Digits and Dreams: Scholarship and computing in the age of abundance

Amy Friedlander
23 September 2008; revised 2 December 2008
Royal Irish Academy
Dublin, Ireland

Thank you for inviting me to join you this afternoon to talk about the future of scholarship. This is my first trip to Ireland, and I am looking forward to walking around Dublin this week. Walking is my favorite way to explore a new place and takes me back to my career in archaeology, historic preservation and land use planning as well to prior trips to visit new places.

About twenty years ago, my husband and I took a road trip through rural England and at one point, found ourselves in Exeter. As we wandered along the High Street, I paused to look over the fence enclosing a construction site, and there was a young man working, about hip deep in an archaeological unit. I noticed artifacts in the dirt piled up beside him, what we call "backfill". In North America, we generally screen the backfill to recover any possible scrap of information, particularly for sites prior to, say, 1750, so naturally I was curious. "Excuse me," I said, "Why aren't you screening the backfill?" The archaeologist looked up, surprised, and replied, well, the practice was not to screen the backfill except in certain situations. So why was this context in Exeter not accorded more treatment? He answered calmly – and quite reasonably – that his team knew that they were digging through the remains of an adjacent fourteenth-century abbey. Since they knew more or less what had been there – there were surviving foundations as well as documentation of the entire medieval complex -- the archaeological remains were not of great interest. Instead, he said, "We're going for Rome."

So much for 1750.

The moral to my story is obvious: Significance is context dependent, and what may seem rare or unique in one setting may be an acceptable loss or construction-related overburden in another. So it is with research, the value we place on it and the judgments we make. Humans do context very well. "Natural language," the linguists' term for the languages we grow up speaking, is all about context, which is why telling jokes or learning slang in a second language is so very difficult. Human intelligence does not always do well beyond a certain scale or when we are asked to observe repetitive actions like watching dials for hours on end. Such situations are frequently where computers become useful. Computers function quite well when we ask them to process lots and lots of information (what is frequently called "scale") or detect a level of detail not otherwise visible (what is often called "granularity"). However, understanding context is something that computers do poorly, if at all.

Not surprisingly, then, one of the central challenges in building systems is optimizing human intelligence and therefore getting the context right. That is to say, what is the right system for the right problem in the right context? For researchers in the humanities, using computers can challenging not because they are necessarily afraid of technology but because the available software tools don't seem to do what researchers want. Which isn't surprising, given that many of the computational tools that we use were developed for either scientific or commercial purposes. So before we even talk about the substance of our research, whether Petrarch or political history, we confront a fundamental question: How do these tools and systems enable us to ask questions in the humanities that advance the interpretations of text, images, and other materials of traditional interest? How does the new environment modify what we understand to be "scholarship"? Note that there is an implicit tension between a

model of scholarship and scholarly results that sees the technology in essentially incremental terms, as allowing investigators to examine traditional sources and questions using new tools, and a model of scholarship that is transformative, that sees the digital environment as a way to use new sources and new methodologies to answer new questions and possibly to communicate in new ways.

Finally, what happens to traditional configurations of roles and responsibilities? At the Council on Library and Information Resources (CLIR), we are particularly interested evolution of scholarship and advanced research and how that will affect the roles of libraries and other cultural institutions. CLIR is small not-for-profit headquartered in Washington, D.C., supported by a combination of generous funding from the Andrew W. Mellon Foundation together with smaller awards from other sources including federal grants and contracts and sponsorship from over 200 academic libraries and scholarly organizations.

A Few Confessions and Caveats

Before we go much further, let me offer a little more autobiography and then a few caveats. I retired from archaeology in the early 1990s, did a series of studies about historical development of large scale, technology intensive infrastructures (railroads, electrical power, communications, and so on), and from there went into writing about information, in particular information in digital form and the associated technologies.

Broadly speaking, people who study science and technology tend to divide into two main camps: the optimists and the pessimists. Pessimists look and see cause for alarm – with some justification. The evidence for global warming, destruction of wildlife habitat, maldistribution of access to water, food, basic health care, and shelter, and, frankly, a decrease in simple civility is real and, to my mind at least, largely incontrovertible. Optimists, on the other hand, see evidence that things are, if not already better, capable of being better. And certainly, there is evidence to support that view. Twentieth-century achievements in vaccines and therapies to combat smallpox, influenza, diabetes and polio have clearly reduced human suffering and contributed to improved longevity in much of the world. A series of engineering achievements tells a similar story. When one-term Congressman Abraham Lincoln made the 800-mile trip home to Springfield, Illinois from Washington, D.C. in 1849, he needed 12 days and three modes of transportation: stagecoach, railroad and steamship. Barely 12 years later, then President-elect Lincoln made the trip in 2 days, wholly by rail.[1] Today, I can drive the distance in about 12 hours.

