

Scholarly Technology in the Humanities at Northwestern University: Some Reflections and Recommendations

By Martin Mueller

Introduction.....	2
Enhanced access to digital archives.....	2
The tool conservatism of humanists	4
Scope of this Report.....	5
Digitizing a cultural heritage	5
Thinking about the curatorial function of humanities scholarship draws attention to the benefits of working with other institutions of cultural memory	6
From scholarship to pedagogy	7
Changing the implicit contract of a print based library in a digital world.....	7
All about bits.....	10
The cost of bits.....	11
Are computers ‘computers’? or All about lists	12
The digital surrogate and its query potential.....	13
Availability as the first-order advantage of digital surrogates.....	13
Second order advantages of digital surrogates.....	13
The MONK principle: Metadata Offer New Knowledge	13
The query potential of digital images	16
Building a scholarly infrastructure for the humanities: What has been and should be done at Northwestern	17
Should we commit faculty resources to the building of digital cultural archives?...	18
A tour of some Northwestern projects	20
Carl Smith’s projects.....	20
The Encyclopaedia of Chicago	21
The Mellon International Dunhuang Archive.....	21
Oyez	22
The Vesalius Project	22
Other archival projects.....	22
Project Pad	23
WordHoard	23
Virtual Orthographic Standardization of Early Modern English texts	23
The people behind the projects	24
What should be done?.....	25
ERANOS: collaborative tools for querying and annotating digital archives.....	26
MONK and ERANOS.....	29
Educating Users	30
Some final reflections on space in the library.....	32
A ‘back to the future’ model for the third floor of the South Tower in the Library .	33
Appendix: List of Recommendations	34

Introduction

Enhanced access to digital archives

The following is a set of reflections and recommendations about scholarly technology at Northwestern University. Think of it as a private follow-up to a task force on Humanities Computing in a Networked Environment that I chaired over a decade ago.¹ This report is based on conversations over the past year with many people inside and outside the institution. I have tried to avoid riding my own hobbyhorses too hard, but I have not tried to produce a consensus document. What I have written is likely to suffer and benefit from the focus that a single observer brings to the consideration of an issue.

For a variety of reasons I think that scholarly technology in the humanities is most productively seen as an extension of library services. The library is the laboratory of the humanities. For a century or more, the American research library has been a very complex logistical system using the technologies of the day to develop a national “Internet.” think of the LC catalog and Interlibrary Loan. Since the sixties, the inventory management of libraries has been computerized; since the eighties, primary and secondary materials have been increasingly available in digital format. These changes have a powerful impact on what and how we read.

A possible tag line for this report is “Enhanced Access to Digital Archives.” We live in the midst of a migration of much scholarly information into a digital medium. Access to digital resources is of course measured first of all by how many of them you have. By and large more is better. But beyond a certain point—which we have probably passed—getting more stuff may add less value than making better use of what we already have. I do not want to suggest that we should stop getting more, but I will argue that the maintenance of a good digital documentary infrastructure must look beyond just getting more. It remains true that you cannot use what you have not got. It is also true that there is no point in having something unless you know what to do with it. Attending to the second truth will, at least for a while, have a better pay-off than focusing exclusively on the first.

Another way of making this point is to draw a distinction between first and second generation digital archives of primary materials. First-generation archives focus on providing simple access. Second-generation archives turn digital documents into highly manipulable objects that support new forms of inquiry. Where first-generation archives have been constructed in a responsible fashion, their transformation into second-generation archives can be done largely by building on what has already been done.

¹ The report is still available at <http://www.it.northwestern.edu/about/committee/hcne/index.html> and is a useful point of departure for gauging progress over the past decade. Ten years ago many humanities faculty lacked network access or network capable computers. Many library services did not work over the Northwestern network. Those problems have been solved. Close cooperation between the Library and the IT organization, in particular Academic Technologies, has become a daily reality far beyond what we expected then. And I understand that we are at last within weeks or months of implementing a state-of-the-art model of a key recommendation of that report, “a digitized image service to support the casual and systematic use of digitized images by faculty and students for instruction and research.”

The enhanced access provided by second-generation digital archives will play an important role in maintaining the scholarly infrastructure that is critical to research and advanced instruction.. I envisage a three-pronged strategy that includes

1. Adding intelligence to existing digital resources, typically in the form of metadata
2. Creating better tools for exploring the query potential of digital resources in an environment that supports collaborative work
3. Educating users to identify and take advantage of the distinct query potential of digital resources

As I write this report, I am aware of considerable interest on the part of the central administration to invest in the humanities, raise their relative strength in the university, improve our ability to recruit excellent graduate student, and enrich research opportunities for upper-level undergraduates, quite a few of whom are capable of work that in imagination and originality yields nothing to the work of graduate students.

The continuing conversation about the best form of such investments has in the past two years produced two high-level reports by groups of humanities faculty, neither of which identified technology as a topic for formal discussion. At a national level, however, there has been a good deal of discussion of this matter. The ACLS last year appointed a Commission on Cyberinfrastructure for the Humanities & Social Sciences. Its draft report, recently released, has much to say about lack of knowledge and appropriate skills among scholars, teachers, students, and professionals as a major cause of relatively slow progress.² It is a telling corroboration of such analysis that at Northwestern—not untypical in this regard—not one, but two humanities committees chose not to pay any attention to technology issues when asked to think broadly about the future of their disciplines.

They may have believed that technological changes can safely be left to librarians and information professionals because these changes do not affect the modalities of scholarly work in interesting ways. This is wrong. On seeing and playing with a very early typewriter, Nietzsche is supposed to have said “Unser Schreibzeug arbeitet auch an unseren Gedanken mit” (our writings tools cooperate in our thoughts). Tools are never mere instruments but define the calculus of the possible. The building of a first-rate scholarly infrastructure—a combination of digital archives with deep query potential and of tools to exploit that potential—is done best through cooperative ventures that involve librarians, programmers, and faculty in equal measure. At the very least faculty should take a strong interest in what the librarians and programmers are doing, since the cumulative effect of

² The draft report is available at <http://www.acls.org/cyberinfrastructure/acls-ci-public.pdf>. Northwestern’s Sarah Fraser has been a member of this commission, which has been chaired by John Unsworth, the dean of the Graduate School for Library and Information Sciences at UIUC.

what they choose to do (or not do) will define parameters and constraints of future scholarly work.

Recommendation #1: The quality, maintenance, and strengthening of an increasingly digital scholarly infrastructure should be a continuing topic of high priority in any discussion among faculty and academic administrators about priorities and investment in the humanities. At a broad planning level, chairs and program directors in the humanities should give to this matter the same attention that chairs in the sciences give to laboratories and equipment.

The tool conservatism of humanists

Across the humanities there still remains diffuse but strong resistance to the sea changes in information technology that are transforming the documentary infrastructure of the profession and will alter ways of doing business, with all the losses and gains that deep change inescapably involves. In the contemporary American university humanists overwhelmingly identify themselves with a progressive left in a wider social and political climate that they perceive as markedly reactionary. But as humanists they are also part of a scribal culture with a long tradition of tool conservatism: in the early sixteenth century, the Abbot of Sponheim fulminated in print against the evils of the printing press. His point was made more succinctly by Filippo di Strata, who observed that “the pen is a virgin, the printing press a whore.”³ The anti-digital versions of such rhetoric are legion. From a slightly different perspective, Julia Flanders has written eloquently about this topic:

Within the conceptual framework of the humanities, the computer and its associated labor occupy the space of the scholar's demoted Other. . . . in the broadest and vaguest sense, the computer is imaginatively identified with the negatively valued side of the body-soul binary: it is pure body without a soul. Through whatever related thematizations we choose to explore this binary, we find the computer on the wrong side, humanistically speaking: it is aligned with factuality rather than imagination or interpretation; with reductivity rather than holism; with the detail rather than with the universalizing consciousness.⁴

If I look at the deep changes that have transformed the world since I was a graduate student four decades ago, the following (and multiply interwoven) themes stand out:

1. Decolonization, post-colonialism, and the establishment of English as a global language
2. Civil rights

³ “Est Virgo Hec penna: Meretrix est Stampificata.” Quoted from Wendy Wall, *Imprint of gender: authorship and publication in the English Renaissance* (Ithaca, 1993).

⁴ In her recent Brown dissertation on *Digital Humanities and the Politics of Scholarly Work*, which is accessible at http://www.diegesis.net/julia/thesis/flanders_dissertation.html.

3. Feminism
4. Globalization and changes in transportation (the Boeing 707 and its successors)
5. Computers and communications technology

In my discipline of literary studies, it is very striking how deeply the first four of these have changed the way we do business, and how much reflection on the discipline ('theory') has been shaped by them. It is equally striking how little practical or reflective attention has been given to the implications of information technology for literary scholarship. One would be ill-advised to boast of one's indifference to postcolonial or gender issues in an English department, but it remains quite fashionable to say that one neither knows nor cares about information technology. Other humanities disciplines do not appear to differ much in this regard.

Scope of this Report

Digitizing a cultural heritage

My focus on this report is on the implications of information technology for scholarly work done in the humanities departments of WCAS. The humanities in the college are not the same as the humanities at Northwestern. Music, theatre, and film are the responsibility of the School of Music and the School of Communication. Information technology plays a powerful role in these fields. While there is overlap, the technological challenges and opportunities in those fields are somewhat different, and in the Northwestern environment thinking and planning about them is most properly the role of faculty in those schools.

From a practical perspective, humanities scholars spend much of their time examining surrogates of the primary documents that are their object of study—whether slides or prints of works of art, parliamentary records, editions of Chaucer or Hobbes, and the like. For students of phenomena younger than twenty years, the primary sources increasingly include materials that are natively digital—a development with deep and largely unknown consequences. But in all fields where the primary materials are older than twenty years, scholars study primary sources that largely or exclusively began their life in a non-digital format. In many fields this is true for most primary materials up to the present moment.

