Analyzing scholarly narratives

I spent several hours this afternoon refamiliarizing myself with Bamboo, and more particularly, its opportunities for the field I am particularly interested in, text-centric scholarship especially on documents from earlier ages that are in the principle in the domain. In what follows, I'll look at the sixteen or so scholarly narratives that fall in this category -- not quite half of all the narratives.

I do, however, want to start with some reflection about my difficulties in connecting these scholarly narratives to the planning process. For the most part, I understand the scholarly narratives and their relations to each other. Most would benefit from more resources, some would benefit from more attention to existing standards. Some projects would benefit from looking closely at each other and perhaps merging. What is not clear to me is how the Bamboo superstructure or infrastructure will help.

If it is a matter of making technical decision about large pieces of infrastructure relating to data storage or network traffic, industry will make those decision or narrow down the choices. Ditto for applications: there are, for instance, two or three relational databases that account for 95% or more of work that uses this technology. It's unclear, then, what needs regulating or coordinating with regard to those choices. Most of the stuff out there is written in six or fewer programming languages, and there may not be much virtue in reducing that limited diversity. Michael Pollan writes very eloquently about the monoculture of corn.

If it is a matter of institutional cooperation, much of it happens, but not nearly enough of it and often not in the right manner either. How to stimulate or guide it in the right direction is a big question. I am not, however, persuaded that a regulatory regime, however communitarian in its rhetoric -- and there is a lot of that in Bamboo -- will help. To put it bluntly, I have not been able to look at any particular project from the perspective of Bamboo planning and recognize how Bamboo would help. I see a lot of overhead, and I worry whether it will distract rather than support.

So I have put down as honestly and politely as I can my sense that the more Bamboo stuff I read the less I know what it is about and the more skeptical I am about its power to do much good.

I turn from this confession to a look from the perspective of the Tools and Content workgroup at the scholarly narratives that concern themselves with the study of primary texts. Language here is the great divider. The work of very few PhD students in English, French, German, Spanish, or Russian will be affected by work on digitizing Tibetan Buddhist manuscripts or inscriptions from the Greek or Semitic parts of the Ancient Mediterranean. Can one talk about tools and content across different languages?

This is a moment to nod gratefully in the direction of Microsoft and Apple. A decade ago multilinguality was still a big headache. Today your standard computer will move effortlessly across many languages both ancient and modern. The remaining bottleneck is the continuing ability of the modal American programmer to be surprised that there are people in the world whose native language is not English.

It is also a moment to acknowledge the extraordinary success of the TEI Consortium at creating standards that cut across languages and millennia. There is a global and tightly knit community there, and a question that arises in the encoding of a French medieval manuscript may find its answer in a practice developed in the encoding of Buddhist manuscripts in Kyoto. To my mind, the TEI community is a remarkable example of a scholarly group that has global breadth, temporal depth, and is united in a common purpose to use technology to help with philological problems of long standing.

I was particularly intrigued by Scholarly Narrative 0051, the construction of a fully digitized corpus of Tibetan Buddhist writings originally encoded on wooden blocks. What value do you add if you can take all or most of the writings of a culture and transform them into a digital corpus that not only emulates the sequence of glyphs in the original but adds bibliographical, lexical, semantic, morphological, and syntactic metadata so that the original body of writings in its digital medium becomes an enhanced surrogate with affordances for inquiry that far exceed the original?

This project takes you through the stages of this process from the first step of digital images through manual transcription (too expensive), optical character recognition (accurate enough?) to algorithmically applied annotation. A wonderful project, although I suspect that in the end its execution will require more editorial human intervention at all stages of the process.

It is useful to remember that the 'thousands' of Tibetan books add up to a quite small collection of highly curated data. Tens (but not hundreds?) of millions of words, but they will all fit comfortably on a laptop, metadata included.

