A TECHNOLOGICAL OUROBOROS
Searching Scholarly Narratives in Hopes of Founding a Cyberinfrastructure for the Humanities

John Laudun
University of Louisiana – Lafayette
laudun@louisiana.edu

Figure 1. Wordcloud of textual corpus

Abstract

An analysis of forty-four of the texts collected in the Project Bamboo Scholarly Narrative Repository reveals the perspective from which humanities scholars, as users or potential users of resources, tools, and services, understand the nature of the change technology may have on their work. In particular, nine keys, seven actions and two properties, that recur throughout the corpus come to the fore. The seven actions — digitize, access, search, manage, collaborate, preserve, compute — reveal that scholars and curators, be they librarians or archivisits, are anxious to perform a discrete set of tasks but that they wish to do so within a framework that possesses the properties of being rich with metadata and that the materials and other users with whom they participate are authentic and authorized.

1. Introduction

This paper analyzes 44 of the texts collected in the Project Bamboo Scholarly Narrative Repository in an attempt to glimpse the perspective from which humanities scholars, as users or potential users of resources, tools, and services, understand the nature of the change, or continuity, technology may have on their work. In particular, this paper isolates nine keys, or motifs, that recur throughout the corpus: seven actions and two properties that scholars return to again and again. It should come as no surprise that in examining a corpus of mostly narrative texts collected in the process of planning a proposed "digital infrastructure for humanities research," as the original proposal for Project Bamboo notes, that the words *digital* and *research* should be among the most used terms. In a corpus of 44 texts *digital* occurs 110 times and *research* 96. What may come as a surprise, or at

least reassurance for the planning effort, is that despite any apparent passivity toward or confusion about technology, humanities researchers imagine it in an active way. That is, they want to do things.

A particular set of things. The breakdown by actions — digitize, access, search, manage, collaborate, preserve, compute — may take some readers by surprise, but it is born out by abstracting as little as possible from both actual and speculative workflows detailed in the texts, which are available both on the secured Project Bamboo Planning Wiki [1] and as a composite text [2]. Glancing again at the word cloud (Figure 1) reinforces the orientation toward action in imagining a cyberinfrastructure for humanities research. The prominence of the words *can*, *use*, and *work* are born out by their occurrences within the corpus: can (79 hits), use (66 hits), and work (64 hits).

The properties — annotated, authenticated — are a bit more diffuse, but that seems to be in keeping with the way they are constructed in the texts themselves, as already and everywhere present in the contents made available through any infrastructure and the people who populate it.

2. Methodology

As of June 2009, the scholarly narratives examined here were all to be found in a repository on the Project Bamboo Planning Wiki. There are 58 numbered entries and an additional 15 entries that had yet to be accessioned fully into the repository for a total of 73 texts. However, when navigating from the repository's main page, the links to 11 pages take the user to "Add Page" placeholders in the wiki hierarchy, accounting for 15% of the repository.

Somehow this too seems emblematic of humanistic digital efforts: a decent infrastructure, the wiki, with only limited functionality for the intended use, a database of texts to be analyzed, pressed into service for lack of a better solution. None of this will be news to those of us who have traversed the extant humanistic webs, revealing as it does what often seems idiosyncratic but is more often a function of pressing available resources into, albeit sometimes awkward, service. The problem for the humanities, of course, is that data isolated in such a fashion creates no real opportunity for the discerning of information and the discovery of knowledge.

Of the 62 available texts, only 44 were examined in the current study. The remaining 18 texts were excluded for two reasons: First, while I was not present for the articulation of the desire to have scholarly narratives at Workshop 2, I followed the resulting activities on the Planning Wiki and joined the refinement of the idea and the articulation of its relationship to other parts of Project Bamboo that occurred at Workshop 3. In particular, we perceived an opportunity to tie scholarly narratives to actual workflows. The series of actions, or steps, that comprised a workflow could then be addressed by a recipe. Recipes by their very nature would be successful abstractions of a group of particular problems and thus able to act as predictors of what services and tools would be needed. A quick totaling up of all the recipes would reveal what tasks and features should be ranked above others in terms of priority to be mapped or built. The goal, obviously, was both to do what could be done quickly and easily as well as to be in a position to assess what had the potential to have the highest impact. Some of this thinking is captured from a graphic prepared at Workshop 3 — and thus should be understood as an artifact of its time — which indicates that the movement from a scholarly activity to an API is one of increasing abstraction, from the point of view of the scholar. (See Figure 2.) Second, it seemed to me, then, that the texts that would be most interesting to examine would be those that were most closely tied to actual scholarly workflows, and, as it turns out, were most like, well, a narrative.