These primary materials are often called 'cultural heritage' materials, especially if they are old. For the twentieth century they include films and sound recordings, but as you go back in time the materials become increasingly textual and visual documents. The central focus of this report is on the challenges and opportunities that the digitization of such materials poses for scholarship.

Thinking about the curatorial function of humanities scholarship draws attention to the benefits of working with other institutions of cultural memory

Workers in the vineyards of the arts and humanities are either creators, performers, or curators.⁵ I understand the curatorial function very broadly as caring for whatever is thought worth keeping.⁶ Such caring extends across a seamless continuum of activities that include physical preservation, restoration, editing, explication, and reflection. These activities are often—and not helpfully—divided into a ‘lower’ and ‘higher’ domain, but it bears remembering that much great scholarship has been done in the ‘lower’ domain and much bad scholarship is of the ‘higher’ kind.

The humanities in a university are a form of institutionalized cultural memory. Historians, art historians, literary critics, and even philosophers spend much of their time as curators in a museum of the human past. The ‘re-mediation’ of cultural memory in digital form presents many similar challenges to the academy and to museums in the narrower sense of the word. These challenges support new forms of cooperation. Chicago is a city of superb ‘museums’, whether you think of the Art Institute, the Chicago Historical Society, the Field Museum, the Newberry Library, or the Chicago Shakespeare Theatre. All of these institutions present opportunities for cooperation—and not incidentally many research opportunities for students and the undergraduate and graduate level. Some of the Northwestern’s most successful humanities computing projects have turned on such cooperation.

Recommendation : Northwestern should increase its cooperative projects with other institutions of cultural memory. Such projects are likely to strengthen, rather than detract from, the building of an appropriate digital infrastructure for scholarship in the humanities at Northwestern.

⁵ Whether performing is a distinct activity or a hybrid of the creative and curatorial is an interesting question.

⁶ My emphasis on the curatorial is not intended to downplay the creative. In thinking about information technology in the humanities, the most important question may well be: What changes will digital technology bring to the creation of the cultural objects whose interpretation has been a central activity of humanities scholars? For several millennia, writing was the only affordable technology for communicating ideas across distance. Writing poems (or composing music) has never required large capital resources. Painting is more expensive, sculpture and architecture much more so.

It must have been a considerable financial sacrifice for George Gershwin’s father to buy the piano on which his gifted son developed his genius. Today, a much lesser sacrifice enables a talented child to combine words, sounds, and images in various ways and ‘publish’ his creation on the Web for everyone to see. In such a world, ‘poetry’ as a ‘text only’ mode of production is no longer supported by economic constraints. Whether it survives in its ‘pure’ form is a matter of choice. The long-term implications of these changes for literature and other forms of art remain to be seen: one would be foolish to predict what will or will not happen ten years down the road.

From scholarship to pedagogy

I say little about directly pedagogical uses of technology. Important as pedagogy is, its practice in a university like Northwestern must be explicitly recognized as occurring in a research-driven environment. Whatever the rhetoric and occasional recognition of teaching or service, your standing among colleagues and administrators is overwhelmingly determined by your research prowess. While this is as it should be, some consequences flow from it. If there is a perception that the use of digital resources in the humanities is sub-optimal, any initiatives for change must be in tune with the institution's prestige hierarchy. Faculty will make use of technology if it helps them with research that is valued by their colleagues. And they are very likely to use in their teaching what they have found valuable in their research. Trickle down will work better than trickle up.⁷

A broad definition of scholarship, however, will include a lot of pedagogy. The kinds of archives and tools that will help the scholar will also find uses in graduate or undergraduate honors seminars. But in thinking about information technology in the humanities at a place like Northwestern, scholarship is a better point of departure than pedagogy.

Changing the implicit contract of a print based library in a digital world

Scientific work occurs typically in laboratories where natural phenomena are variously bottled and manipulated by Bunsen burners or their high-tech descendants. The science library holds the published records of such inquiries. For the humanities scholar the library is both laboratory and library. It contains versions of the 'primary' materials that are the objects of scholarly attention—whether texts, slides, scores or other media--and it contains the 'secondary' literature that records the scholar's interpretative encounters with primary materials. The distinction between 'primary' and 'secondary' is relative and fuzzy at the edges, but it is useful for many purposes.⁸

Libraries have changed very deeply over the past three decades. For scientists, these changes amount to changes in the way the 'literature' about their discipline is managed. For humanists, however, the changes also affect the representations of their primary materials. Information technology driven changes affect both the 'laboratory' and the 'library' of the humanist. Librarians are generally more aware than humanities scholars of the challenges and opportunities of these changes.

⁷ The situation is different with faculty whose jobs are defined as predominantly pedagogical. In the field of language instruction Franziska Lys, Janine Spencer and the technical staff of the Multimedia Learning Center have done world class work in creating various kinds of computer-assisted and media-rich environments for language learning. They have also made substantial contributions to the methodological discussion of such tools.

⁸ Not all scholarship in the humanities is 'document-centric.' Philosophers may be thought to confront the world directly, and some secondary materials become primary. But most scholars operate with some version of a distinction between sources and literature about the sources.

Crudely speaking, a library in a print environment rests on an implicit contract where the librarians provide the books, and readers provides the access tools, i.e. their reading skills. If librarians buy and catalogue the books, put them on the shelves, help readers find what they need, and implement circulation policies that maximize a fair distribution of access to the books, they have done their job. What readers do with those books is their business. It is certainly not the librarian's task to teach the reader how or what to read.

In a digital world, some aspects of this contract are subject to change, especially as it concerns access to primary materials. Digitization has the potential for enhancing access to such materials. In my view such enhancements of access should become part of a new contract between librarians and scholarly readers in a research environment. Sophisticated programmers—whether in a library or in an IT organization—are the people with the skills to implement such enhancements.

Let me develop this view by looking at the implications in a digital world of Ranganathan's charming and deceptively simple five laws of library science:⁹

1. Books are for use
2. Every reader his or her book
3. Every book its reader
4. Save the time of the reader
5. A library is a growing organism

'Book' here stands for any entity intended to be 'read' in the broadest sense of the term: books, slides, scores, web sites, etc. are all 'books', and a 'library' includes any institution intended to mediate access to 'books' in that sense—an interpretation that follows directly from the last law, where the meaning of 'growing' surely includes 'change' as well as increase in size.

"Books are for use" and "Save the time of the reader" interact interestingly if you look at these laws from a digital perspective. Whatever we do depends on some initial assessment of the likely time cost. "Had we but world enough and time," we would sooner or later do everything. But we don't. Most things take longer than we think. Many things go undone for lack of time. Some things get done only because we fooled ourselves into thinking that they would take much less time, but then we 'made' the time to finish them anyway. But whatever we undertake is supported by some belief that we have the time to do it. And time is the ultimate currency.

Digitization changes the time calculus of many activities. Separately these changes may be trivial. Together they produce a change in kind. Just about any interesting aspect of digitization can be represented as flowing from calculations about the time it takes to do this or that. For a particular section of an article I should really look at that pamphlet in the Bodleian. But I cannot go there right now, and I drop that section. If my library has a

⁹ S. R. Ranganathan (1892-1972), a mathematician turned librarian is the father of modern library science. His five laws of library science are to libraries what the ten commandments are to the Bible.

microfilm of that pamphlet I need not travel to Oxford (fortunately or unfortunately). If my library has a subscription to Early English Books on Line (EBO), I can look at it over a broadband connection from home at 2 am (I am unlikely to wax nostalgic about the lost pleasures of working with microfilm).

It is easy to interpret decisions to purchase the microfilm or replace it with EBO as applications of Ranganathan's fourth law. But two quite different scenarios follow from the law. In the first, readers save time by doing more quickly what they would have done anyhow. In the second scenario readers spend time on something they would not have done otherwise because it suddenly seems possible or, quite literally, 'worth their while.' It seems to be a relatively straightforward matter to calculate the savings that flow from the first scenario.¹⁰ It is much harder to estimate the expenditures that flow from the second, not to speak of the benefits that accrue over time when people start doing things they would not have done otherwise.

Humanist scholars of 'primary' documents typically encounter the objects of their attention in relatively unmediated forms. The texts, scores, or images they read or look at are rarely the originals, but the surrogates pretend to be close to the originals, and the scholar behaves as if they were. The digital representation of such documents supports their transformation (or deformation) into a variety of different shapes that in principle support answers—and speedy answers at that—to questions that it was previously impracticable to ask. "Books are for use," Ranganathan says, and the digital 'book' (text, score, image, etc) supports new uses. Ranganathan also says "Save the time of the reader." If enhanced in certain ways, the digital 'book' acquires new properties, and if such enhancements are extended to large collections of digital 'books', there is a strong multiplier effect: the query potential of the sum of such books is much greater than that of their parts. The enhancements 'save' time for known procedures, but more importantly they encourage readers to 'spend' time on things previously considered impracticable.

With current technology, it is in principle possible to turn large portions of the humanist's primary archives into highly manipulable digital surrogates in something like a laboratory environment. Spending affordable amounts of time, scholars can then force such surrogates into variously abstracted or reduced forms that may produce new insights. But for such procedures to work, you need new tools and methods for manipulating the enhanced 'books' (texts, scores, images) in productive ways. The contract between librarians and readers requires further revision, for the built-in tools readers bring to their task—their readerly skills of various kinds—need to be supplemented by all manner of laboratory equipment—e.g. search engines, data mining routines, image manipulation tools and the like—that need to be purchased or developed and certainly maintained at non-trivial expense.

None of this may sound very inviting either to the humanist or to the librarian. But if you look closely at the implications of digital technology for Ranganathan's laws in a chang-

¹⁰ I say 'seems' because devices for saving time invariably create opportunities for squandering it. It is not easy to argue with those who say that more time is wasted than saved and that as a result we are busier and more distracted than we were before.

ing world, you are led to the conclusion that in their stewardship of primary documents, librarians in research universities should think of such documents as increasingly existing in highly mediated laboratory environments where they are subject to complex manipulations with sophisticated tools.