Digitization is in many ways easier with dead cultures where you have a limited number of documents, and the project of digitizing all or many of them is a manageable task in terms of scale -- never mind the scholarly labor that classicists, Assyriologists, or Anglo-Saxonists don't mind lavishing on the objects of their research. The Perseus project, which includes just about all major texts of ancient Greek from Homer to the Second Sophistic, is just five million words. The Thesaurus Linguae Graecae (TLG), which includes most written Greek from Homer to late antiquity, adds up to ~100 million words but will fit comfortably on the flash drive of a digital camera.

Questions of scale are raised in Gregory Crane's 'memography' (Scholarly narrative 0033), where I find myself in a very familiar disciplinary terrain of scale analysis. Trace Plato through myriads of pages of European writing, perhaps the 50 billion words printed annually in American newspapers. For me it is a moment to nod again in the direction of Microsoft and Apple. The Thesaurus Linguae Graecae (TLG) is a very big project, although I suspect that it will require more editorial human intervention at all stages of the process.

It is useful to remember that the 'thousands' of Tibetan books add up to a quite small collection of highly curated data. Tens (but not hundreds?) of millions of words, but they will all fit comfortably on a laptop, metadata included.

Digitization is in many ways easier with dead cultures where you have a limited number of documents, and the project of digitizing all or many of them is a manageable task in terms of scale -- never mind the scholarly labor that classicists, Assyriologists, or Anglo-Saxonists don't mind lavishing on the objects of their research. The Perseus project, which includes just about all major texts of ancient Greek from Homer to the Second Sophistic, is just five million words. The Thesaurus Linguae Graecae (TLG), which includes most written Greek from Homer to late antiquity, adds up to ~100 million words but will fit comfortably on the flash drive of a digital camera.

Questions of scale are raised in Gregory Crane's 'memography' (Scholarly narrative 0033), where I find myself in a very familiar disciplinary terrain of scale analysis. Trace Plato through myriads of pages of European writing, perhaps the 50 billion words printed annually in American newspapers. For me it is a moment to nod again in the direction of Microsoft and Apple. The Thesaurus Linguae Graecae (TLG) is a very big project, although I suspect that it will require more editorial human intervention at all stages of the process.

It is useful to remember that the 'thousands' of Tibetan books add up to a quite small collection of highly curated data. Tens (but not hundreds?) of millions of words, but they will all fit comfortably on a laptop, metadata included.

Digitization is in many ways easier with dead cultures where you have a limited number of documents, and the project of digitizing all or many of them is a manageable task in terms of scale -- never mind the scholarly labor that classicists, Assyriologists, or Anglo-Saxonists don't mind lavishing on the objects of their research. The Perseus project, which includes just about all major texts of ancient Greek from Homer to the Second Sophistic, is just five million words. The Thesaurus Linguae Graecae (TLG), which includes most written Greek from Homer to late antiquity, adds up to ~100 million words but will fit comfortably on the flash drive of a digital camera.

Questions of scale are raised in Gregory Crane's 'memography' (Scholarly narrative 0033), where I find myself in a very familiar disciplinary terrain of scale analysis. Trace Plato through myriads of pages of European writing, perhaps the 50 billion words printed annually in American newspapers. For me it is a moment to nod again in the direction of Microsoft and Apple. The Thesaurus Linguae Graecae (TLG) is a very big project, although I suspect that it will require more editorial human intervention at all stages of the process.

It is useful to remember that the 'thousands' of Tibetan books add up to a quite small collection of highly curated data. Tens (but not hundreds?) of millions of words, but they will all fit comfortably on a laptop, metadata included.

Digitization is in many ways easier with dead cultures where you have a limited number of documents, and the project of digitizing all or many of them is a manageable task in terms of scale -- never mind the scholarly labor that classicists, Assyriologists, or Anglo-Saxonists don't mind lavishing on the objects of their research. The Perseus project, which includes just about all major texts of ancient Greek from Homer to the Second Sophistic, is just five million words. The Thesaurus Linguae Graecae (TLG), which includes most written Greek from Homer to late antiquity, adds up to ~100 million words but will fit comfortably on the flash drive of a digital camera.