Not surprisingly, many technologists are optimists, the dreamers; people who build things tend to be of that mindset. On the other hand, those who look at the consequences of technology can be more pessimistic and, frankly, fearful. So the optimists look at the computationally rich environment and the stunning advances in communications, life sciences, and robotic manufacturing and see promise. They see drudgery taken away and human creativity released. Pessimists worry about loss, over-simplification, seemingly rampant plagiarism, and faltering standards. Although I can find justification for these fears and more, I will admit to cautious optimism. No, I do not expect to solve problems of global warming, world peace, or maldistribution of resources, but humans are quite resilient. Indeed, they built the very tools and resources to which we will shortly turn our attention. Yes, we can cite (unfortunately) many stories of students who buy or steal papers on the web. But we can also find senior citizens like my husband's Aunt Naomi, who at the age of 80-plus downloads recipes to try for holiday meals and e-mails her family regularly. Or my father,

[1] John F. Stover, *Iron Road to the West, American Railroads in the 1850s* (New York: Columbia University Press, 1978), pp. 1-5.

who read the Windows operating system documentation when at the age of 87 he set up his second machine.[2]

A few more confessions and caveats are in order before we proceed much further. First, I am necessarily talking from a U.S. perspective, not because I think it is the only one but because my career as an historian, archaeologist, editor and now funder has been centered in the U.S. Second, most of my work in the last 15 years has taken place in the context of science and technology and my exposure to the humanities has been quite recent and is probably limited. Consequently, my remarks will reflect broad commonalities across scholarship, science, technology and humanities.

Now the caveats: Trends in the nature of research are taking place in the context of a broad re-structuring that consists of at least three related elements:

- a re-thinking and re-organization of the system of higher education in the U.S. over the next 30 years, including support for advanced research, systems of scholarly communication and relationships with the private sector. This has several pieces: (1) the degree of federal funding versus funding in the private sector either through grants to universities and research institutes or through research in corporate laboratories; (2) the balance between long term research and near term development; (3) changes in the demographic profile of the undergraduate population; and (4) the so-called "pipeline" problem which embraces both the number of young people electing to study science and mathematics at the secondary and undergraduates levels and the willingness of new Ph.D.'s to go into or to remain in careers in research.

- a re-thinking of notions of literacy, how it is measured and what it means to be literate. In a white paper for a recent workshop on humanities and computing at CLIR, Maureen Stone argued the case for literacy in visualization because it is integral to understanding the graphical representations of information with which we are surrounded.[3] Improper representation of information arising from failure to appreciate context or carelessness can result in serious misinformation. Consider the hypothetical example Stone, a computer scientist, gives of pricing information over time that failed to control for inflation (or price indexing). Or the historical example of a map of the western United States, promulgated in 1884 during a presidential election, in which the proposed route of the transcontinental railroads through public lands was indicated by a thick black line 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.[4] Even our old friend text can take on different meanings. A recent report by the Pew Internet and American Life project found that U.S. teenagers differentiate between formal writing and the argot they use in text messaging and e-mails. They do not think of composing electronic

[2] As of May 2008, 35 percent of U.S. citizens, age 65 and over, had Internet access, substantially less than the 73 percent of the total adult population of the U.S. which reported Internet access; see Demographics of Internet Users, updated July 22, 2008; http://www.pewinternet.org/trends/User_Demo_7.22.08.htm. Interestingly, interest groups associated with the elderly (AARP, for example) encourage use of technology to offset other barrier, mindful, however, of the vulnerability of this constituency to scams, identity theft, and other predatory behaviors.

[3] Maureen Stone, Information Visualization: Challenge for the Humanities, [October 2008], http://www.clir.org/activities/digitalscholar2/stone11_11.pdf..