Recommendation #2: It is useful to distinguish fairly sharply in the humanities between primary documents and secondary literature. Secondary literature—the books and articles of scholarship—is subject to generic library routines that are fairly constant across disciplines. The maintenance of primary documents—in whatever medium— should be envisaged as taking place in a laboratory like environment, where large document archives, enriched by metadata in discipline-specific ways, can be manipulated with the help of sophisticated tools.

Recommendation #3: The maintenance of such a laboratory environment, which provides enhanced access to digital archives, should be the joint responsibility of the Library and of Academic Technologies.

All about bits

I am a great admirer of Janos Starker both as a cellist and as a teacher. On the several occasions that I went to his master classes I was invariably struck by how few of his remarks were concerned with the “higher” criticism of interpretive choices and how many of them addressed the question of how the student’s bow touched and moved across the string: the height of the chair or length of the cello pin, the student’s posture, his way of holding the bow, and so forth.

I learned from this that it is always worth attending to the basic building blocks that enable and constrain whatever you do. While it is true that systems of complexity have emergent properties that cannot be predicted from those building blocks it is also true that the building blocks define the outer limits of the possible. Books and bicycles are marvelously transparent technologies: a child can understand how they work and what you can or cannot do with them. By comparison, computers are very opaque objects, but it is very helpful to ask what is the computer’s equivalent to the bow touching the string and how, across many intermediate technical steps, that basic activity enables and constrains what you can do with this remarkable tool.

I like to think of the computer as the ultimate monomedia machine. It can take in real bits of the world, as long as they are represented to it in the form of **binary digits** or bits. You can move, manipulate and store these digital bits faster, more flexibly and in almost unimaginably smaller spaces than the ‘real’ bits.¹¹ And when you’re done with it you can

¹¹ The German critic Walter Benjamin was fascinated by miniatures, and somewhere he talks about the entire Torah being inscribed on a small object like a thimble. How would he have reacted to a recent article in the New York Times (<http://www.nytimes.com/2005/12/12/technology/12flash.html>), which describes the competition of Toshiba and Samsung for market share in a new type of flash memory chips that “are the size of a fingernail and can store two gigabytes, the equivalent of every word and image printed in

transform the bits into something you can make sense of. The output may be similar or different from the input. While email message goes from text to text, you can also go from score to sound file, equation to graph, or the other way round. But the enormous flexibility of the computer as a multimedia tool derives from the fact that it turns varied bits of the world into its singular reality of digital bits, moves these bits around in various ways and turns them back into some medium that a human can make sense of. The more 'multi', the more 'mono'. Whatever happens inside a computer is all about 'bits'.

The cost of bits

If you think of digital bits as a currency in which you pay for the act of digitization, an interesting relationship emerges. The more symbolic the phenomenon, the cheaper its digitization. Numbers are cheaper than text, and both text and numbers are much cheaper than images, not to speak of time-based media such as audio and video files. A numerical representation of a million takes half the space of the text string '1000000'. The jargon for this is low and high 'bandwidth'. Text is low bandwidth: you can store texts of all of Shakespeare's plays and poems in the space that it takes to store a one minute high-quality recording of a Shakespeare song. A high-resolution image of a two-page image from the *Times Atlas of the World* takes up as much space as fifty copies of Shakespeare.

Bandwidth is a constraint that operates with different power at the level of storing and moving bits. It is one thing to store a billion bits somewhere cheaply; it is quite another to move them quickly from one place to another. In the United States the cost of storing bits cheaply has dropped much more rapidly than the cost of moving them quickly. What is called "broadband" refers to the ability to move data fast enough from one place to another. Broadband is now getting a big push from the entertainment industry. The Holy Grail is an environment in which middle income people can afford to download movies over the Internet. Most of the families that send their children to places like Northwestern will be able to do this by the end of the decade. If you have a world in which the middle class can download trashy movies on demand, you have a world in which the cost of storing and moving bits no longer imposes serious constraints on what humanities scholars can do with digital technology.¹²

nine years of a newspaper"? To put this in a more scholarly perspective, a device based on such a chip, which you will be able to buy for about \$200 and attach to your keychain, will hold all the remains of Greek literature, the entire English Poetry database, and three additional Chadwyck-Healey databases that include some 200 novels of Early English Prose fiction, about a hundred eighteenth-century and 250 nineteenth-century novels.

¹² There is an important wrinkle to this. In Northwestern's campus network and similar institutional environments information travels at the same speed, regardless of direction. Affordable broadband connections in an off-campus or residential environment, whether broadband or DSL, typically have different speeds at which bits travel to or from you (download vs. upload). The service providers assume that for most users 'uploads' are simple commands like 'get me that movie', where it doesn't really matter how quickly this very short command travels across the Internet. Download speeds may be faster by a whole order of magnitude. This begins to matter where a humanities scholar, often working from home, uses programs that involve a lot of to-and-fro traffic. Relatively slow upload speeds are unnoticeable with single commands, but they make a difference when an activity that the user perceives as a single step is in fact triggers a shuttle of ten or a hundred steps. This is one example why it is always useful to remember that whatever you do with a computer you're asking it to move bits around, and what it can do for you is constrained by the number of

Are computers ‘computers’? or All about lists

Until about twenty years ago, the technological and economic constraints on the number of bits you could store or the speed at which you could move them around were such that for most practical purposes computers could only handle highly symbolized data like text and numbers.¹³ And in the earliest days of the computer, even text was expensive, and the only affordable digital operations were addition, subtraction, multiplication, and division of numbers.

If bits are binary digits there is certainly a way in which calculation may be seen as a computer’s native activity. And if it is, then computers are seen, as Julia Flanders has argued, on the wrong side of the humanist fence. On the other hand, ‘lisp’, the name of famous early computer language, is an acronym for List Processor. In a characteristically humanistic computer program like the search engine Philologic, the central processing unit spends most of its time sorting lists, which is not a particularly mathematical activity.

Irving Klotz, one of the most distinguished chemists in Northwestern’s history and a man of sardonic wit, enjoyed wondering in public what kinds of errors humanists would have avoided if Einstein had stuck to his original plan and called his famous theory the Theory of Invariance.¹⁴ In a somewhat similar vein one might wonder whether humanists would feel differently about ‘computers’ if they were called ‘list makers’. We all make lists all the time: wish lists that grow and do list we don’t get through. An academic book typically ends with an important list called a bibliography and another list called an index.

List making may be seen as a humble activity, hardly worthy of serious or ‘theoretical’ consideration. The place of bibliographies and indexes at the end of a book appears to testify to its ancillary status. But list making is also a ‘primitive’ activity in the sense in which John Unsworth has talked about scholarly ‘primitives’.¹⁵ If you think of list mak-

bits it can store and the speed at which it can move them around the slowest section of the “information highway.”

¹³ Digitization of catalog records was the first step towards digital libraries. Jim Aagard, an electrical engineer at Northwestern, played a crucial role in this process, and his contributions to a digital catalog are the basis of the most widely cataloguing software today. When Aagard did his pioneering work, it was quite a task to squeeze the content of two million catalog entries into a format where the bits could be stored and moved around with reasonable speed. Today, many catalog entries include pointers that take you to full text records of the book stored here or somewhere else. Ten years from now, it is quite likely that a sizable portion of the books in the NU library and most of the books out of copyright will be accessible digitally. They will be stored in a ‘box’ that may not be bigger than a roller bag that fits in a plane’s overhead compartment.

¹⁴ The science of this is way beyond my competence, but if I understand it correctly, the theory of relativity depends on a number that is like pi, in it can be defined at increasing levels of precision, although never absolutely. So much for ‘relativity’.

¹⁵ “Scholarly Primitives: what method do humanities researchers have in common and how might our tools reflect this?” Cited from <http://www.iath.virginia.edu/~jmu2m/Kings.5-00/primitives.html>. Unsworth’s admittedly provisional list consists of “discovering, annotating, comparing, referring, sampling, illustrating, representing.” Given the fact that the primitives are represented as a list, listing should probably be a primitive as well.

ing as an activity that deeply informs a scholarly project at each of its stages, the computer may cease to be a profane intruder on the sacred precincts of scholarly inquiry. If it can help you make, keep, and sort lists of vast length much more quickly and accurately than you can do on your own, it becomes a useful helper with a constitutive and deeply familiar activity of our daily and scholarly life.

As a list maker, the computer draws attention to the fact that complexity is built from the combination and iteration of very simple steps. Conversely, 'programming' is the deconstruction of complex activities into sequences of simple steps. All this has more to do with common sense and logic than with mathematics or numbers. But do you really want to put the humanities at one sense of a divide and logic and common sense on the other? The computer of course cannot by itself make the leap from list to insight. Therein, to vary the doctor in *Macbeth*, the scholar must minister to himself.

The digital surrogate and its query potential

Availability as the first-order advantage of digital surrogates

A digital version of a document that originated in another medium is a surrogate of a putative 'original'.¹⁶ A surrogate is never the real thing, but it may be better in some respects. The most striking advantage of the digital surrogate is its easy availability. We call this 'democratizing access', which is exactly what Filippo di Strata meant when he contrasted the virginal manuscript culture of the pen with the whore of the printing press, a metaphor that accurately current scholarly anxieties about real and imagined problems of access to cultural materials in a new medium.

What difference does such easy availability make to the nature and quality of scholarship? This is a hard question to answer. Given a good microfilm image and an accurate digital scan, there is little cognitive difference between examining the original in the Bodleian, a microfilm in your library's reading room or an EEBO image on a web browser at 2 am in your study. The differences appear to affect the accidentals of the activity but do not bear on the intellectual activity of figuring out whether or how the passage you are looking at is relevant to your inquiry. But there is the cumulative effect of more people getting at stuff in the first place or getting at it more quickly. As the Scots say, many a mickle makes a muckle.