For my purposes, I used a well established definition of narrative that defines it as simply a *sequence of clauses which are temporally ordered* [6]. This made it easy to remove from my consideration a number of texts which were simply abstracts of projects. A secondary restriction I developed was that the relationship of the agent to the action had to be clear. This meant removing from consideration a number of texts which were proposals of some kind or another, and usually written with a level of abstraction that made it difficult to discern who was doing what. This left me with 41 texts that I had provisionally described as either *narratives* or as *scenarios*. In my survey of the texts, I reserved the use of *narrative* to first-person accounts that were either in the past tense, and thus revealed an authentic or habitual action that

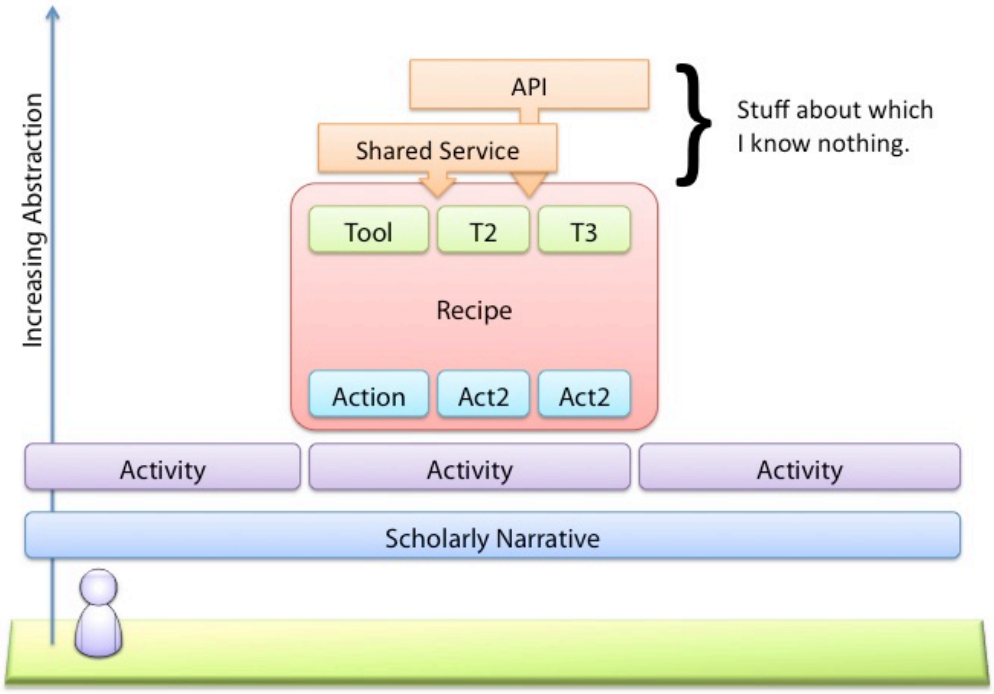

Figure 2. From scholarly activity to API

had taken place or, when in the future tense, revealed the narrator capable of imagining themselves taking such an action. I ascribed *scenario* to texts that were written in the third-person and often worked with a kind of composite scholar, which may or may not be based on personal experiences.

