Next Generation Finding Aids: Creating the Polar Bear Expedition Digital Collections

Elizabeth Yakel, Seth E. Shaw, Magia Krause, and Jeremy York, University of Michigan School of Information Ann Arbor, MI; Polly Reynolds University of Michigan Bentley Historical Library Ann Arbor, MI; James Sweeney, Consultant, Ann Arbor, MI (USA)

Abstract

Although the advent of the Internet has increased the amount of information about primary sources online, the intellectual accessibility of archival materials and researchers' ability to effectively reuse digital archival materials and surrogates is not known. Many of the online representations of archival information, e.g., finding aids, mirror their paper counterparts both in look and in functionality and do not take advantage of the electronic environment. This paper describes the design, implementation and launch of a new type of archival access system, the Polar Bear Expedition Digital Collections (http://polarbears.si.umich.edu) which provides traditional searching, browsing, and interlinking as well as social navigation features such as collaborative filtering, commenting, and awareness of other registered visitors.

The Next Generation Finding Aid Project and the Polar Bear Expedition Collections

Archives and special collections have made great strides in mounting information on the Internet; however the ability of researchers to effectively reuse digital archival materials and surrogates is not known. What we do know about the researchers' ability to understand and use EAD finding aids is not encouraging [1, 2]. Studies have found that researchers are hampered by such diverse problems as poor site navigation and archival jargon. Since most archivists are still struggling with the basic representation of online finding aids, few have begun to reenvision the online representation of finding aids or experiment with Web 2.0 functionalities. Web 2.0 refers to the second generation of Internet-based services. These are characterized by their interactive and collaborative nature and feature shared control as found in sites such as social networking sites, wikis, and folksonomies.

In 2005, faculty and students from the University of Michigan School of Information (SI) began a research project investigating "Next Generation Finding Aids" with the idea of rethinking and enhancing traditional finding aid structure and functionality. While many repositories employ Encoded Archival Description (EAD) to generate their online finding aids, none take full advantage of the properties afforded by EAD and XML-based systems. This paper reports on a pilot project to redesign and implement a "Next Generation Finding Aid." Our guiding research question is: 'do new types of access tools enhance the accessibility of primary sources?' To answer this we concentrated on three aspects of this endeavor: 1) the reuse of existing metadata, 2) rethinking how traditional architecture can best be used to provide access to finding aids, and 3) how Web 2.0 features can enhance visitor interaction in the system. This project is part of a larger research project investigating how archives and special collections can take better advantage of the web and new technologies.

The Polar Bear Expedition Collections, a group of 64 manuscript collections at the Bentley Historical Library (BHL) on the University of Michigan – Ann Arbor campus, document the "American Intervention in Northern Russia, 1918-1919," (nicknamed the Polar Bear Expedition). During World War 1, the U.S. military was ordered to Northern Russia to fight the Bolsheviks. Since military units were geographically assembled at that time, many of the soldiers involved hailed from Michigan. The Bentley has been collecting materials concerning this event since the 1960's making this one of the deepest and broadest collections on this incident. The BHL collections consist of diaries, letters, photographs, oral histories, and a motion picture film. In 2004, these collections were digitized as a digitization experiment as well as for preservation reasons.

We selected Polar Bear Expedition digital collections for our initial experiment for several reasons: the depth of the collections, the uniqueness of experimenting with entire digitally reborn collections rather than selected documents, and the Polar Bear Expedition collections' established following of genealogists and historical researchers interested in this unique historical event.

Planning for the Polar Bear Expedition site began in January 2005 and the site officially went live in January 2006. The site is implemented using the Everything Development Engine (www.everything2.com), a content management system. Everything2 was chosen because at the time at the time, it best supported the social navigation features with which we wanted to experiment. We use SQL as the backbone of the content management system and ImageMagick and Zoomify to render the images. The site is designed around a series of cascading style sheets. In terms of functionality, we were inspired by socio-technical systems in everyday use, including Amazon.com, Flickr.com, and deli.cio.us.com. In the end, we selected the following set of features and functionalities to enhance the finding aid: 1) bookmarks, 2) comments, 3) link paths, a form of collaborative filtering used by the Everything2 engine, 4) browsing, 5) searching, and 6) user profiles.