Second order advantages of digital surrogates

The MONK principle: Metadata Offer New Knowledge

From a conceptual perspective, the more interesting powers of the digital surrogate, however, derive from a radical extension of the familiar advantage that books acquire by be-

¹⁶ I put original in quotation marks because the original is often itself a surrogate in a chain of substitutions with no fully definable origin. Consider the First Folio text of *Julius Caesar* or the Venetus A manuscript of the *Iliad* as records of a set of authorial intentions that in my view have a real presence but can never be grasped with complete precision.

coming part of a library. The general characteristics of this advantage are most easily seen with text documents, but they apply with important qualifications to other documents as well. Consideration of this advantage returns us to the central role of the computer as a tool that makes, keeps, and sorts lists. It also draws our attention to another fact. For the digital surrogate to release its full query potential it is not enough to produce a replica of the assembly of real bits that make sense to humans. You must create an equivalent of digital bits that can be processed by computers in such ways that human beings can derive from these equivalents forms of knowledge and insight that cannot be derived from the inspection of the real object. Digitization is ideally a process of highly flexible modeling that allows inquirers to throw arbitrary aspects of the original into sharper relief.¹⁷

A printed page is a set of encodings for a rather slow but exquisitely subtle decoder, the human reader, who brings a set of complex and largely tacit skills to the task of making sense of the frequently underdetermined signs on the page. In particular, human readers have no trouble distinguishing between signs that are ‘part of the text’ and other signs that are not (or not quite) part of the text, such as page numbers, chapter headings, tables of content, title pages, or indexes. The formal and rigid distinction between ‘data’ (a text, a picture, or some other object of attention) and ‘metadata’ or descriptive information about the data is a cardinal aspect of digitization. It is not much of an exaggeration to say that a digital surrogate is only as good as its metadata. One can certainly say that the potential advantage of the digital surrogate consists largely in the quality of its metadata.

‘Data’ and ‘metadata’ lead us into terminology that humanists enjoy detesting. The underlying concepts, however, take us far back in time, long before computers or for that matter, print. The Hellenistic poet Callimachus earned his living as the head of the library at Alexandria. His name is prominently associated with the history of cataloguing, and one of his works is called *Pinakes* (‘boards’ or ‘tablets’). It is an annotated list of the books in the library. The purpose of these ‘metadata’ is apparent. Each item in the list is a grossly reduced representation of a book, but the reduction serves the purposes of establishing an order among the books and articulating the links between them. Callimachus’ list is a very primitive avatar of a modern library catalog, but it takes the critical first step of creating a library as an entity that is more than the sum of its books. In modern technical jargon, the list makes the books at least partially ‘interoperable’. If you prefer a more humanistic rhetoric, it follows the advice of *Howards End*: “Only connect.”

I owe to John Unsworth the formulation of MONK principle, and it is worth pausing over the details of this charming and very informative acronym. In what way do Metadata Offer New Knowledge? Metadata about a single item may help me choose to read this rather than that, but the true power of metadata emerges from the links they enable between items I know about and items I don’t know about. The medieval aura of the acronym also has a point beyond its charm. C.S Lewis writes somewhere that medieval

¹⁷ Creative deformation is a powerful theme in Jerry McGann’s thinking about digital humanities, and he has given it a strongly playful twist in his ‘Patacriticism’ and the literary computer game Ivanhoe. If this does not sound quite serious enough, consider how much insight has been gathered over the ages by ‘playing’ or ‘fooling around’ with data.

monks liked nothing better than making lists. MONK draws attention to the attractions of organized knowledge as a very old human interest.

As lists of summary data about data, metadata have a long history. Generations of scholars have benefited from such lists, whether in the form of bibliographies, dictionaries, concordances, encyclopedias, anthologies, or catalogs of other kinds. If you want to say that scholarship is lists all the way down you are not far off the truth. Digitization, however, extends the power of such lists in ways that turn differences in degree into a difference in kind.

Digitization of libraries began in the sixties with the digitization of catalog records, formal representations of a book that reduce it on average to 0.1% of its size. It was a very difficult task to manipulate millions of such metadata within the very limited bit-storing and bit-moving of the computers of those days. Today, however, there is no reason why metadata should be smaller than the data about which they provide information. They can exceed the size of the original data by orders of magnitude.

Consider the linguists' practice of corpus annotation. The reader who sees strings of letters on the page effortlessly provides the knowledge that turns these strings into meaningful sequences of nouns, verbs, and other parts of speech. Linguists have found it useful to create digital surrogates that add the rudiments of readerly knowledge in the form of metadata about each individual word in the text. The opening sentence of *Emma* in such a digital surrogate begins as follows:

Emma_NP Woodhouse_NP, handsome_ADJ,_PUN clever_ADJ,_PUNC
and_CJ rich_ADJ

This tells the competent reader nothing s/he does not already know but makes the text unreadable unless you strip the suffixes in a display for a human reader. But a text annotated with such very granular and verbose metadata can be processed in a variety of ways to yield information about lexical, thematic, or even syntactic properties of a text or text archive. There are computer programs that can produce sufficiently accurate versions of such annotation at the rate of a million words per hour.

This takes you into a new and strange world where the digital surrogate of the book you read is surrounded by layers of metadata, the purpose of which is to support links between different texts at different levels of organization. Consider the hypothetical, but entirely feasible, example of a digital archive of some thousand English and American novels between *Robinson Crusoe* and *Ulysses*. Each of these novels has metadata at the top level of the document—the kind of information typically found in a bibliographical record. But each of these novels also has metadata at the 'molecular' level of word occurrence. If you now combine the metadata information at the top and bottom levels of a document, every word in every novel can be made to tell a story. Novels are very powerful records of the 'struggles and wishes of an age', even more so in the aggregate than individually. Metadata of this kind let you trace the comings and goings of words across novels quickly and arbitrarily grouped by author, genre, time, or place. Much new

knowledge can be gained by such inquiries. Imagine a bright undergraduate who gets interested in phrases of the type ‘handsome, clever, and rich’ and can within minutes compile the materials for a very interesting honors thesis on the changing fortunes of the three-adjective rule in English fiction.

The study of classical Greek offers a useful example in this regard. The Thesaurus Linguae Graecae (TLG) has over the course of the past four decades created plain vanilla transcriptions of virtually every Greek text from Homer to Nonnus and beyond. The Perseus Project has woven a web of quite sophisticated metadata around the most commonly read texts. Put together, the TLG and Perseus offer a very solid philological platform to any student of Greek anywhere in the world for no money (Perseus) or very little money (the TLG). If you want to contribute to the study of ancient Greek, this platform has substantially shortened the time it takes for an ambitious and talented student to contribute to the discipline.¹⁸

If you wanted to replicate something like Perseus for English and American Letters between Chaucer and World War I, you would have to construct a consistent set of metadata around a corpus of somewhere between one and five billion words that have been already encoded or are scheduled for encoding in a manner that meets reasonable scholarly standards. The addition of such metadata is a large but entirely feasible project, and if you think of it as a Book of English, it pales in scale or difficulty by comparison with the hugely successful cooperative genome projects that have transformed the study of biology. It can be done collaboratively by a group of research libraries each investing relatively modest five-figure sums over a period of five to seven years.

The query potential of digital images

The power of the digital surrogate for scholarly purposes derives from the fact that it can be creatively deformed in ways that generate insights about the original. Textual and visual documents are both similar and different in this regard. You can take pictures of objects from different perspectives or under different lighting conditions. You can manipulate the resultant images in programs like Adobe Photoshop and submit them to various deformations for cognitive or aesthetic purposes. You can see things in a digital surrogate that you could not see in the original object, as in the charred sections of the Beowulf manuscript.

The Mellon ARTstor project uses a very striking technology of image manipulation for exploratory purposes. Images are stored at very high resolution and can be broken into ‘tiles’ that let you zoom in and out of parts of the image without loss of image quality. Thus the computer becomes a very elegant magnifying glass supporting a closer and more flexible examination image than is possible in a book or slide.

¹⁸ I was a student of the distinguished linguist Fred Householder, who was reputed to have learned Greek on the subway to Columbia and liked to say that ‘the first two thousand pages of Greek are a waste’ in the sense that they just get you to the point where you might say something useful about a passage. You can do better than that now, although “Ancient Greek in 21 days” is probably not in sight.

The very striking visual power of image manipulation program may suggest to the casual user that computers can really do a lot more with images than with texts. On the other hand with current technology, images are much less searchable than texts. The causes and implications of this difference are worth some elaboration. While computers derive their power from their ability to resolve everything ultimately into ‘bits’, at an intermediate level pictures and texts decompose quite differently. The computer by and large uses a typewriter model of language in which texts are divided into lines and lines are divided by space or punctuation characters into words. In the world of computers, a text therefore decomposes ‘naturally’ into a self-indexing inventory of its parts from which you can easily learn an astonishing amount about its content.¹⁹

By contrast a computer image decomposes into a set of pixels from which it is much harder to get usable information. There is intensive research on something called QIBC or ‘querying images by content’, but it will be quite a few years before you can ask your computer to go through a list of photographs and find all the pictures with cats in them. The query potential of images is largely determined by verbal metadata that must be added by humans in a very labour-intensive fashion. Nor is there a common controlled vocabulary that works across different collections. Thus the meticulously annotated and deeply searchable images in the Blake Archive use a taxonomy specially developed for that project.

Building a scholarly infrastructure for the humanities: What has been and should be done at Northwestern

It is difficult for the humanities in any university to be better than its library. In the future the quality of a library in the humanities will be increasingly measured by the scholarly potential of its digital infrastructure and in particular by the quality of the laboratory environment it creates for complex analyses of primary archives. The maintenance of such an infrastructure will take place in an environment that will remain unsettled for many years, and it will require different forms of cooperation between scholars, librarians, and programmers across many institutions.