Questions of scale are raised in Gregory Crane's 'memography' (Scholarly narrative 0033), where I find myself in a very familiar disciplinary terrain of scale analysis. Trace Plato through myriads of pages of European writing, perhaps the 50 billion words printed annually in American newspapers. For me it is a moment to nod again in the direction of Microsoft and Apple. The Thesaurus Linguae Graecae (TLG) is a very big project, although I suspect that it will require more editorial human intervention at all stages of the process.

It is useful to remember that the 'thousands' of Tibetan books add up to a quite small collection of highly curated data. Tens (but not hundreds?) of millions of words, but they will all fit comfortably on a laptop, metadata included.

Digitization is in many ways easier with dead cultures where you have a limited number of documents, and the project of digitizing all or many of them is a manageable task in terms of scale -- never mind the scholarly labor that classicists, Assyriologists, or Anglo-Saxonists don't mind lavishing on the objects of their research. The Perseus project, which includes just about all major texts of ancient Greek from Homer to the Second Sophistic, is just five million words. The Thesaurus Linguae Graecae (TLG), which includes most written Greek from Homer to late antiquity, adds up to ~100 million words but will fit comfortably on the flash drive of a digital camera.

Questions of scale are raised in Gregory Crane's 'memography' (Scholarly narrative 0033), where I find myself in a very familiar disciplinary terrain of scale analysis. Trace Plato through myriads of pages of European writing, perhaps the 50 billion words printed annually in American newspapers. For me it is a moment to nod again in the direction of Microsoft and Apple. The Thesaurus Linguae Graecae (TLG) is a very big project, although I suspect that it will require more editorial human intervention at all stages of the process.

It is useful to remember that the 'thousands' of Tibetan books add up to a quite small collection of highly curated data. Tens (but not hundreds?) of millions of words, but they will all fit comfortably on a laptop, metadata included.

Digitization is in many ways easier with dead cultures where you have a limited number of documents, and the project of digitizing all or many of them is a manageable task in terms of scale -- never mind the scholarly labor that classicists, Assyriologists, or Anglo-Saxonists don't mind lavishing on the objects of their research. The Perseus project, which includes just about all major texts of ancient Greek from Homer to the Second Sophistic, is just five million words. The Thesaurus Linguae Graecae (TLG), which includes most written Greek from Homer to late antiquity, adds up to ~100 million words but will fit comfortably on the flash drive of a digital camera.

Questions of scale are raised in Gregory Crane's 'memography' (Scholarly narrative 0033), where I find myself in a very familiar disciplinary terrain of scale analysis. Trace Plato through myriads of pages of European writing, perhaps the 50 billion words printed annually in American newspapers. For me it is a moment to nod again in the direction of Microsoft and Apple. The Thesaurus Linguae Graecae (TLG) is a very big project, although I suspect that it will require more editorial human intervention at all stages of the process.
Two scholarly narratives (0034 and 0039) concern ancient inscriptions. 0034 begins with the sentence "Doing research with inscriptions is much easier with a digital corpus." Indeed, inscriptions, by their nature scattered documents, do not yield their full query potential until organized in a firm data structure. That was the central insight of Mommsen's great 19th century project of gathering and publishing inscriptions across the entire Roman empire, which revolutionized the study of Roman administrative history. Digitization adds powerful affordances to print publication.

While I am not an epigrapher, it seems to me that the digital opportunities of this discipline -- and of the broader discipline of archaeology -- are quite well understood by its practitioners. The real problems re resources and the fact that the disciplines are often what the Germans call *Orchideenflächer* or 'orchid subjects.' Will Bamboo smile on or ignore projects like a (not entirely hypothetical) dictionary of 'West Tocharian colloquialisms'? An important question, and I wish I could be assured of the smile. But I worry more that in joint planning, however well-intentioned, the little and abstruse subjects in odd languages will get pushed even more to the margins.

There are several scholarly narratives that have the ring of truth to them, but it is less clear what one could do other than just live with them. You're interested in a 19th century Italian archaeologist/orientalist who will make up a chapter in your book on Victorian orientalism (Scholarly Narrative 0045). You scour the web and your library for what you can find and manage the data in your head until they begin to make some sense. Welcome to the world of scholarship. Nothing that Bamboo does will help with this aporia.