One of the goals of the project was to experiment with data reuse. In addition to the approximately 12,000 digital images from the 64 collections, we populated the content management system with information from three existing data sources: EAD finding aids, MARC records, and a database listing over 6100 of the soldiers in this campaign.

Literature review

Usage of Web 2.0 features has been haphazard in archives and special collections and little evaluation of whether these features have enhanced users' visits to repository websites has been done. In lieu of a traditional literature review, we review selected Web 2.0 implementations in archives and with primary sources outside of formal archives.

The Science and Technology in the Making (STIM) Project (http://sloan.stanford.edu/index.htm) was the first attempt of which we are aware that envisioned and incorporated more interactive features into the development and presentation of online archival collections. This series of five projects, begun in the late 1990's, aimed at determining whether web-mediated scholarship was possible. One project, which appears to have had the longest active life, is centered on the history of the Blackouts of 1965, 1977, and 2003. It includes the ability to contribute personal stories to the site. Although groundbreaking, the projects were hampered by the limits of the technology. The threaded discussion lists and comment features were not robust enough to support the community of users that project directors envisioned. The final report notes that "It is difficult to transform communities that have a life outside of the Web into communities that work on the Web unless they believe they are doing real work…to be successful… projects should attempt to integrate the work of the community they wish to address…To be 'sticky', sites will need to be self-sustaining and provide an archival presence. By self-sustaining, we mean having the ability to continue to generate interesting and useful content; by archival, we mean having the ability to manage that content for the benefit of its users" [5]. The STIM project is the only one for which we have found any attempt at systematic evaluation, even though as these comments reveal, the results were not totally positive.

A more recent example of taking scholarship to the web is the September 11 Digital Archive (http://www.911digitalarchive.org/) which solicits stories from visitors about their activities on 9/11, their feelings about the event, and how it has affected them. Projects such as the Blackout and the September 11 Digital Archive seek narratives from people to help shape the historical record and to provide a counter narrative to official records and the voices of those in power.

A related example from the museum world is the Art Museum Social Tagging Project (http://www.steve.museum/). This endeavor allows visitors to "tag" or assign their own one-word descriptions to art objects. The Art Museum Social Tagging Project aims to democratize the description of artwork with the hope of increasing the potential audiences for art. Researchers studying this project have found that social classification, known as "folksonomy," can complement traditional museum documentation practices and provide unique access points. Evaluation of the Steve museum is ongoing [6].

Since the 1960's archives and special collections have collected more broadly. However, providing detailed access to the myriad photographs, papers, and other media is difficult. Social navigation tools potentially provide some of the answer, taking the description burden off of archivists and potentially allowing experts and interested visitors in offering their own interpretations and descriptions. While social navigation features have not been utilized broadly in archives, the three examples below demonstrate how different mechanisms result in different types of interactions and elicit different forms of knowledge from visitors.

The first example of an archive employing such interactive technologies is the Haags Gemeentearchief (Municipal Archives of The Hague, http://www.denhaag.nl/smartsite.html?id=37609). The Hague municipal archives has implemented a comment feature for a collection of local photographs. Visitors can describe photographs, respond to other visitor's comments, and / or correct the municipal archives descriptions. The Hague project provides a voice in a very different way than the Blackout or September 11 Digital Archive, which sought first-hand accounts of and public reactions to historical events.