I say ‘unsettled’ because there are not at the moment clearly understood ways through which faculty or academic administrators communicate institutional goals that are then translated into appropriate policies in libraries or IT organizations. Something like that was the *modus operandi* for most of the twentieth century when the documents were print based, the different players had a good understanding of who did what, and emerging digital technologies were limited to inventory management. There was a time not so long ago when scholars wrote, editors edited, publishers published, librarians purchased and catalogued and scholars read and triggered another iteration of this virtuous cycle. But technology has changed every activity in this value chain, not to speak of the relations

¹⁹ This process is called ‘tokenization’, and it underlies many procedures in information retrieval. From a properly humanist perspective, there is something obscenely brutal about ignoring a writer’s intentions and counting his words without any regard to their explicit ordering in a text. But the procedure can be marvelously expressive: the list of Homeric nouns in descending order begins with “man, ship, god.” Could there be a better title for a book about Homer?

between them. It will take quite a while for these activities and their relationships to evolve into new and settled patterns, if they ever do. In the interim we will live in a world where Ranganathan's fifth law has special relevance: the library is a growing organism.

Jaroslav Pelikan, a very distinguished historian of Christianity and former dean of the Yale Graduate School, has written very eloquently on this matter in a paragraph that on the surface says nothing about information technology. He writes about the need to re-think the relationship of 'faculty' and 'staff' in an increasingly interdisciplinary world and argues that

the university library ... must be seen as a collegial part of a total university network of support services for research, and the network must be seen as a free and responsible community if it is to be equal to the complexities that are faced by university-based research. . . . Scholars and scientists in all fields have found that the older configurations of such services, according to which the principal investigator has the questions and the staff person provides answers, are no longer valid, if they ever were; as both the technological expertise and scholarly range necessary for research to grow, it is also for the formulation and refinement of the questions themselves that principal investigators have to turn to "staff", whom it is increasingly necessary -- not as a matter of courtesy, much less a matter of condescension, but as a matter of justice and of accuracy -- to identify instead as colleagues in the research enterprise.²⁰

The work of librarians and programmers could be called a form of scholarly logistics. Scholars are apt to underestimate its importance—a mistake unlikely to be made by generals or Walmart executives. From the scholar's perspective it is tempting to think of oneself as thinking about or doing things with stuff, while others are merely engaged in keeping and fetching it. The limited truth of that distinction does not take you very far. How you think about, or what you can do with, stuff is deeply enabled and constrained by how the stuff is kept and fetched. In a digital world the basic activities of keeping and fetching turn into highly complex systems whose design and maintenance require very high degrees of intelligence, ingenuity, and imagination.

Should we commit faculty resources to the building of digital cultural archives?

It is clear to me that much of the most consequential scholarship in the humanities over the next two decades will consist of the rebuilding or 're-mediation' of a documentary infrastructure in a digital world. It is much less clear to me in what institutional environments this work will be done. If you look for historical analogues, you can point either to the migration from manuscript to print during the first century of print culture, or you can look to the period from the mid-nineteenth to the mid-twentieth century, during which the documentary infrastructure of modern scholarship was built by generations of editors and archivists. Theodor Mommsen's masterminding of the systematic collection of inscrip-

²⁰ *The idea of the university* (New Haven, 1992).

tions from all parts of the Roman Empire is a powerful example of the transformative effect caused by changes in the documentary infrastructure of a discipline.

Work of this kind was once rooted in university departments and libraries. In the American academy, this work began losing its prestige in the sixties. If you were a gifted scholarly editor in the fifties, your chances of getting an assistant professorship in a top twenty English department were quite high. Today they are close to zero. One reason for this is probably the judgment that much of the important work had been done, that the crucial methodological problems had been solved, and that the remaining work was of a journeyman variety that required little inventiveness or imagination.

Changes in information technology have challenged such judgments, whatever validity they may have had. Cultural archives of all kinds need to be rebuilt in a digital environment in such a way as to perpetuate traditional forms of inquiry and to enable new ones. This work requires a lot of dedication and ingenuity. Whether it is done well or poorly has a considerable impact on future scholarship in a discipline.

Where should such work be done and who should take the lead? These are difficult questions, and no widely accepted institutional solution has yet emerged—or is likely to emerge any time soon. Two points are fairly clear.

First, you misunderstand the nature of the work if you think of it as a merely technical task of moving stuff from one medium to another. The work requires cooperation between people who know about the stuff and people who know about the technology. The point of a scholarly digital surrogate is to enhance the query potential of the original in certain ways. For this you must understand the original, have a sense of what questions scholars want to ask now or may want to ask in the future, and what kinds of questions are enabled by creating the digital surrogate this way rather than that way. Good solutions to these problems do not happen by themselves.

Secondly, if a university wants to be in the forefront of humanities research, it must be an active player in the construction of the digital infrastructure for scholarship in its targeted disciplines. If it is not, it will lack critical elements of the expertise that are necessary for research and for the training of graduate students and advanced undergraduate students.

There are two different models for a university to build and maintain that expertise. You can build it into the departmental structure through regular appointments. Alternately, you can build it into the Library or IT organization through research faculty appointments. In either model, the university adds discipline-based researchers who either have high computing skills or know how to work with people who have them.

The two models each have things to recommend them, and they are not mutually exclusive. If you look around the country, there is a fair amount of experimentation with new ways of organizing scholarly labor. A particularly interesting hybrid model is underway at the University of Nebraska, where a humanities technology institute is jointly chaired by the curator of the Rare Books Library and a professor in a humanities department,

with an appointments plan that systematically straddles these different parts of the university.

But no experimentation gets around the fact that you cannot spend the same dollar twice. In the relatively static financial environment of private universities like Northwestern, an appointment made here is an appointment not made there. My own view is that humanities departments would be well advised to recognize the construction of complex digital surrogates as an important part of their curatorial responsibility and to invest some faculty resources in them. There are approximately 130 tenure line faculty in the WCAS humanities departments, about 60 in Literary Studies of various kinds, some three dozen in History, and about a dozen or fewer in each Philosophy, Art History, and Religion. You can make a very strong case that it would be a good idea to redirect between three and five appointments in Literary Studies, somewhere between 5% and 10% of faculty resources, towards the high-tech scholarly curatorial tasks of what used to be called 'lower' criticism. I would certainly take issue with the idea that somehow the humanities have advanced decisively beyond the merely curatorial into a permanently critical or theoretical sphere. That way delusion lies. The work of 'higher' critics over the next three decades will be powerfully enabled and constrained by the 'lower' work of rebuilding the documentary foundations of scholarship and criticism in a digital world.²¹

A tour of some Northwestern projects

Northwestern has made notable contributions to a new documentary infrastructure, and each of these contributions has an interesting story about the increasingly uncertain boundaries between 'scholarly', 'professional', and 'technical' labour. These contributions also point to the growing role of inter-institutional cooperation. It is also worth pointing out that such work has brought significant resources to Northwestern—some two million dollars to Academic Technologies over the past two years, including a recent half-million dollar grant from the Mellon Foundation for work on the Shuilu'an Temple in Western China.

Carl Smith's projects

Carl Smith's *Chicago Fire* was one of the first scholarly web sites at Northwestern to receive national attention. Like its successor, *Dramas of the Haymarket* it grew out of collaborative work with staff at the Chicago Historical Society. The latter project was closely linked to a massive digitization project undertaken by the Library of Congress.

²¹ There is a practical argument for investing faculty resources in this field. If you look in the history of the Department of English for the most successful 'producer' of PhD's with good careers, you will discover quickly that it was Harry Hayford's Newberry-Northwestern edition of Melville. It was a veritable job machine from the late fifties into the early eighties. Its most successful alumnus, and probably the most successful scholarly alumnus of the Department of English in the last generation, has been Thomas Tanselle, who made very significant contributions to editorial theory in the sixties and is now the senior vice president of the Guggenheim Foundation. There is a 'back to the future' component to my argument. I think that technology-driven changes will give strong and positive impulses to collaborative and project-oriented work in environments that have some of the qualities of a good science lab and will produce good employment opportunities inside and outside the academy.

As scholarly works with a marked public dimensions, the *Chicago Fire* and *Dramas of the Haymarket* sites are like museum exhibits with formal published catalogues, forms of scholarship that are in Art History but have no analogue in Literature or History. Smith has always been very eloquent about the deep ways in which what he wrote was affected by the design and technical work of his collaborators.

The Encyclopaedia of Chicago

In more recent years, the print and electronic publication of the *Encyclopaedia of Chicago* has been an effort involving the Chicago Historical Society, the Newberry Library, and Northwestern University. Janice Reiff, one of its general editors and an historian at UCLA, ran Northwestern's microcomputer store at some point in the eighties—a fact not entirely irrelevant to the tangled web of this project's long history. While Northwestern's contribution to this project has been predominantly 'technical', its impact on our local scholarly infrastructure is likely to be substantial. It has played a big role in the 'project-tool cycle' by which the development or improvement of tools and procedures employed in one project help to solidify and extend the infrastructure that enables more complex projects to be built in the future.

More specifically, the Encyclopaedia of Chicago extends the use of FEDORA, a set of rather complicated technical routines that were pioneered at Cornell but whose implementation now rests on collaborative work involving academic technology professionals at Cornell, the University of Virginia, Tufts, Northwestern and others. FEDORA will play a critical role in the development of a university-wide digitally based cultural image service at Northwestern. This is a project that has been on the institutional wish list for well over a decade, but various technical and copyright issues have hampered its implementation. As in many other universities, the art history department's slide collection has been the central repository of cultural images. But the last slide projector ever made by Leitz has either rolled or will shortly roll off the production line in Wetzlar. The age of the slide projector is over.

The Mellon International Dunhuang Archive

The most significant digital project to come out of Northwestern is probably *The Mellon International Dunhuang Archive*, a multi-year, multi-national, and multi-million project to create digital surrogates of Buddhist cave shrines in Dunhuang, located on the Silk Route in the Gobi Desert.²² The project was funded by the Mellon Foundation and is now part of Mellon's ARTstor collection. For several years it preoccupied the work of the art historian Sarah Fraser, the photographer Harlan Wallach, and the programmer Bill Parod, as well as other staff of Aca-

²² This project is a spectacular demonstration of the query potential of the digital surrogate. First of all, the Dunhuang Caves are far away. If you get there, the caves are dark. If you bring a good enough flash light, you will find that many of the images are so high on the wall that you cannot see them very well. The web site, on the other hand, uses virtual reality display and the zoom-in technology of redrawn images to let you examine the paintings and sculptures from many angles and at arbitrary levels of magnification.

demic Technologies and the MultiMedia Learning Center. In scope, complexity, or intrinsic interest this project yields little to the signature projects of the University of Virginia, whether the Valley of the Shadow or the Blake and Rossetti archives. It is regrettable that for a variety of innocent circumstances the deep association of this splendid project with Northwestern has never quite captured the publication imagination.