Before proceeding any further, it might be useful to provide some example texts that bear out some of these distinctions. The first example is of a text of a type that I chose to exclude:

> The Internet provides unprecedented new ways of compiling and publishing this information as a dynamic, collaborative, ongoing process. A seed catalogue containing known information, perhaps from an already well-documented collection or exhibition catalogue can provide a model for the entries and can form the basis for gathering new works and information. One of the important characteristics of this method of research is to make the seed catalogue discoverable on the Internet. This can be a potent way of attracting potential collaborators and also of enabling non-researchers — dealers, buyers, sellers and private collectors - to discover the project and contribute information on works they own or that have passed through their hands. Some may become researchers in their own right and contribute directly to the growing catalogue. Others may provide information via more traditional methods — letters, email, telephone discussions, visits. Support for a creative commons approach might be the default for this tool. However, there would need to be a minimum level of access control for inviting contributors and approving contributions. At some stage in the process, a version of the Catalogue Raisonné might still be published in the form of a high quality printed monograph. [SN-0002]

No doubt, the text proposes an interesting idea, but it in no way compares to the following text in terms of concrete detail that gives us insight into how humanities scholars go about their work:

> I am investigat[ing] a group of 16th century (Ming dynasty) Chinese painters … I need to perform the following tasks:
>
> 1. Find all mention of these artists in texts that date to the sixteenth and seventeenth centuries. Such material primarily includes the collected writings of individuals, local and imperial histories, and gazetteers. Read and translate such material.
>
> 2. Because these painters were categorized with the label "Zhe School" at some point in the 17th century (this label was construed as perjorative), I also need to find all uses of the term Zhe pai 浙派 in texts that date to the sixteenth and seventeenth centuries. Read and translate such material.

> 3. Examine all extant attributions to these painters, with particular attention to any inscriptions and seals by other contemporary figures who either saw or owned the work.
>
> 4. Examine anonymous paintings attributed to the Song dynasty and anonymous paintings of the Ming dynasty that exhibit the styles of these artists in order to look for seals of sixteenth century individuals. [SN-0013]

In this particular instance, the contributor has gone so far as to enumerate the tasks, but the more important dimension of this text is the use of active verbs that take objects: "I need to … find all mentions" or "I need to … examine paintings." With such a text we have a very clear and vivid sense of what it is the scholar has done, does, or will do. It clearly is derived from authentic experience, and, arguably because of this derivation, the flow of actions taken is quite evident.

A final form of text that I decided to include within the working corpus for this analysis is one that I provisionally dubbed a *scenario*. My working definition for this genre was that it was often composed in the third person and was sometimes written in the conditional, e.g. "a scholar would do this, were it available." An example of this kind of text, which actually uses the term scenario to describe itself, is:

> Now in our scenario, faculty members in Computer Science, East Asian Languages and Cultures, and South and Southeast Asian Studies team up to address the problem. First, the digital images are stored in the campus archiving repository, which provides improved speed of access, reduced costs, and a guarantee of permanence. Achieving the requisite level of accuracy will itself require the development of new OCR techniques by Computer Science Professor 1 (CS-Prof1) guided by syntactic and semantic models co-developed with East Asian Language and Cultures Professor 1 (EALC-Prof1). Metadata on authorship, woodblock location, etc., is added to the corpus. [SN-0051]

Time, and availability, permitting I would like to run these texts through something like DocuScope to see if it makes similar generic distinctions. For now, I am confident that these provisional genres will serve their purpose in allowing us to survey the data collected so far and give suggestions for future data collection, methodologies, and analyses.

3. Analysis

Across the 44 texts, there were a consistent set of seven things that scholars wanted to be able to do and two properties that they wished to permeate their actions. The seven actions were: digitize, access, search, manage, collaborate, preserve, and compute. In addition to actions, there were two properties, which in some sense pervaded almost every text, annotated and authenticated. (In the on-line database of this corpus, both actions and properties are labeled as keys.)

3.1 The Seven Actions

I have ordered the presentation of the actions by the sequence in which they seemed to appear in the texts, but it should be noted that not every text addressed all actions nor did every text necessarily sequence the actions in this order. The composite text of all 41 narratives and various visualizations are available on the Alphaworks site[2].

Digitize. Many scholars expressed the importance, and often the difficulty, of digitizing materials for study. (A correlative observation to this point of the importance of digitizing materials for study is that the studies themselves are assumed to be born-digital in some fashion. That is, most scholars are working digitally, even if their inputs and output(s) is non-digital.) Objects ranged in nature from Tibetan books written on woodblocks to handmade boats witnessed in their working environment to theatrical performances to fragile manuscripts. Each object type posed a unique challenge but none, at least within the narratives themselves, seemed insurmountable in terms of supporting humanistic study. In many instances, the anticipated difficulties with rendering an authentic digital version of an object were really a function of the problem of representation and thus something that will have to be decided within ongoing disciplinary conversations over established, or emergent, conventions. This latter point perhaps indexes the difficulty of achieving that utopian vision that sometimes effervesces through the narratives as well.