Similarly a project about 'keywords' in American culture (0040), an imaginative and acknowledged riff on Raymond Williams' book, is just a matter of colleagues working together and using whatever tools of communication, digital or otherwise, are available.

0043 is a very interesting narrative about digitization of personal research collections. Somewhere in Minnesota a professor has over a lifetime assembled an archive of ~70,000 postcards, images, and other memorabilia of American Indian culture. You don't want this stuff to be tossed or stowed away in file cabinets in some basements. Problems of this kind arise daily all over the world from sleepy local history societies to Research I libraries. Do I remember that Mellon once funded the development of mobile and easy-to-use digitization equipment that at least secures the good enough digital capture of the original stuff?

Two scholarly narrative (0026 and 0027) stress collaborative annotation. A recent ARL report (November 2008) about Current Models of Digital Scholarly Communication stresses the importance of "annotated content" for humanists. In a very interesting draft chapter of his dissertation (http://www.peterboot.nl/onlineannotationdraft.pdf), Peter Boot raises the question why there are still no good annotation tools for scholarly work and argues that the collaborative tools of the wiki world despite their many virtues do not meet the requirements of scholarly commentary. The fundamental problem is that in scholarly annotation the target of an annotation is typically defined with greater precision than in typical digital notes and that existing software is largely ignorant of citational schemes and their importance. The greatest potential advantage of digital annotation is that it can transcend the 'ad locum' limitations of print annotation and provide annotational schemes that may be 'concept-bound' as well as 'location bound'. Here is a software module that would do much for scholars in many disciplines if it were designed in a sufficiently flexible and comprehensive manner.

One thing I miss in the scholarly narratives is attention to well-curated archives of texts written in English, especially texts before 1923 that are in the public domain. Interoperable documents, textual or otherwise, are highly desirable, and 'interoperability' has certainly been mentioned often in the two Bamboo conferences I attended. The ability to range quickly among many texts may well be the most important advantage of the digital medium. If you have one book and one reader, the computer adds comparatively little. If you have one researcher and a thousand texts or more, the digital medium shines.

American universities have taken the lead in creating highly curated and interoperable textual archives such as the TLG, the Cuneiform Digital Library Initiative, or the Tibetan Buddhist archive with which I began this posting.

Where is the sufficiently comprehensive, sufficiently well curated, and fully interoperable archive of texts in English? There are tons of digital texts here and there, but it is difficult or impossible to get them to play well with each other.

From the perspective of textually oriented research, this very anecdotal review of some scholarly narratives raises some general agenda items for scholars and librarians. "Only connect," the motto from *Howards End* may be a good tag line. A digital edition should always be done on the basis of a standard, in practice the TEI. If goals cannot be realized within that standard, it should use the standard as far as it will and use extension beyond it, but in such a way that the edition can be 'stepped down' to be interoperable with other texts.

At the level of library collections, much more attention needs to be given to making text archives interoperable. Even more importantly, libraries need to start thinking of digital 'repositories' beyond the simple model of a digital shelf from which readers will pick this or that book for reading. The digital repository needs to be rethought as a laboratory in which entire collections or subsets become objects of complex manipulations.

The tools for manipulation have been familiar to corpus linguists and experts in information retrieval for decades. Their introduction into the domain of the humanities has its problems for a variety of technical and cultural reasons. The SEASR project is the most aggressive effort to speed up that introduction. Three things are very clear to me. First, these tools are very powerful and promising. Second, their application in a different domain will require many adjustments in the tools, in their use, and in the treatment of the data on which they are used. Third, new tools are nearly always more successful if they are deeply embedded in an environment of old and familiar tools.

How can Bamboo help scholars and librarians do a better job managing the challenges of digital textuality? If Bamboo plays the proverbial role of the onion in the stew or the role of a skilful host who introduces guests and then gets out the way of their conversation, much good can come of it. If it develops a large agenda of its own and is very visibly there, it may spend a lot of money but do little good.