Oyez

At the border of the humanities and social sciences, and also at the border of 'cultural heritage' and contemporary projects lies *Oyez*, Jerry Goldman's Supreme Court Web site, and its sibling History and Politics Out Loud (HPOL). These are probably the most widely accessed Northwestern sites with a scholarly core. From a technical perspective, their most distinctive contribution has been the use of audio files and the development of sophisticated procedures for coordinating audio files with written transcripts. Good and user-friendly techniques for doing this have a wide variety of uses in many disciplines, including linguistics or the study of music.

The Vesalius Project

The Vesalius Project, developed by Dan Garrison in Classics and Malcolm in the Medical School and published by the Library, invites reflections on the history of technology. Andreas Vesalius' *On the Fabric of the Human body* (1543) is not only the first anatomy book, but it is a striking demonstration of an often overlooked aspect of the print revolution. In addition to making it faster and cheaper to disseminate written text, print technology enabled a much more extensive and complex coordination of graphic and textual materials than had been possible in manuscript culture. What is often thought as a distinctive feature of Web, the ability to bring text and image together, is more properly a constitutive feature of print culture. Vesalius had elaborate ways of explicitly linking parts of this text to particular regions in his illustrations. The digital surrogate mimics these links but adds to its query potential by letting the reader manipulate and magnify the images in various ways.

Other archival projects

Several other projects deserve a fuller discussion than the cursory acknowledgment I give them here. They include

- digitization of Edward Curtis' North American Indian, developed at Northwestern and now included in the American Memory Project
- The Siege and Commune of Paris, which digitizes Northwestern's extensive collection of photographic images from that period of French history
- The MMLC project about the Picpus Cemetery, the grave of many victims of the Reign of Terror in 1794
- The Winterton project, a largish project, still in an early phase, which will digitize a superb collection of some 6,500 photographs taken in East Africa between

1860 and 1960. This project will be FEDORA based and is likely to play a role in defining the scholarly requirements of a digital image repository.

Project Pad

Project Pad is a tools oriented project with considerable potential for scholarly and pedagogical activities. It has been under development for several years and is the work of Jonathan Smith and a small team working with him.²³ It began as a project for annotating the Web and has developed very elegant procedures for annotating images or time-coded media, such as audio or video files. With ProjectPad I can attach a note to the mouth of the Mona Lisa or to bars 17-19 in Tosca's "Vissi d'arte."²⁴ I can keep this note to myself or share it with you. If I do, the existence of annotation in a given place is made visible to whoever cares to see it. This tool is in the vanguard of collaborative annotation software, an area that is receiving much attention.

WordHoard

WordHoard is a Mellon funded project that is now drawing to a close. It surrounds a limited set of highly canonical literary texts (Homer, Chaucer, Spenser, Shakespeare) with complex layers of lexical, morphological, prosodic, narratological, and semantic metadata. It provides analytical tools for exploring these metadata. In particular it includes statistical tools that are commonly used in natural language processing (NLP). WordHoard probably has richer metadata than any other text analysis program, with the possible exception of some Bible software. The project leaders for WordHoard have been Martin Mueller and Bill Parod. Much of the software has been designed and written by Phil Burns and John Norstad, with important text processing contributions from Jeff Cousens.

Virtual Orthographic Standardization of Early Modern English texts

The CIC libraries have funded and sponsored under the CLI umbrella a project to provide virtual orthographic modernization and part-of-speech tagging for the growing archive of fully transcribed early modern English texts in the Text Creation Partnership project (TCP). This is a joint project by the Library and Academic Technologies with Martin Mueller as the initial project leader. This project grew out of series of conversations among librarians and programmers at Michigan, the University of Chicago, and Northwestern. It is a project with some growth potential. If fully implemented it will create orthographic, morphosyntactic, and lexical metadata for several billion words of English written between 1470 and 1900. Prototypes of virtual orthographic modernization are available at <http://panini.northwestern.edu/philologic/shakespearesources.html> and <http://www.lib.uchicago.edu/efts/EEBO/search.html>.

²³ Project Pad is a Web based collaboration system that works with FEDORA repository objects and is now available as a Sakai-compliant tool. The Project Pad server is written in Java and the Project Pad client is expressed by a Flash ActionScript engine. Project Pad 2.0 will be released in January 2006 as an open source (GPL) application at <http://projectpad.northwestern.edu>.

²⁴ The underlying program knows nothing about the Mona Lisa or bars 17-19, but depends on identifying a pixel range or time code.

The people behind the projects

This tour of recent and current projects at Northwestern suggests that the scope and pace of humanities-related digital projects have increased. Externally funded projects have helped, but more internal resources from Academic Technologies have been directed to projects that directly or indirectly strengthen the infrastructure for scholarly work in the humanities.

Virtually every project in humanities computing at Northwestern has benefited from the skills and ingenuity of Bill Parod. He is the only programmer whose work over the past decade has predominantly focused on humanities projects. He designed the data architecture for the Dunhuang project and The Encyclopaedia of Chicago History. He designed the architecture and wrote most of the software for the Homer and Shakespeare projects that have now been subsumed under WordHoard. The interface of the Vesalius Project is his work, and at different times he has played an important role in Oyez. As important as any of these particular achievements is his integration of different tools and approaches into an informal but fairly sturdy platform of tools and procedures for tackling the next project whatever it is.

Harlan Wallach was a crucial member of the MMLC team for many years before moving to Academic Technologies, where he is the leader of media based projects. He is at heart a photographer and has an extremely good eye for functional and aesthetically pleasing design. The quality of his work is apparent in every image of the Mellon Dunhuang Archive.

Jonathan Smith came to Academic Technologies after quite a few years with Roger Schank's Institute for Learning Sciences and its commercial successor Cognitive Arts. Much of his work has revolved around Project Pad and Oyez, and in these areas he has worked closely with Chris Karr, a very gifted young programmer and student of the legendary Brian Kernighan. Smith's and Karr's contributions to humanities computing lie below discipline specific projects, but their work has considerable impact on the strength and complexity generic platforms on which particular projects are based.

Phil Burns' and John Norstad's careers at Northwestern reach back to the glory days when the University funded its computing centre with royalties from the operating software it had developed for Control Data mainframe computers. Both of them are well known among Windows or Macintosh developers for particular programs. Until recently both of them spent most of their time on very complicated systems and network infrastructure projects that users only notice when things go wrong.

Norstad is a programmer with a remarkable grasp of interface issues and very deep skills in typographical layout, qualities apparent from a very cursory look at WordHoard. He writes better documentation than anybody else I know.

Burns is a statistician who wrote a number of modules for SPSS. He has a very deep understanding of the application of quantitative procedures to natural language processing. This is likely to become a more field in text processing, sometimes in explicit ways, but

just as often in ways that are not apparent to end users who might well be surprised to learn that the result of a particular query rested on a statistical procedure.

Jeff Cousens is a young programmer who over the past couple has moved from systems administration into project related work. He has done excellent work in various areas of text processing.

If you go outside Academic Technologies, Matt Taylor, the chief programmer of the MMLC is a very versatile developer who prefers finding his solution for your problems as opposed to finding your problem for his solution. This is unusual for programmers.

Academic technologies staff have worked increasingly closely with library staff over the past five years. The decision to move Academic Technologies and the collection management of the Library into adjacent has worked even better than people expected at the time. Stu Baker , the head of Library Management Systems, and Claire Stewart, the head of the Digital Media Group, are two key players in the continuing conversation between librarians and programmers. Their interest and attention has contributed much to the success of a number of projects.

There is an anecdote from the early days of Museum Science when the curators in one of the great Berlin museums wondered about how to attract attention to their treasures. One of them said: ‘Well, we can always move the stuff across the road and call it an exhibit.’ If we took the work of the MMLC, Academic Technologies, and the Library in the field of humanities computing and bundled past achievements, current projects, and the people working on them under the ‘special exhibit’ category of a center or institute, we would look pretty good. Not very many universities in the country at this moment have a group of technology professionals that work on humanities projects and can match the knowledge and skills set collectively represented by Burns, Cousens, Karr, Norstad, Parod, Smith, Taylor, and Wallach. This is a very impressive group of people with a deep understanding of computer languages and procedures that govern

- the logistical frameworks that undergird projects of any complexity
- general interface design
- manipulation and display of images
- text processing and searching
- typography and layout in a multilingual environment
- statistical procedures for document analysis

What should be done?

We will do very well indeed if we can maintain an environment in which the very talented and diverse group of programmers that are currently involved in humanities related projects can continue to use a significant portion of their talents and energy on projects that directly or indirectly bear on strengthening the scholarly infrastructure in the humanities. An important ingredient of success will consist in the ability to abstract from the requirements of particular projects, to look at underlying functionalities and relationships,

and to maintain a virtuous project cycle in which the next project is built as much as possible on the platform of the previous and helps to strengthen or extend it.