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

communications as "writing." In fact, they believe that good writing is an essential skill and say they would welcome more – but different – in-school training.[5]

- the increased exposure in the conduct of scientific research. Scientists have shared their results and interpretations for millennia. In the last 30 years, collaborations and cooperation have begun to occur earlier in the research process, driven and enabled by advances in information technology. This phenomenon is particularly obvious in large scale scientific experiments, such as the organization of international research teams at C.E.R.N. and elsewhere, and in the construction of shared databases that are fundamental to the research process. At the same time, the research process has become more visible to non-scientists and in some well-known examples more inclusive and participatory.

Within these concurrent processes lies a set of interactions and feedbacks with advances in information technology such that the information technology, in its many expressions, is both cause and effect. In this context, it is useful to think of the information technologies as creating a computationally rich environment with a set of properties and affordances that allow or encourage people to do things rather than as a set of technologies (for example, networking or word processing) or toolkits (for example, geospatial, statistical or text analysis packages) that implement capabilities but that may be changed or become obsolete as a result of innovation. As previously mentioned, many of these systems were built to support scientific research, for example to collect sensor data, to communicate data from satellite and terrestrial earth observing instruments to labs where scientists could analyze it, or to render the data in an interactive image – that is, to visualize it – in a way that was meaningful and interpretable. The challenge for humanities is to understand what these capabilities mean for their teaching and research, what questions humanists can ask, and how their kinds of questions can push the development of the technology. Recognize that technology development is not a linear process from idea to research to prototype to product. Rather, it is an iterative process of experiment, design, re-thinking and re-design in which computer scientists seek difficult problems and to which humanists can make contributions as well as reap benefits.

Humanists ask deceptively simple but very hard questions, frequently because the answers are context dependent or depend upon highly heterogeneous sources in many languages, media, and locations. Moreover, the phenomena are not inherently quantized in the way that physics is inseparable from mathematics and chemistry is wed to the periodic table of elements. Rather, research in the humanities rests on human expression, traditionally creative expression in literature and art (we'll leave out music, for the moment.), and context is always in front of us. Of course context matters in scientific research. Whereas scientists have devised agreed-upon ways of controlling for context, the relationship between the individual and the context may be precisely the subject of humanists' inquiry.

Not only is context the name of the game, but computers do context poorly. The raw material of humanities analysis– text, images, sound, and so on – can now be rendered in digital form and is thus amenable to computer processing. Humanists would like to find ways either for computers to do context better or to provide the processed information to us so that *we* can put the data in context. This is very hard, and computer scientists like hard problems, thus creating an opportunity, a market, so to speak, in hard questions. For this reason, the Council has re-engaged with an enduring theme in modern Western thought: the

[5] Amanda Lenhart, Sousan Arafeh, Aaron Smith, and Alexandra Rankin Magill, Writing, Technology and Teens (Pew Internet and American Life Project and The National Commission on Writing, April 24, 2008), http://www.pewinternet.org/pdfs/PIP_Writing_Report_FINAL3.pdf.

relationship between the two cultures of science and humanities. This intellectual problem assumes a different character in the new environment where everything is digits, processing is cheap, and data is abundant. It doesn't matter whether the corpus of text to be searched is the current scientific literature or the combined works of Jane Austen, Immanuel Kant, and Leo Tolstoy all in their native languages and all perhaps housed at different locations. The investigator still wants the answer (or answers) in a sea of otherwise impervious material.

Finally, I should also observe that making generalizations is always a perilous business. There are almost certainly many counter-examples and nuances that could be introduced to qualify some rather broad assertions about the digital environment and promises it offers. But let us go forward nonetheless

A Quick Tour of How We Got Here

We will start our journey with a quick tour of the technology. The second half of the twentieth century saw advances in networking, software, and processing power, which converged to create a computationally rich environment realized first in the universities and large corporations but soon thereafter widely distributed in small businesses and at home. The Internet component of this story tends to be frequently told. Equally if not more important are developments in microprocessor technology and software. Advances in microprocessors meant that smaller and smaller devices could do more and more things. Advances in software de-coupled the function of the task from the physical equipment so that one machine could do many things. At the same time, advances in software led to new analytical tools to run on those machines, namely, databases, spreadsheets, visualization, and word processing and probably most importantly the ability to integrate across multiple packages and applications. Finally, there has been an immense increase in storage capacity and memory, both the chips on which data is stored and the software to manage and retrieve the data. For research, one of the major achievements was the ability to do simulations, which require immense processing power and large data sets, frequently larger than can be stored on a single machine. The network became important initially as a means of piping data around to support complex data hungry operations and then as a way to provide access to sophisticated analytical power from remote locations.