The goal of digitizing an artifact — be it a text, event, or artifact — was closely tied to other keys: to make it accessible, to make it searchable, to make it (and larger sets of items like it) computable, and to be able to preserve it. The connected nature of these

activities is born out below in the discussion of the other keys, but it bears some emphasis here: a word tree of all 44 texts reveals that the term *digitize* occurs in the following instances:

> **digitize**, catalog, and upload content to the digital repository
>
> **digitize** most of the non-digital artifacts
>
> **digitize** the bulk of the texts I need to reference

Other forms of digitize were: OCR, scan, and automate. All of them reveal that narrators were aware of the wealth of data available and, at times, felt overwhelmed by the possibilities. Attitudes seem to have changed from "working with what one has" to something more like "if one is going to do this right," indicating that at least for these narrators, some sort of switch has occurred in terms of how one imagines the nature of scholarly action.

Access. Access can mean a variety of things, but for the sake of this analysis, I am constraining the term to mean the ability of a researcher to avail herself of pre-existing digital materials. It was a fairly common complaint within the corpus: digital sources were available; they just weren't accessible. The reason for the lack of access varied. On the one hand, there were financial difficulties: the narrator's institution did not subscribe, or could not afford to subscribe, to a particular source. On another, there were logistical difficulties — sometimes despite, but sometimes because of, technology: there were a number of accounts of data being trapped within clumsy interfaces, of having to log into multiple sites in order to access very similar data, of data not being easily ported out of its repository and into a space where it could be *computed*, *managed*, or *annotated* by the user. Both of these cases are underlined by the fact that over half (21 of 37 instances) the use of the word *access* itself occurred within the phrase *access to* as seen in Figure 5.

Narrators were not simply interested in access but sought access in order to get *to* something within a data store. This may seem a trivial point, but it does reinforce the notion that the potential users of a Project Bamboo infrastructure are interested in acting upon it in order to get certain kinds of work done. They are not interested, by and large, in the structure itself. In fact, to this point, the dominance of *access to* here, in conjunction with the content of some of the narratives as noted above, emphasizes the importance of transparency.

Also the place where intellectual property issues most often arose.

Search. In an examination of the most commonly occurring words within these texts, the two most