ERANOS: collaborative tools for querying and annotating digital archives

In thinking about Enhanced Access to Digital Archives critical to remember that the ‘access’ given to the scholar is a function of the structure of the data, the tools for getting at them, and the relationship between data structure and tools. I have said a fair about data structures in the discussion of metadata and the MONK principle. Let me turn for a little while to collaborative query and annotation tools, which I will bundle under the name ERANOS. That is an ancient Greek word for potluck. Whatever its etymology, one may conveniently relate it to various Greek ‘er’ roots that define related semantic fields of desiring, asking, and speaking.²⁵

Exponential growth in size is a characteristic feature of digital archives. Even simple query tools let scholars search across large document archives of texts or images. Such tools extend control over large archive and make it possible for scholars to identify documents to read or look at closely. Extended control can also come from collaborative software or ‘groupware’, which makes it easy for groups of researchers to share queries and results, whether in a pedagogical environment like a seminar or in a scholarly activity like an editorial. It is probably the case that some inquires into very large archives will benefit from explicitly collaborative patterns of organization.²⁶

If the whole thing is an ERANOS, its individual steps break down as follows:

1. A group of people come together for the purpose of studying something
2. The something may be given to them or they may have to create, prepare, and transform it into a sharable archive
3. They ask questions about the things in their archive and draw connections between them
4. They say and write things about these things and relationships
5. They share what they say and write with each other and perhaps with others as well

Consider the simple example of an American Studies seminar that looks at history through family pictures. The requirement for the seminar is that every student bring three dozen pictures spanning three generations. In this situation, the students will spend much time on Step 2, transforming a set of pictures into digital images with the kinds of metadata that will add up to a searchable archive. This differs from a seminar in which participants begin their work with an archive that is rich in metadata, such as texts mediated through WordHoard.

²⁵ It is an ancient philologic habit to make convenience rather than truth the principle of etymological explanation.

²⁶ It is often said, and with some justice, that compared with the sciences, scholarship in the humanities is isolated and individualistic. But if you stand back a little and analyze the web of citations, acknowledgements, and rebuttals that make up the bulk of scholarship in any field, the humanist, too, plays a competitive and cooperative game.

Step 2, which concerns fundamentals of scholarly logistics, has taken much of Bill Parod's attention in recent years. Working with colleagues from Tufts and the University of Virginia in the FEDORA environment he has worked on routines that ultimately govern how users access, fetch, store, or manipulate images (including page images of texts) or digitized texts. This work continues and has recently included representatives from the Digital Library Federation (DLF) Aquifer project. The goal is deep access interoperability across and repositories.

Steps 3 -5 have been key concerns of WordHoard development. Steps 4 and 5 have been the focus of Project Pad. Outside of Northwestern, Jerry McGann's COLLEX software may be the fullest realization of a scholarly ERANOS. The name is a portmanteau of 'collect' and 'exhibit'. COLLEX lets users "collect, tag, analyze, and annotate trusted objects (digital texts and images vetted for scholarly integrity); reorganize and publish objects in fresh critical perspectives; share these new collections with students and colleagues, in a variety of output formats; and, without any special technical training, produce interlinked online and print exhibits using a set of professional design templates."²⁷

Annotation software

Collaborative work on digital archives is an eranos or, if you prefer a more American metaphor, a form of scholarly barn raising. It consists of the acts of building, exploring, and discussing archives. Discussion often takes the form of annotation, which is one of John Unsworth's scholarly primitives.²⁸ There was much discussion of it at as Digital Tools Summit sponsored by the Institute for Advanced Technology in the Humanities at the University of Virginia in September 2005. Given the importance of groupware and content managements systems (CMS) in many walks of life it is noteworthy that collaborative annotation programs have not yet become established software tools.

The discussions of annotation software at that conference ended with the following preliminary specifications for a 'highlighters tool':

- works with text, image, time-dependent media, annotated resource, or nothing, online or offline
- allows you to demarcate what's of interest (visually/by hand, according to a user-specified rule, or according to an existing markup language specified in some standard grammar)

²⁷ An excellent description of this Mellon funded initiative is found in Bethany Nowviskie, "COLLEX: semantic collections & exhibits for the remixable web," <http://www.nines.org/about/Nowviskie-Collex.pdf>.

²⁸ "Scholarly primitives: what methods do humanities researchers have in common, and how might our tools reflect this?" Part of a symposium on "Humanities Computing: formal methods, experimental practice", sponsored by King's College, London, Mary 13, 2000. <http://www.iath.virginia.edu/~jmu2m/Kings.5-00/primitives.html>.

- allows you to type the highlight (according to your own evolving organizational structure, or according to an existing ontology/taxonomy specified in some standard grammar)
- allows you to cluster, merge, link, overlap the demarcated parts
- allow you to attach annotation to the demarcated parts
- allow you to attach annotations to clusters, links
- allow you to search at least your annotations
- output to this needs to be available as input to this
- output needs to be exportable in some standard grammar
- allow you to do these things in arbitrary order
- allow you to zoom to arbitrary levels of detail

In addition, annotation software needs to be integrated into a permission regime that governs who may add, edit, approve, remove, and see annotation.

In thinking about digital annotation in a scholarly environment, it is best to start from conventions of scholarly writing in the print world. The classic unit of annotation is the footnote. The classic genre of annotation is the scholarly commentary, which is little more than a sequence of footnotes. Digital annotation in a scholarly environment must support at least what can be done in the print world, but it should also aim at leveraging three distinctive qualities of the digital medium:

1. The footnote and the commentary are space constrained. In the digital world the most important constraint is increasingly Ranganathan's fourth law.
2. The footnote must be attached to a single place in a document, though it may be cross-referenced. The digital note in principle has multiple points of attachment or retrieval.
3. Digital annotation can be done collaboratively in a distributed environment.

There is an ambiguity in the term 'annotation' that is worth spelling out. It turns on the oppositions of private/public and process/product. Annotation may be private—a note to myself—and it may be process-oriented—the 'brainstorming' of people working together on a project. But annotation may also be public and product oriented—a considered and formal way of saying something about something. Both forms of annotation must be supported in a digital ERANOS.

There are two aspects of annotation that programmers identify as difficult. If the annotation attaches to a unique part of a digital object—a line in *Hamlet*, a part of a picture, a sequence of notes in music—it is relatively trivial to answer the question: Where do I put and how do I display the annotation? But if the annotation is about something that has multiple occurrences or is not localized at all the question becomes much harder to answer. A particular word in Shakespeare ('incorporate'), a repeated phrase or grammatical condition in Homer ('wine dark sea', aorist optatives) may well be captured as distinct entities what programmers call the 'object model' of a given digital project. But it is harder to figure out how and

where to attach the annotation. It is even more difficult to find good ways of making its existence visible in the right places both unambiguously and unobtrusively.

Secondly, if the annotation is a simple gloss or short paragraph it is easy to provide editing tools for writing annotations. But if the annotation has a structure of its own, includes multiple paragraphs, footnotes, cross-references or links to other documents, it is quite a challenge to create an environment in which different contributors do the same thing in sufficiently similar ways for a body of annotations to be managed and displayed in a reasonably coherent fashion. On the other hand, the scholarly genre of the commentary makes use of all these features, and it is worth the programmer's while to find digital solutions that mimic and exceed the capabilities of the print commentary.

MONK and ERANOS

MONK and ERANOS are the Janus face of Enhanced Access to Digital Archives : metadata-rich archives and collaborative query and annotation tools to explore and report on them. My central recommendation in this report is that an investment in the humanities should involve the vigorous pursuit of initiatives in both directions:

Recommendation #4: Building on their various and quite considerable achievements in various humanities related digital projects, the different players in these projects, notably Academic Technologies, the University Library, and the Multimedia Learning Center should abstract common elements from particular projects and use them as guides for further development. A possible list of such abstractions might include:

1. A general 'scholarly logistics' platform (in practice FEDORA) that governs the ways in which documents or 'digital objects' in various media and at various levels of granularity are kept in 'digital repositories' and that supports the fetching or linking-to of such digital objects across different collections and institutions.
2. A campus-wide cultural image service that supports the retrieval and manipulation of large archives of digital images likely to be of interest to research and teaching in the humanities.
3. The gradual transformation of current digital collections of English and American letters from the late Middle Ages to the early twentieth-century into a seamless archive surrounded by layers of metadata that support sophisticated exploration of textual data. This work should build on current collaborative links with UIUC, the University of Michigan and the University of Chicago.
4. The development of collaborative annotation software that supports both scholarly projects and pedagogical projects in seminar environments for graduate students and advanced undergraduates. Such work should build on ProjectPad, the annotation module of WordHoard, and should seek collaboration with annotation projects elsewhere, in particular McGann's COLLEX project at the University of Virginia.

Several ancillary recommendations flow from this central one:

Recommendation #5: The University should do some internal and external marketing of the considerable achievements of its various digital humanities projects. It should consider joining some important consortia, such as the

Digital Library Federation, and its programmers and project leaders should be encouraged to display their wares at the leading international ‘fairs’ of humanities computing.²⁹

Recommendation # 6: Building on the excellence of past projects and the very deep talent pool of programmers currently involved in digital humanities projects, there should be a coordinated effort to seek outside funding for a set of interlocking projects that pursue the priorities outlined in Recommendation #4.

Recommendation # 7: In the context of general investment in the humanities, the Central Administration should provide seed money of approximately \$300,000 over three years for the initial development of projects specified in Recommendation #6, on the assumption that Academic Technologies and the Library will between them redirect significant resources to the pursuit of these projects. Seed money of this magnitude presents an equity investment that will substantially improve the chances of winning major grants.

Educating Users

For the past few years, in a cooperative venture between WCAS and the Library, Ruth Reingold and Harriet Lightman have run very successful day-long digital orientation sessions for incoming graduate students in the humanities and social sciences. It is fairly clear from these sessions that few graduate students are sophisticated users of technology. They are not as a rule very good at spotting or exploiting the full analytical potential of digital archives. Nor can it be assumed that the relevant skills teach themselves. It may even be the case that modern office tools, by making simple routines very simple, make it harder for most people to ramp up gradually and insensibly to more sophisticated uses but encourage them instead to stay at the level that allows them to get by.

If you talk with librarians about the bibliographical skills of their patrons they will tell you that many users at all levels, from freshman to full professor, are not very good at navigating information resources. They will also tell you that digitally based tools are very complex. They do more, they cost more, and once you go beyond simple routines, they take longer to master than card catalogues and print bibliographies.

