation already online), translation assistance (provide links from individual words and phrases to dictionary entries and annotations and from syntactic constructions).
Humanists may be able to rely on basic services for modern languages from commercial providers (e.g., Google's rapidly evolving machine translation). Aside from the morphological and syntactic analyzers mentioned above, humanists must provide the machine readable dictionaries and parallel texts (e.g., matching Greek source texts and English translations) for cultural heritage languages.

4. Customization and Personalization: Customization includes user-driven choice (e.g., emphasize transportation technologies such as canals and railroads vs. anti-slavery activists) while personalization describes system initiated adaptations (e.g., visitors who ask about Lydia Maria Child may also be interested in George Luther Stearns). Customization and personalization might thus identify background that a particular user needs for the Russian music video about Saint Petersburg (e.g., emphasizing music for one and historical context for another).
Many humanists are familiar with the recommender system built into Amazon.com (customers who bought the book you have just ordered also ordered the following additional books). Much can thus be done automatically, but humanists will also need to create their own domain specific modules: language students may want to track the vocabulary with which they are familiar and receive prioritized lists of new vocabulary (or technical terms) in an unseen passage. Customization is represented most closely by two articles in the current issue: Hunter and Cheung, who present work that enables scientists to construct scientific publication packages, and Candela et al., who detail the DILIGENT architecture that scientific communities can use to create custom digital libraries to support specific research needs.
5. Continuous user contributions: We need to be able to collect contributions large and small, produced not only by individuals and by collaborative groups but also smoothly aggregated from many small contributions into a useful whole. This includes not only digital analogues to traditional publications but small corrections and modifications to the many automated systems that underly emergent electronic systems. In the humanities these include simple corrections of OCR errors [17] and sophisticated analyses of syntactic structure [3]; examples from the world of digital libraries can be found in the current issue in the articles by Collins et al. (who present work on collaborative eScience libraries) and by Barros et al. (who describe work on allowing field ecologists to contribute to an eScience repository).

Throughout the twentieth century humanists enjoyed a stable print-based infrastructure that took recognizable shape in the nineteenth century and relied upon efficient printing, mature page layouts, effective citation schemes, and well-organized libraries. Our instrumentation has changed relatively little and humanists seem unprepared, for the most part, to grapple with the deeper implications of emerging digital environments. The first generation of digital applications are, for the most part, classic incunabula: productions that replicate in new media the limits of the old [9]. We continue in the early twenty-first century to produce documents and to pursue research agendas that, in form if not in content, are not so different from those of the late nineteenth.

After a century of relatively stable instrumentation and information infrastructure, humanists must develop new organizational structures to develop and maintain the services on which we increasingly depend. At present, humanities projects are too small in scale – we have many cottage industries solving the same problems in ways narrowly optimized for particular projects. While we cannot predict the organizational structures that will evolve, we can begin to see ecological niches emerging. We need disciplinary centers: classicists, for example, have their own specialized needs that involve the languages on which they focus. At the same time, we cannot have a flat organization, with each discipline managing its own infrastructure. A relatively large humanities discipline such as classics might be able to support its own unique systems, but that would only condemn us to an underfunded infrastructure that we could not sustain over time. Greek and Latin are particular examples of historical languages and can share much of the same architecture that would support Akkadian, classical Arabic, and other historical languages. While disciplinary centers can contribute their own expertise to their individual services, we must reach out to our colleagues both within the humanities and in the sciences in order to define the shared cyberinfrastructure that we all need. As the articles in this special issue demonstrate, our problems are largely the same.