In the second example, the Everglades Digital Library (http://cwis.fcla.edu/edl/SPT--Home.php), takes another approach and allows users to rank items in the digital library. This form of social navigation serves as a recommender system. The more people who use and rate an object, the better the system becomes at identifying the best resources. This mechanism is also the closest to the familiar practice of peer recommendations and citation chaining (or working backwards from the footnotes of books or archives to see what sources others have used) [7].

A final example of the use of social navigation tools in archives is the Ohio Memory Project (http://www.ohiomemory.org/). The Ohio Memory site allows users to create and share online scrapbooks derived from photographs, which the Ohio Historical Society has digitized and posted online. This approach is similar to shared bookmarks (as used in del.ici.ous.com) and provides a way for visitors to easily reuse archival information and share it with others.

While these are all interesting approaches to incorporating social navigation features and shared authority in archival sites, there has been little evaluation of these endeavors. Thus, our goal was to build assessment into the project from the beginning.

Methods

In order to answer our research question, do new types of access tools enhance the accessibility of primary sources? We have utilized a multi-methodological approach combining quantitative and qualitative elements. The data collection mechanisms in place include: web analytics (internal transaction logs since January 2006 and Google analytics since August 2006), visitor surveys (March 2006 and March 2007), interviews with visitors, and a content analysis of visitor contributions to the site. Thus far, the strongest evidence comes from the site itself (the transaction logs and the contributions/comments of visitors). Response to the initial survey was sparse (6 respondents probably because of an overly conservative randomized survey pop-up on the site). Of those surveyed, only three visitors to the Polar Bear Expedition Digital Collections site agreed to be interviewed. As a consequence, the survey and interview data should be viewed as anecdotal and illustrative of usage patterns identified through the web analytics and content analysis. Therefore, we have triangulated these data in our evaluation to analyze both the quality and quantity of use of the websites' features and functions, particularly those that support social navigation.

Findings

Three aspects of the design and implementation of the Polar Bear Expedition Digital Collections address accessibility: reuse of existing metadata, enhancing traditional finding aid functionality and applying Web 2.0 features to the finding aid.

The reuse of existing metadata is a behind-the-scenes element, but it affected our ability to enhance both traditional features of finding aids and to add Web 2.0 functionality. As previously noted, we imported legacy data from MARC records, EAD finding aids, and a Filemaker Pro database of soldiers' brief biographical data. Since EAD is an archival descriptive standard in the United States, we made a decision early on to work with EAD and to study how easily it could be reused. EAD reuse was the most complicated and the most important for the site.

The digital images were delivered to us with minimal metadata so we individually described each item. We then used scripts to tie descriptions to images during processing. Two aspects of the EAD immediately became apparent. First, EAD's flexibility and lack of normalization hampered our ability to ingest these data into the system consistently and required both scripting and hand coding. Second, to support the rich browsing structure we envisioned, additional coding within the EAD finding aid was necessary. This also entailed hand coding as well as authority control work to merge like concepts as well as discrepancies between the Russian and English place names. Balancing human intervention and trying to optimize machine processing became a key managerial decision. As will be seen, this did pay off in terms of usage of the browsing features; however, we realize that intensive labor renders this kind of descriptive intervention impractical for most projects.

In displaying the EAD, we also made a decision not to display everything originally encoded in the EAD. We eliminated duplication, like broad subject terms and decided not to display some information which we thought was not relevant or redundant for researchers, such as collection identification numbers and extent. In the latter case, because the collections were small the extent was readily apparent. As a result, our EAD display is streamlined and minimizes redundancy.

As we have discovered from our findings, traditional features such as browse, search, and bookmarks can be enhanced to improve accessibility, supporting site navigation and information discovery. Previous research on online finding aids demonstrated that no conventions for either search or browse are used. Browsing in particular is often hampered because most browsing structures rely on users knowing the exact title or creator of a collection, rather than more useful browse lists such as access by subject [7]. Search is often thought to be the more utilized navigation feature on websites. However, research has demonstrated that this preference for search is often the result of inadequate browsing structures and that, if done correctly, browsing would be the chief method of navigation on a website [8, 9]. Therefore, we decided to create a rich browsing structure for the Polar Bear expedition website.