The payoff has been substantial. Scientists simulate the behavior of tectonic plates or climate data under various conditions without waiting for a disaster, thus enabling development of precise prediction and hence warning systems. On the humanities side, we saw and are seeing extremely interesting visualizations and interpretations of archaeological sites. In one of the earliest, Bernard Frischer and his colleagues at the Cultural Virtual Reality Laboratory at the University of California, Los Angeles, built a model of the Roman Forum, based on available archaeological data, to simulate the classical city center in ways that let "visitors" see it at various scales and different perspectives, including from inside of individual buildings, so that we can re-experience the play of light and space.[6] Stephen Murray, an historian of French gothic cathedrals, uses a mix of capture and display technologies to re-create 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 pedagogical technique removes the cathedral from its status as a fully formed and static object

[6] Bernard Frishcer, Diane Favro, Dean Abernathy, and Monica De Simone, The Digital Roman Forum, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXIV-5/W10 [2003], http://www.iath.virginia.edu/~bf3e/revision/pdf/FrischerEtAlRomanForum.pdf.

represented by a slide in a darkened lecture hall and allows students to understand these were works in progress over a period of decades, embodying countless choices and decisions.[7]

Speaking as a sometime historian, the societal adoption of computers as consumer devices together with Internet connectivity is the interesting phenomenon. While there is substantial debate about the causes of this broad and rapid diffusion, first of computers and then of the network, one very simple economic fact stands out: Computers and applications got cheaper as well as more powerful at a time when the cost of energy was still low and the market was large enough to support an initial launch. One well-known example is the spreadsheet,[8] which is generally considered responsible for take-off in the lucrative personal computer business software market in the 1980s. The critical attribute of the spreadsheet was not that we didn't have spreadsheets in paper or on mainframe computers – we did – but that the interactive, easily changed implementation of a spreadsheet that ran on Apples and Windows machines was far more useful, less time-consuming to work with, and could be used to budget, monitor performance and forecast. A spreadsheet can also be used to maintain lists, which isn't even a financial operation. So it possessed important characteristics: It was familiar; it let people do things they wanted to do better than they were able to do them already; it was context appropriate; and it was multi-functional, capable of adaptation by users as they became and still become more comfortable with it.

The advent of the consumer version of the Internet ten years later converged with access to relatively inexpensive equipment and relevant software packages, so that diffusion of Internet connectivity rested on fairly widespread diffusion of personal computers as well as the existing infrastructure of cheap electricity and telephony. There was both an unrealized and unmet demand with a flexible infrastructure to support deployment and development as the networking technology improved. In sum, from roughly 1980 to 2000, computational behaviors expanded beyond science and technology in the universities and corporations first to an international community of hobbyists, many of whom were graduate students at the labs where the technologies were developed, and then to ordinary people. Thereafter, it was a feedback of lower costs, interesting content and growing demand interacting in ways that have yet to be teased out.

As of the most recent data I could find in mid-August 2008, just under 22 percent (21.9%) of the world's population had some form of Internet access, a three fold increase since 2000. Internet penetration, an unlovely term which measures the proportion of a population with access to the net, was highest in North America (73.6%), followed by Oceania/Australia (59.5%), and Europe (48.1%). But growth in the same 8-year period was fastest in the Middle East (1,176.8%), Africa (1,031.2%), and Latin America/Caribbean (669.3%).[9] The top three languages, by the way, are English, Chinese and Spanish.[10] As of December 2007, Internet penetration in Ireland had reached 50.1 percent.[11] The most recent information posted to the page maintained by the Central Statistics office showed that in the first quarter 2007, 57 percent of households had Internet connectivity, up from 45 percent in 2005, and that 31 percent of households had broadband, up from 7 percent two years earlier. Sixty-five