References

1. Robert Allen, Andrea Japzon, Palakorn Achananuparp, and Ki J. Lee. A framework for text processing and supporting access to collections of digitized historical newspapers. in *HCI International 2007* Volume 4558 of *Lecture Notes in Computer Science* 2007.
2. Shlomo Argamon and Mark Olsen. Toward meaningful computing. *Commun. ACM*, 49(4):33–35, April 2006.
3. David Bamman and Gregory Crane. The Latin Dependency Treebank in a cultural heritage digital library. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 33–40, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

4. David Bearman and Jennifer Trant. Converting scanned images of the print history of the world to knowledge: A reference model and research strategy. *Russian Digital Libraries Journal*, 8(5), 2005.
5. Tobias Blanke, Stuart Dunn, and Alistair Dunning. Digital libraries in the arts and humanities– current practices and future possibilities. In *INSCIT 2006*, 2006.
6. Jiangping Chen, Yuhua Li, and Gang Li. The use of intelligent information access technologies in digital libraries. In *Web Information Systems WISE 2006 Workshops*, pages 239–250, 2006.
7. Sayeed G. Choudhury, Tim Dilauro, Robert Ferguson, Michael Droettboom, and Ichiro Fuginaga. Document recognition for a million books. *D-Lib Magazine*, 12(3), 2006.
8. Paul Courant. Scholarship and academic libraries (and their kin) in the world of Google. *First Monday*, 11(8), August 2006.
9. Gregory Crane, David Bamman, Lisa Cerrato, Alison Jones, David Mimno, Adrian Packel, David Sculley, and Gabriel Weaver. Beyond digital incunabula: Modeling the next generation of digital libraries. In *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006)*, volume 4172 of *Lecture Notes in Computer Science*. Springer, 2006.
10. Cathy N. Davidson. Data mining, collaboration, and institutional infrastructure for transforming research an teaching in the human sciences and beyond. *CTWatch Quarterly*, 3(2), May 2007.
11. A. Dimitriadis, M. Kemps-Snijders, P. Wittenburg, M. Everaert, and S. Levinson. Towards a linguist's workbench supporting eScience methods. In *Second IEEE International Conference on e-Science and Grid Computing (e-Science'06)*, pages 131–141, December 2006.
12. Peter Gietz, Andreas Aschenbrenner, Stefan Budenbender, Fotis Jannidis, Marc W. Kuster, Christoph Ludwig, Wolfgang Pempe, Thorsten Vitt, Werner Wegstein, and Andrea Zielinski. Textgrid and eHumanities. In *E-SCIENCE '06: Proceedings of the Second IEEE International Conference on e-Science and Grid Computing*, Washington, DC, USA, 2006. IEEE Computer Society.
13. Martyn Jessop. The application of a Geographical Information System to the creation of a cultural heritage digital resource. *Literary and Linguistic Computing*, 20(1):71–90, March 2005.
14. Gareth J. F. Jones, Ying Zhang, Eamonn Newman, Fabio Fantino, and Franca Debole. Multilingual search for cultural heritage archives via combining multiple translation resources. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 81–88, Prague, Czech Republic, 2007. Association for Computational Linguistics.
15. Marijn Koolen, Frans Adriaans, Jaap Kamps, and Maarten de Rijke. A cross-language approach to historic document retrieval. in *Proceedings of the 28th European Conference on IR Research (ECIR 2006)* Volume 3936 of *Lecture Notes in Computer Science*, pages 407–419. 2006.
16. Shawn Martin. Digital scholarship and cyberinfrastructure in the humanities: Lessons from the Text Creation Partnership. *Journal of Electronic Publishing*, 10(1), 2007.
17. Gregory B. Newby and Charles Franks. Distributed proofreading. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 361–363, Washington, DC, USA, 2003. IEEE Computer Society.
18. Roy Rosenzweig. Collaboration and the Cyberinfrastructure: Academic collaboration with museums and libraries in the digital era. *First Monday*, 12(7), 2007.
19. Dean R. Snow, Mark Gahegan, Lee C. Giles, Kenneth G. Hirth, George R. Milner, Prasenjit Mitra, and James Z. Wang. Cybertools and archaeology. *Science*, 311(5763):958–959, February 2006.
20. Paul Watry, Ray Larson, and Robert Sanderson. Knowledge generation from digital libraries and persistent archives. volume 4172 of *Lecture Notes in Computer Science*, pages 504–507. Springer, 2006.