Some people think that this is a problem of time to be solved as generations of “digital natives” arrive. If that were the case, freshmen and sophomores should on average be more sophisticated than they are. A few of them are very good indeed, but on average they do not differ a lot from older users. Like them, they can do the few things that are

²⁹ Two thirds of the top twenty universities in the country and more than half of the CIC institutions, including UIUC, the University of Chicago, and the University of Michigan are members of the Digital Library Federation, which is emerging as a standards setting body for many of the issues addressed in this report.

necessary to get by, but they are largely clueless about how to take full advantage of digital tools and resources.

There is a problem here, and as is often the case, recognition and attention are the first and most important steps towards a solution. The most discouraging aspect of the digital orientation sessions for graduate students, which have been thoughtfully designed and meticulously executed events, has been the lack of buy-in by faculty who should take an interest. I'm not sure I ever saw a director of graduate studies at any of them. I certainly did not see a handful of them—and neither did the students who were asked to attend these events.

The goal of user education should not be to turn all humanists into hackers who are wizards at Ruby on Rails. But we should move towards an environment in which scholarly users in the humanities on average have a better calculus of the possible than they currently have, think about the ways in which digital tools and resources can help them, have a sense of what it would take to get them from here to there, and know where to go for help.

There should probably be follow-up events to the current orientation sessions for graduate students, and there is no reason why such events could not include faculty or motivated undergraduates as their target audience. A number of universities have been quite successful with workshops or short courses that cover a 'curriculum' of scholarly technology practices from an introductory to an advanced level. General techniques of text processing, image manipulation, and data management typically make up the bulk of such activities. There is much better open source or cheap software available for such routines than was the case five years ago.³⁰

Developments in videoconferencing and other online technologies make it quite possible to think of such follow-up activities as being based partly on talent elsewhere. There are a number of people at other universities who have strong reputations for their pedagogical skills in these fields.

In the early days of the Web Northwestern ran a very successful training program for faculty (TiLt). It was labor intensive but much appreciated by the people who took it. It may be worth asking whether some version of that program deserves to be resurrected.

Easily available help with small-scale projects should be another building block of a user education program. Sarah Fraser has argued eloquently for such an approach. If you have a modest project in mind, whether pedagogical or scholarly, and this project would benefit from a programmer's assistance for hours or days you might well do it if you can get help quickly. You are probably not going to do it if you have to wait weeks or months for the help. It would be a challenge to develop and manage such a program of quickly

³⁰ The oxygen XML editor may in fact be called a breakthrough in this regard. I have heard from several that this very transparent and relatively user-friendly has made it much easier to get novice users to a fairly sophisticated grasp of the basics of text technology within a time frame of days.

available help. For it to work economically and fast it would have to rest on firm understandings of supported software tools and protocols. But it would be worth doing.

Recommendation #8: User education should be identified as an important issue requiring continuing attention from the different players in scholarly technology and particularly more active support and encouragement from faculty and academic administrators. A coordinated set of activities in this area should:

1. Continue and strengthen the digital orientation session for incoming graduate students
2. Develop follow-up activities consisting of workshops or short-term courses that add up to a curriculum of basic scholarly technologies
3. Provide quickly available technological support for small scholarly or pedagogical projects

Some final reflections on space in the library

The library is the laboratory of the humanities. I began this report with that observation, much of my analysis has concerned itself with the implications of looking at primary archives as if they were the scientist's objects of experimentation, and I return to it at the end.

Northwestern's Library is a particularly striking architectural example of something you can observe on many university campuses. There is an early twentieth century building of a certain size—typically with architectural ambitions that link it to an imagined past. There is a building from the second half of the twentieth century that makes no such claims and is bigger almost by an order of magnitude. At Indiana and the University of Toronto, the new libraries are in different locations and make the old place look like cottages. At Northwestern, the new library is an 'annex' that dwarfs the original, which continues to say with much eloquence, "I may be smaller, but I'm much prettier than you are."

As soon as all those mid-century libraries were built they began to run out of space, and it became clear that there was no way of realizing the dream of keeping all or most books within the confines of one building. And computers have added new twists to the question "where is the stuff I need?"

The design of the Northwestern library was never a particularly good solution to the problem of storing books: shoebox layouts work better for that. But its modular design may lend itself to repurposing in a digital age. Let's face it: the real estate of the nine stack floors in the very heart of the campus is far too valuable to serve as a warehouse for many books that are rarely if ever used. It is far better to think of the library as a people space where individuals or groups find the information, tools, and support they need for their work. Apply to the Library as a whole the kind of redesign that has turned the old cataloging space into an information commons.

The Library has nine stack floors holding, I believe, some three million volumes. I would redefine the stack function of the Library and argue that it should hold a maximum of 1.5-2 million volumes most likely to be used by the Northwestern research community.³¹ The stack floors of the Library would then become an interlocking set of “at hand” collections on the model of European seminar libraries. These collections would be limited to a maximum size and would be subject to a constant process of weeding and feeding.

About a third of the current stack space might then become available for people-centered activities. Or to put it differently, on each floor the researchers in a discipline would find themselves in an environment where information sources and tools of different kinds or media would be available ‘seamlessly’ and in a manner most suitable to the needs of a particular community of scholars at a particular moment in time. A stack floor would be somewhere between a seminar library and a laboratory. As Ranganathan said, “A library is a growing organism”

A ‘back to the future’ model for the third floor of the South Tower in the Library

Applying such principles to the humanities sections of the library opens up some intriguing choices. I will focus on the third floor of the South Tower, which is adjoined to the Ver Steeg Lounge, a handsome room of uncertain purpose and wretched (but fixable) acoustic properties. Think of that floor as both museum and laboratory.

The museum function would be served by creating in that space some representation of a cardinal event in the history of Northwestern’s library, the acquisition of some 20,000 volumes collected by Johannes Schulze, a Prussian contemporary of Hegel’s. Some of the distinctive strengths of Northwestern’s library resources in the humanities derive directly from this collection.³² If you were to create some version of that collection in 3S you would not only commemorate a signal event in the university’s history but at the same time give modern scholars a very useful overview of learning in the midnineteenth century.

The laboratory function would be built by moving into 3S the group of developers whose careers have revolved around humanities computing projects. Some of the needed space would be of a back office kind, but other space would be designed for interaction with users and for the educational activities envisaged in Recommendation #9.

I offer this not as a detailed plan but as a stimulus for thinking along certain lines. Humanists will work for many years to come in technologically distinct environments.

³¹ How do you define the “books most likely to be used?” It would be a trivial but quite informative exercise to keep a frequency-weighted count of all bibliographical references in dissertations by Northwestern students. Map those references to catalogue areas, combine this information with other sources, continue the exercise on three-year cycles, and you have a pretty good basis for spotting current trends and occasionally anticipating new ones.

³² It is called the Greenleaf collection to commemorate Luther Greenleaf, a nineteenth-century real estate developer, who provided the funds.

Space planning should make it as easy as possible to move from one to the other and to be aware of the layered technological structure of the humanist's work.

Appendix: List of Recommendations

1. The quality, maintenance, and strengthening of an increasingly digital scholarly infrastructure should be a continuing topic of high priority in any discussion among faculty and academic administrators about priorities and investment in the humanities. Chairs in the humanities should give to this matter the same attention that chairs in the sciences give to laboratories and equipment.
2. It is useful to distinguish fairly sharply in the humanities between primary documents and secondary literature. Secondary literature—the books and articles of scholarship—is subject to generic library routines that are fairly constant across disciplines. The maintenance of primary documents should be envisaged as taking place in a laboratory like environment, where large document archives, enriched by metadata in discipline-specific ways, can be manipulated with the help of sophisticated tools.
3. The maintenance of such a laboratory environment, which provides enhanced access to digital archives, should be the joint responsibility of the Library and of Academic Technologies
4. Building on their various and quite considerable achievements in various humanities related digital projects, the different players in these projects, notably Academic Technologies, the University Library, and the Multimedia Learning Center should abstract common elements from particular projects and use them as guides for further development. A possible list of such abstractions might include:
 - a. A general 'scholarly logistics' platform (in practice FEDORA) that governs the ways in which documents or 'digital objects' in various media and at various levels of granularity are kept in 'digital repositories' and that supports the fetching or linking-to of such digital objects across different collections and institutions.
 - b. A campus-wide cultural image service that supports the retrieval and manipulation of large archives of digital images likely to be of interest to research and teaching in the humanities
 - c. The gradual transformation of current digital collections of English and American letters from the late Middle Ages to the early twentieth-century into a seamless archive surrounded by layers of metadata that support sophisticated exploration of textual data. This work should build on current collaborative links with UIUC, the University of Michigan and the University of Chicago
 - d. The development of collaborative annotation software that supports both scholarly projects and pedagogical projects in seminar environments for graduate students and advanced undergraduates. Such work should build on ProjectPad, the annotation module of WordHoard, and should seek

collaboration with annotation projects elsewhere, in particular McGann's NINES project at the University of Virginia.

5. The University should do some internal and external marketing of the considerable achievements of its various digital humanities projects. It should consider joining some important consortia, such as the Digital Library Federation, and its programmers and project leaders should be encouraged to display their wares at the leading international 'fairs' of humanities computing.
6. Building on the excellence of past projects and the very deep talent pool of programmers currently involved in digital humanities projects, there should be a coordinated effort to seek outside funding for a set of interlocking projects that pursue the priorities outlined in Recommendation #4.
7. In the context of general investment in the humanities, the Central Administration should provide seed money of approximately \$300,000 over three years for the initial development of projects specified in Recommendation #6, on the assumption that Academic Technologies and the Library will between them redirect equivalent resources to the pursuit of these projects.
8. User education in scholarly technologies should be identified as an important issue requiring continuing attention from the different players and in particular requiring more active support and encouragement from faculty and academic administrators. A coordinated set of activities in this area should:
 - a. Continue and strengthen the digital orientation session for incoming graduate students
 - b. Develop follow-up activities consisting of workshops or short-term courses that add up to a curriculum of basic scholarly technologies
 - c. Provide quickly available technological support for small scholarly or pedagogical projects