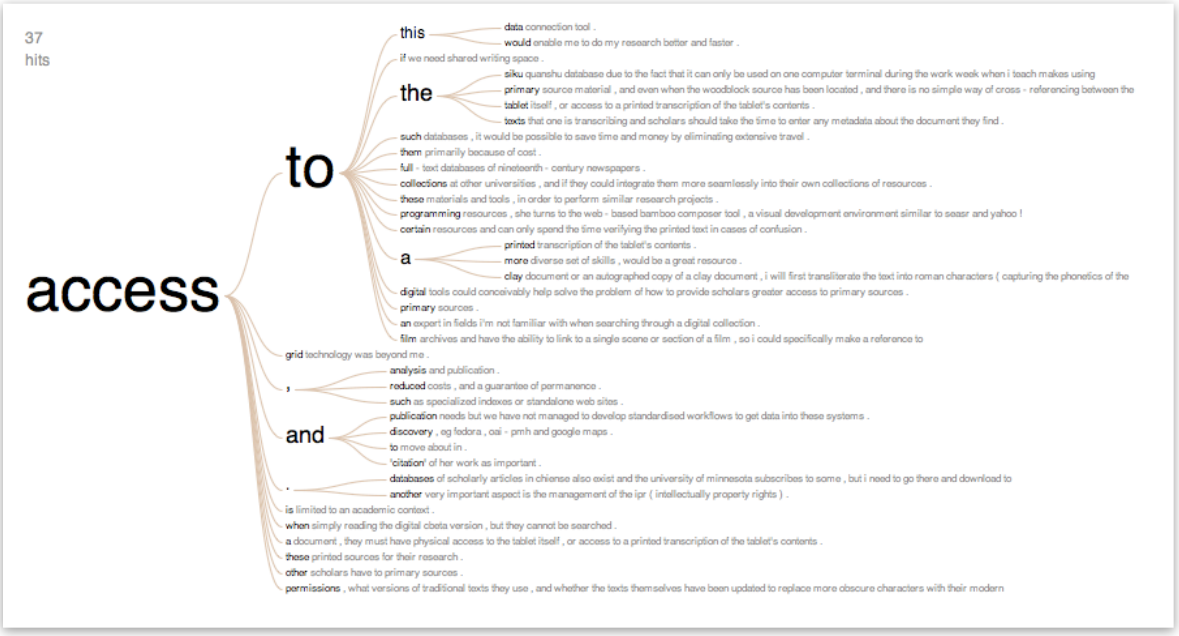

Figure 3. Access

prominent were *digital* (110 occurrences) and *research* (96), neither of which should come as a great surprise given the ostensible focus of Project Bamboo to build a digital infrastructure for humanities research. Tying in popularity were the terms *data* and *search*, each occurring 69 times in the corpus and bookended by the terms *can* (79) and *use* (66) discussed above.

Search is perhaps the most obvious of the actions that scholars should wish to perform, resonating as it does with the prominent place of searching in almost any use of the internet. Indeed, the first technological word to achieve the role of a verb, the strongest role a word can possess, is *google*, eponymously named for the information company that dominates the on-line search landscape.

Manage. What is being managed here is data; the management of people and projects, though certainly there are dimensions of the latter contained here, is grouped under collaborate below. While digital asset management (DAM) is a well established product category, most of those applications focus on media, principally images and video. However, while the narratives treated here reveal that media plays, and will play, an ever increasing role in the work of humanities scholarship, what most researchers imagined is a much broader spectrum of data to be managed: primary materials like texts, image, audio, and video as well as secondary materials — some of similarly complex nature — as well as personal notes and data. Such agglomerations of materials are very different from the kinds of collections curated in archives or organized in libraries and repositories. Researchers create sets of diverse materials that cut across a variety of conventions and/or disciplinary boundaries in order to achieve the kind of synthesis that leads to genuine innovations within their fields. The texts hinted at the possibility for genuine innovation and breakthroughs texts and yet, at the same time, were haunted by a sense of being overwhelmed by so much material of such diverse nature. One narrator provided a terrific inventory of a collection she was working with, noting she was faced with:

> Interviews with Muscogee Creek individuals in Oklahoma. Microfilms of fourteenth century Southern French notaries. Tapes, transcripts and translations of hundreds of interviews spanning 20 years of research. Rare photographs and ethnographic objects. A map collection. Archaeological artifacts, primarily ancient ceramics but also original field notes, inventory

Figure 4. Search

> cards, and other primary excavation records. Approximately 40-50,000 slides, covering four continents and 30 years, of mostly urban architecture. A small number have been digitized. Russian newspapers from the 80s and 90s that I don't know what to do with. Paper survey data from 20-year longitudinal sociology study

And summed it up quite well by noting:

> The organization and storage of research materials is a subject of significant interest for researchers. They concern themselves not only with methods of classification developed by librarians and archivists, but also, and most immediately, with their own methods of organizing and habits of collecting.

Collaborate. Between a quarter and a third of the texts in the current corpus raised collaboration either directly or indirectly. While a number of users were interested in digitizing, streamlining, or automating certain aspects of collaboration, it is also interesting to note that users also saw this is one of the most available spaces for innovation. This is perhaps reflected in the number of texts that were classified as scenarios in genre, speaking as they did of some future scholar, typically narrated in the third person — and a number of times as Scholar A interestingly enough, e.g., 0027, 0051, 0054c. Users also revealed an interest in doing something that almost contradicts one of the hallmarks of the humanistic tradition, that of the long view, in seeing the opportunity to move quickly, to create temporary groups to perform discrete tasks and then disperse:

> I will often try to set up a a working group around an idea or project with a few grad students and colleagues. The colleagues might be at other institutions. These working groups are for new ideas for which we don't have grant funding to travel or pay for infrastructure to be set up. We need a mix of ways to communicate, share files, collaborate on writing grants. We need to be able to meet online regularly. We need to be able to set these up quickly and to be able to add communication tools as we need them. We need to be able to do this with a minimum of bureaucracy. We need to be able to close them down and archive stuff. Sometimes when we move we need to actually move the hair ball to another university. [SN-0009]

For these users, a digital infrastructure lowers the costs of bringing together and organizing people and materials.