[7] Stephen Murray, Art History and the New Media: Representation and the Production of Humanistic Knowledge, [October 2008], http://www.clir.org/activities/digitalscholar2/murray11_11.pdf.
[8] A concise history of electronic spreadsheets can be found in D. J. Power, "A Brief History of Spreadsheets", DSSResources.COM, World Wide Web,http://dssresources.com/history/sshistory.html, version 3.6, 08/30/2004. Photo added September 24, 2002.
[9] World Internet Usage Statistics News and World Population Stats, Page updated 8 August 2008, http://www.internetworldstats.com/stats.htm.
[10] Top Ten Internet Languages – World Internet Statistics, [1st quarter 2008], http://www.internetworldstats.com/stats7.htm
[11] Internet Usage in Europe, [2008], http://www.internetworldstats.com/stats4.htm#europe.

percent of all households owned at least one computer and 87 percent of these computer-owning households were connected to the Internet.[12]

That's a substantial audience. Moreover it is interactive: Internet users upload as well as download information, unlike earlier, one-way forms of mass communication: newspapers, books, magazines and broadcast. As a result, more and more information has been pushed out to more and more people, leading, among other things, to a devaluation of information and intense competition for attention. In its entirety, the web constitutes a vast market for human attention, what physicist Bernardo Huberman of HP Labs calls "social attention".[13] Rumors do ricochet in echo chambers of screaming messages, and sourcing in the traditional sense can – and does – break down. At the same time, the logic of the technology has been to push more and more capability to the individual, because of powerful capacities locally or because local capacity affords fast remote access to even more data and more power. Information is plentiful and cheap, and human attention is scarce and valuable, and the line between local and remote disintegrates into a paradox: an empowered local user who is dependent on a complex and largely unseen infrastructure. (I am writing this paper in my study at home but with near-instantaneous access to global resources, whether I want to check the spelling of a word or confirm the existence of papers in esoteric journals.)

Users tend to ignore this paradox and focus on two principal features: access and interactivity. *Access* has several meanings. Access enables end-users either to have substantial processing power on multiple personal devices, from the desktop to the handheld or to be able to get such power remotely – or both. Access also embraces access to information, which is becoming not only abundant but also heterogeneous. *Interactivity* means that end users can find like minded souls, provide feedback, and form and dissolve communities. For scholars, this combination has exacerbated the tendency to identify with a discipline rather than an institution, like the university, particularly in the sciences, although they implicitly expect the institutions – notably the libraries and centrally administered IT systems – to support them. This trans-institutional behavior is most evident in the sciences – FermiLab in Illinois, CERN in Switzerland, the telescopes in Chile and Hawaii are but a few examples.

Unlike the physical world, the virtual world is highly transient and unstable in both social and informational senses. The Washington Post website is updated continuously through the 24-hour day so the site is not identical from one hour to the next although the designers work hard to maintain consistent branding elements. But consistent brand is not enough for research. The ability to validate sources or replicate the experiment is unquestioned. But how do we expect scholars to use the resources on the net if we cannot guarantee the preservation of the content they find there? Thus, a challenge for scholarship is both to benefit from the interactivity but preserve sufficient stability where it is needed, notably in sources.

For those of us who manage resources on behalf of scholarship, merely coping with information abundance is a challenge, particularly when we consider the need to preserve sufficient stability in sources. Sustainable long term preservation and access is a major problem. Well-managed scientific datasets are probably not at risk, but smaller collections, whether in the sciences or the humanities, are more difficult. They are widely dispersed, subject to different rights regimes, highly heterogeneous in content and format; and in many cases, are owned by scholars who may be unaware of broader value or reluctant to release

[12] Central Statistics Office, Information Society Statistics, First Results 2007, 30 November 2007, http://www.cso.ie/releasespublications/documents/industry/current/iss.pdf
[13] Bernardo A. Huberman, Social Attention in the Age of the Web, [October 2008], http://www.clir.org/activities/digitalscholar2/huberman11_11.pdf

their research material even to the custody of the library. At present, there is a vague sense that the libraries will eventually take on the custodial and preservation function, whether they do it directly or manage it with a series of third party services.

That is scary.

The technology evolves; information quite literally decays, so that archival tapes can no longer be read, and rights regimes vary and seem, in some instances, to be insufficient to enable cultural institutions to exercise their preservation mission on behalf of scholarship. Finally, the costs are unclear, although the LIFE project in the U.K. has begun to make substantial progress on modeling costs of long term digital preservation. In the U.S., a task force has been formed with international participation to ask the questions, how much does it cost? And who should pay? Scholars tend to identify most closely with their disciplines and not necessarily with their home institutions. Yet their home institutions provide substantial components of the research infrastructure, notably the university library, resulting in a disconnect between scholars' research orientation and the resources to support that work. Implicit in the second question, who should pay, are issues of public policy as well as economics so much is on the table. However those issues revolve, it is very clear that as digital information grows more abundant and heterogeneous, so too do the issues of managing it on behalf of scholars for whom replication of results and reliable validation of sources are critical.

Stovepipes and the Challenge of Abundance

Early projects in the humanities tended to focus on breaking down barriers, especially barriers created by limited access. Initially, there was enormous interest in efficient and reliable digitization of source materials through a combination of scanning and optical character recognition, followed by tool development with admittedly uneven results. Humanist scholars took the lead in the development of mark-languages, which were systems of codes that permitted the idiosyncrasies of many historic materials to be rendered or displayed on screen with reasonable or perhaps high degree of fidelity and without resorting to page images, which cannot be searched or analyzed and therefore could not support constructing sophisticated concordances or similar sorts of text analysis. With greater or lesser degrees of sophistication there arose a number of digital collections, some of them thematically organized; others organized around an object, like Beowulf or Roman de la rose or a figure or group of figures; and others representing digital equivalents of pre-existing collections such as Early English Books or papyri. Most of them required and still require extensive manual input – what we call "post processing."

Two of these early projects, the Electronic Beowulf and the Perseus Project, are instructive. In the Electronic Beowulf, a project sponsored by the British Library, the surviving 11th century manuscript, which had been damaged in a fire in 1731 and conserved in 1845 in a display case, was removed from the case, digitized, and subjected to fiber-optic and ultra-violet readings that displayed characters and features no longer visible to the human eye. Beyond enabling a new reconstruction of the manuscript itself, the current project has expanded to encompass digital images of the Beowulf Manuscript, images of the eighteenth-century transcriptions, copies of the 1815 first edition with early nineteenth-century collations of the manuscript, a comprehensive glossarial index, and a new edition and transcript, and fairly sophisticated search facilities. The Perseus project, a multi-lingual, multi-cultural library of digital texts, images, and reference tools, begun in 1985, is another particularly long-lived example. Its editor-in-chief, classicist Gregory Crane, originally conceived of it as stand-alone product distributed on discs, it was migrated to the web in 1995. Its content has expanded from its original conception as a library of texts in classics to

include a broader range of material and the resource has witnessed a series of engineering upgrades.[14]

These two projects, The Electronic Beowulf and Perseus, illustrate the two major impulses behind early humanities projects: depth and breadth. Sophisticated cameras and digital representation meant that a rare artifact like the 11th century Beowulf manuscript could be reconstructed and reinterpreted and then the results made available to a broader community of scholars. Fragility alone was reason enough to restrict access the physical artifact even assuming that scholars could make their way to London. Perseus attacked the barrier of geographical access somewhat differently by assembling related material in a single source that again could be made broadly available together with increasingly sophisticated glossaries, concordances, commentaries and tools.

Both of these started out as disc-based distribution projects. The web, which began to diffuse broadly among libraries and archives in roughly 1992-1995, by its very nature as a distributed architecture meant the user could experience a single collection that in fact might be supported by a number of physically distinct institutions. The distributed structure of the content also has the practical benefit of distributing responsibility for maintenance of the resources among a number of organizations. Finally, ever improving search technology afforded users detailed access to digitized materials whether that meant merely finding them by submitting a query in a search engine or meant examining a collection more closely with a customized interface.

These are only two examples; there are many others and many refinements as advances in the study of texts, in particular, improved the ways that materials could be represented and analyzed. For example, researchers at the Institute for Technology and the Humanities at the University of Virginia experimented with ways to juxtapose text and images so that materials would display consistently across a variety of browsers; this was important for certain classes of works, like those of William Blake, where the visual quality of the artifact was important to understanding the significance of the work. Very interesting work is also underway in archaeology and city planning, employing geographical information systems and visualization technologies. The scale of such collections has become increasingly ambitious as scholars become more familiar with the collaborative organization and technologies and see the value of assembling collections and services. Recently, for example, Rice University in Houston, Texas, forged a collaboration with the University of Maryland and the Instituto Mora in Mexico City to organize a digital collection and services around digitized archival collections focused on the Americas and now owned by these three geographically distinct institutions. Besides aggregating geographically dispersed materials in multiple languages in a single searchable resource, the Our Americas Archive Partnership encourages a broad range of scholars to view their research in hemispheric terms and "to pry it loose," in the words of project director Caroline Levander, from the self-limiting assumptions of the nation state. [15]