Evident in the quotation above is also the desire to stretch the boundaries of collaboration to include not only students but also the research subjects themselves, as was the case with one project working with Australia's aboriginal peoples (SN-0011) which were not only sources of information but potential colleagues in analyzing and classifying.

Even when users described simply collaborating with other scholars, they tended to imagine, in some way, new, more granular, forms of scholarly communication. Two particular cases that focused on the work of interlingual translation of texts — Latin in one case (SN-0026) and Tibetan in the other (SN-0027) — focused on the ability of scholars working across international borders to arrive at acceptable horizons of understanding for key terms and passages. This work gets done now, but often occurs within the casual space of correspondences or conference hallways or in the constricted space of articles which often has to await enough other material before moving to publication.

Preserve. I have chosen the descriptor *preserve* here in order to avoid the larger domain, and sets of issues, associated with *archive*. Many users conflate archives as places with the curatorial work that occurs in archives and with the action of archiving something for preservation — or with the task of asset management which is broken out above. Also, given the place of librarians and archivists within the Bamboo collective, it seemed wise to choose a word that was separate from existing usages and users.

Preservation was most often the focus of users that most often dealt with artifacts or events that were subject to loss or destruction through the simple vagaries of time: wood blocks, clay tablets, musical and theatrical performances, and intangible cultural forms like local history and creation myths are good examples. Most often, as noted in the discussion of *manage* above, there was some concern about how best to preserve non-digital artifacts within a digital realm. In most instances of the use of the word *preserve* specifically, it was tied to making items available:

> "preserve and make available a body of information" [SN-0030]

> "preserve and make accessible increasingly complex research collections" [SN-0043]

> "preserving the descriptive vision of the performance and making the materials as accessible as possible" [SN-0057]

Finally, it should be noted that some uses of *preserve* were in reference to the scholarly record itself. The humanities are a collection of fields with a long view of the past, and, it seems, of the future as well.

Compute. While early incarnations of the digital humanities focused on harnessing the computational power of information technology either to what we now know as data-mining or to some other hermeneutic task, it is not an overwhelming focus of the current set of texts. Eight texts, however, did focus on the use of computers as, well, computers. In a number of instances, this computation power takes the form of optical character resolution (OCR) of nineteenth-century newspapers (SN-0014) or of Tibetan woodblock books (SN-0051) or of cuneiform on clay tablets (SN-0063). Other computing uses include genre recognition through word use. The analysis of Shakespeare using DocuScope is a particularly interesting example:

> What they found was that Shakespeare's comedies and histories were written with distinctly different diction. For example, the comedies have the most "Interacting" words, which Hope said is plausible because of the witty dialogue that often takes place in comedic plays. A more surprising finding was that comedies had a higher frequency of "First Person" words. [SN-0021]

Other texts tagged as compute described a kind of object-oriented approach to analytical problem, even referencing programming as a model:

> Because she is not a programmer and does not have access to programming resources, she turns to the web-based Bamboo Composer Tool, a visual development environment similar to SEASR and Yahoo! Pipes, to graphically connect the content resources with the timeline widget. She expands upon the content with her own research, which includes a unique approach to medieval maps, and creates her new application. She embeds the timeline within her web page and exposes her tool, Timeline of Anglo-Saxon England (TASE), to the Bamboo Community for others to use. [SN-0052]

3.2 The Two Properties

Annotated. References to *metadata* (alternate spelling *meta-data* included) occurred in 15 out of the 44 texts. It should come as no surprise that humanists are heavily invested information about data. One boat or wood block text may very well look like another to the untrained eye, but with a little metadata significant differences become apparent. In other instances, important historical or cultural dimensions of an artifact or text are not necessarily available within the object itself and require further glossing or, as it is keyed here, *annotation*, which is used here to emphasize the particular way metadata was used or described within the corpus.

There were a total of 38 usages of *metadata* (plus 5 of *meta-data*). Most instances were in relationship to an artifact or kind of artifact, e.g., a book r inscription, but a few also paired *metadata* with a scheme, structure, or, in one case, ontology. What is perhaps most striking about the uses of metadata, and why it is broken out as a property and not an action to be performed, is that text after text assumed, or projected, a space in which metadata was pervasive, in which every object has a cloud of information about it. In fact, what was somewhat interesting was the degree to which this richness of the infosphere was imagined as already having occurred. There were far fewer accounts of the construction of such a rich space, of the work necessary to achieve it. A simple example will do here:

> Digital corpora allow searching on both content and metadata for relevant information. The print indices aren't as rich. [SN-0034]

The assumption of metadata having already happened is born out by the construction of the assertion in two simple sentences: on the one hand the digital realm is rich, while the physical realm is poor. Left out in the ideational rhyme is where the additional richness will come from. (To some degree, the active nature of the digital realm discussed above may account for assumptions users make about the properties objects within it will possess.)

There was some difference between librarians and humanities scholars in awareness of the differences between more technical usages of *metadata*, as, for instance, in the application and maintenance of

controlled vocabularies, and a looser usage which could simply mean "with notes."

Authenticated. Another feature that the texts assumed would be pervasive in some form is the *authentication*, or *authorization*, of users and information. That is, users project a trusted system from which they can draw data and information and to which they can safely contribute data and information. In some sense, the underlying question issue is who gets to contribute to data or knowledge and how are we to interpret-understand-evaluate their authority-expertise within a diffuse network? It should probably come as no surprise that users who have spent years acquiring expertise, by authoring publications, in order to be considered authorities on particular subjects should be concerned with how authorization will occur. At the same time, there was, as the discussion of *collaborate* above demonstrates, a remarkable openness to who can be *authenticated*, or *authorized*, to contribute to humanistic research. One text foregrounds the process in some detail:

> Scholar's A and B then invite scholar C to contribute to the project, as a collaborator. Scholar C goes to the site, identifies himself and provides his titles, affiliations and a list of his publications and presentations most relevant to the project. [SN-0027]

Most descriptions were not this granular, in part because a number of text assume that any cyberinfrastructure will extend or transform current hermeneutic communities of practitioners, many of which are already familiar with each other, by name if not in person, making *authentication* a pre-existing condition.

4. Conclusions

Seven actions, two properties, forty-four texts. Obviously such numbers make it clear that this work is provisional in nature, a start toward more granular analysis with greater potential power to refine the nature and scope of Project Bamboo's mission. Because one of the discussed uses of the stories was to use them as bases for recipes, one of the potential outcomes might be simply to help organize the recipes, oriented as these actions are not from a technological perspective of how to get things done but from a user's perspective of what needs to get done.

Even within the narrow scope of this analysis, it is quite clear that much more work needs to be done, both in terms of statistical analysis of the corpus and in terms of close readings of particular texts. The latter could be especially useful in potentially determining user profiles, which would, in turn, carry forward the idea operating behind this paper, that humanistic analysis can be an effective way to arrive at ideas and practices that could frame a cyberinfrastructure that was itself focused on supporting humanities research.

5. References

[1] https://wiki.projectbamboo.org/display/BPUB/Home

[2] http://manyeyes.alphaworks.ibm.com/manyeyes/users/johnlaudun

[3] Katz, Stan. 2008. The Emergence of the Digital Humanities. *Chronicle of Higher Education* (2008 April 7): http://chronicle.com/review/brainstorm/katz/the-emergence-of-the-digital-humanities.

[4] Laudun, John. 2000. "There's Not Much to Talk about When You're Taking Pictures of Houses": The Poetics of Vernacular Spaces. *Southern Folklore* 57(2): 135-158.

[5] Laudun, John. 2001. Talk about the Past in a Midwestern Town: "It Was There At That Time." *Midwestern Folklore* 27(2): 41-54.

[6] Laudun, John. (In press). Following The Way of the Masks. *Journal of Folklore Research*.

[7] Labov, William and Joshua Waletzky. 1967. Narrative analysis. In *Essays on the Verbal and Visual Arts*, 12-44. Ed. J. Helm. Seattle: University of Washington Press.