A number of these early projects provoked the formation of centers where scholars from disparate departments within the university might congregate around shared interests. As the scope for collaboration has expanded, many of the early single institution, stand-alone collections and collections have become seen as stovepipes.[16] Not only do the centers seem to inhibit disciplinary, trans-institutional collaboration, which is how scholars increasingly

[14] http://www.perseus.tufts.edu/help/versions.html
[15] Caroline Levander, The Changing Landscape of American Studies in a Global Era, [October 2008], http://www.clir.org/activities/digitalscholar2/levander11_11.pdf
[16] Diane Zorich, A Survey of Digital Humanities Centers in the United States. Prepared for the Council on Library and Information Resources (CLIR). 2008; http://www.uvasci.org/current-institute/readings/dhc-survey-final-report-2008/

interact, but these collections are expensive to create, requiring substantial manual coding and other treatment. While there have been advances in automating aspects of these processes, the model of highly curated collections is now challenged by abundance, specifically the massive output of large scale digitization projects supported by Google, Microsoft, the Million Books project and others, to convert to digital form tens of millions of printed books, including the entire general collections of major research libraries. The consequences are obvious: We move from well-defined, well-managed, and labor-intensive digital collections to literally floods of data with minimal post processing treatment. Inconsistencies in editions, possible errors in scanning itself, or even variations in the original texts are presented without or with minimal explanation to the user. It is reasonable to expect that advances in technology will improve the process, leaving us with the more interesting question of what happens when we go from a world in which information is abundant rather than scarce?

Abundance, it turns out, is itself a challenge. I've already mentioned the scary prospect of digital preservation. Scientists are worried about managing ever increasing quantities of data, measured in petabtyes and exabytes – figuratively speaking, single collections the size of truckloads of discs. More precisely, estimates of total information created exceeded total storage capacity for the first time in 2007. For years, the mantra was "storage is cheap." Perhaps, but clearly capacity is not infinite. The conversation in the storage community revolves around the question of ability of current approaches to support massive scale in data and the prognosticators are not always positive.[17]

So much for the pessimists. We optimists are relishing the challenge of scale and unruly data because they offer opportunities to re-examine assumptions and to undertake new research to answer new questions or to answer old questions in new ways. Professor Crane has outlined a series of capabilities that, if realized, would help us deal with language: tools, for example, that recognize the differences between "Washington" the man, "Washington" the state, "Washington" the city where I live, and all the other possible "Washingtons". I ran a Google search with term "Washington" and got 544,000,000 hits. Thankfully, Washington, D.C. was in the first 10. Imagine the possibilities when we scale up to collections of tens of millions of books in multiple languages, including volumes in non-Roman scripts. Coming to terms with abundance, Dr. Crane argues, requires technological advances in two broad directions: one, the ability to dig very deeply into more heterogeneous texts; and two to increase the audience for scholarly work, that is, to expand the breadth. In his own field, depth and breadth would allow "those of us who have dedicated our lives to the study of the Greco-Roman world," he writes, to have access to the tools and scholarship of those "who see the Persian Empire as their cultural heritage and to be able to study the sources on which that present perspective rests."[18]

There exist several decades of computer science that go under the heading "automated language processing" that can be brought to bear on these kinds of problems and that would find such huge and messy corpora fascinating. When you get inside these techniques, you quickly discover that they are based on probability theory, which is a different model of

[17] J. Gantz, C. Chute, A. Manfrediz, S. Minton, D. Reinsel, W. Schlichting, A. Toncheva, The Diverse and Exploding Digital Universe; an Updated Forecast of Worldwide Information Growth Through 2011. (March 2008) An IDC White Paper, Sponsored by EMC. http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf.; M. Peterson, G Zasman, P. Mojica, JPorter, 100 Year Archive Requirements Survey. Storage Networking Industry Association, Data Management Forum, 100 Year Archive Task Force (2007).

[18] Gregory Crane, Alison Babeu, David Bamman, Lisa Cerrato, and Rashmi Singhal, Tools for thinking: ePhilology and Cyberinfrastructure, [October 2008], http://www.clir.org/activities/digitalscholar2/crane11_11.pdf

reasoning, where we talk about statistical likelihood, degrees of confidence, and the presence of error. These are precise terms with well-defined mathematical meaning. The bad news for some humanists, computer scientist Douglas Oard tells us, is that "humanities scholars are going to need to learn a bit of probability theory."[19] So much for those of us who may have fled into the humanities after a bad calculus day.

Beyond doing more better, we have an opportunity to do new things differently, including rigorous study of the World Wide Web as itself an object of communication – as well as a technological achievement. Bernardo Huberman of HPLabs argues that the web of linked information carried by the web can be viewed as a social phenomenon. Can you look at the extent to which social networks as they function through the web mediate the allocation of limited human attention – the scarce resource -- he asks? Can one discover social networks through the web and then show how ideas propagate and eventually decay?

So far, he and his team have demonstrated two major effects: (1) They identified and graphed relationships among people based on patterns in purchase and recommendation of medical books and Japanese graphic novels using data from Amazon, containing 15 million recommendations of books recommended by more than 5 million people who purchased them. The team was able to replicate these results using collections of e-mail, thus validating the technical research, the algorithms, and the baseline argument: that the collection of information embodies a web of social relationships that cluster around information objects. (2) They mathematically modeled the propagation of ideas contained in a collection of news stories within a group of one million users and showed that a few ideas command a burst of interest and enthusiasm, as the ideas propagated within the social networks of readers and are reinforced through recommendations. Ideas, they also showed, decay slowly.[20] In certain communities, a particularly esoteric branch of physics is his example, this type of winner-take-all vetting allows a market place for ideas to emerge in which "intense chatter", as Huberman characterizes it, "serves as a good quality filter." This model also explains how shrill voices can escalate and why looking across communities that form around information make for a web that seems to be in a constant state of hubbub. Artists, writers, those in the arts and letters, he argues, have a key role as the creators of "attention structures," the vehicles in which information is embodied and conveyed.

Neither social network theory nor study of the diffusion of ideas is new, and Dr. Huberman does not claim that they are. What is significant is the scale – millions of users and millions of transactions – and rigor as well as confirmation of the importance of the medium as an incubator and shaper of opinion. Thus, while the information technologies are frequently and correctly characterized as multi-purpose, the way that they are deployed and the human behavior that they support and encourage – the speed, the distribution, the ability of almost anyone to upload information – means that the web becomes an independent actor and the technology is not strictly neutral. As the web continues to grow, so too will its opportunities to foster, and not merely support, new communities and forms of communication that will shape the way we view ourselves in the physical and digital worlds.

Whither?

I mulled over several ways to end this paper, and since I was coming to Ireland, I hunted around in poetry for an appropriate reference, perhaps something elegant about new worlds. A search on "Donne" and "America" did bring up the Elegy as well as a pop-up advertisement

[19] Doublas W. Oard, A Whirlwind Tour of Automated Language Processing for the Humanities and Social Sciences, [October 2008], http://www.clir.org/activities/digitalscholar2/oard11_11.pdf
[20] Bernardo A. Huberman, Social Attention in the Age of the Web.

for lingerie. It was mildly amusing and perhaps instructive of the limits of available technology but probably not appropriate. In the end, I have settled for one more story of my own. This one goes back several years to my time as an editor of a magazine on information and public policy, when I interviewed Gordon Bell. Dr. Bell did fundamental work in the engineering of modern computers and played a critical role in the organization of funding for computer science research at the U.S. National Science Foundation, which is the premier sponsor for non-defense related basic research in science in the U.S. He is delightful and gracious but was also a difficult person to interview because he speaks quickly and seemed to believe that I, as the reporter, really could understand the technical papers he e-mailed ahead of time.

So there we were in a relatively quiet corner of a large hotel in Washington doing the interview, which had been scheduled for an hour and went on for two. I was trying to push him to justify basic research, and he said – and now I'm paraphrasing – that he didn't want to discuss the justification for basic research. Rather, he said, he preferred to imagine things he'd like to do, to "dream" of things to do – and "dream" was the verb he used – and then to go build them. Which he indeed did and with very good results.

Not all of us can expect to be Gordon Bell (or John Donne, for that matter). But everyone can dream.

Thank you for listening to mine.