Visitors to the site can browse using 7 different categories: collections, individuals, military units, geographic locations, subject, media type, and organizations. Interviewees liked the browsing feature. As one individual stated, "My preference has basically been to browse by just going through type of subset and then either alphabetically or whatever then go down, scroll down the list until I find what it is I'm looking for" (Interview 1, section 19). The web analytics overwhelmingly support this browsing preference. In a 6-month period, browse was selected three times as much as search. Table 1 shows the frequencies of accessing the search (in italics) and browse (in bold) features. The table demonstrates that of the top 20 pages accessed, 6 of the 7 browsing categories appear. Interestingly, the access point typically provided in most archives' finding aids browse lists, "browse by collection" is third in frequency for the browse categories.

Table 1: Browse vs. Search (15 August 2006 – 15 February 2007)

Page Title	Unique Views	Page Views
Homepage (Welcome)	5411	7457
browse by: **geographic location**	1859	2911
new advanced *search*	1822	4162
browse by: **individual name**	1566	3614
browse by: **collection**	1448	2578
browse by: **media type**: Photographs.	1275	2291
Polar Bear History	1126	1344
browse by: **military unit**	830	1579
United States Army Signal Corps photograph collection	692	1600
Frank J. McGrath photograph album.	592	740
About this site	498	623
Search Results	435	682
Levi Bartels papers	422	754
browse by: **media type**	383	571
Earl V. Amos papers	378	467
Aldred S. Buckler photograph collection: folder 1	322	418
Polar Bear Association photograph collection	316	493
Frank J. McGrath photograph album: folder 1	305	719
Aldred S. Buckler photograph collection	296	485
browse by: **subject**	274	429

Bookmarking, one design feature we experimented with in order to enhance basic access, was less successful. In the first year, only 8 people utilized the bookmarking feature, creating a total of 35 bookmarks. The number of bookmarks ranged from one person with 19 (an outlier) to 4 visitors with 1 bookmark each (the mode). The interviews and surveys revealed interest in the bookmarks, but the web analytics data on their usage tells another story.

Sites such as del.icio.us.com have demonstrated the value of shared bookmarks and how networks develop around common interests. One explanation for the low usage of bookmarks may be that, unlike the Ohio Memory Project, they are not sharable (a decision we made for privacy reasons); however, they do enable registered users to return directly to favorite pages or items on the site. In designing the site, we saw this as a way to reduce the load on users' memories and to create a sense of customization for visitors to the site.

In addition to enhancing traditional features, the Polar Bear Expedition Digital Collections support social navigation through user profiles and user awareness, link paths, and comments. We selected such features because they represented a combination of direct and indirect interaction. Direct social navigation represents explicit action on the part of the participant. Examples would be either asynchronous conversations through a comment function or synchronous online chat. Individuals engaged in direct interaction are aware of each another. Indirect interaction is unobtrusive and relies on mechanisms that indirectly provide suggestions or recommendations to people [11]. Examples of indirect social interaction are recommender systems such as the one featured on Amazon.com that informs buyers what other individuals who bought one book also bought or the rating system utilized the Everglades Digital Library. On the Polar Bear Expedition site, the user profiles and comments enable direct social interaction; the link paths are indirect. Use of these features varies.

Visitors to the Polar Bear Expedition site have the option of registering. Registration gives visitors several benefits: the ability to contribute comments, bookmark information or images, see simultaneous visitors, and provide a user profile. As of February 2007, 221 individuals had registered. Of those registrants, 9% or 19 people created user profiles. Analysis of these profiles revealed that most (13) have a family connection to the materials. Of the remaining 6 individuals with profiles, 4 have an interest in the history of the event or the time period and 2 provided no information about their interests. The information provided in the user profiles went beyond biographical data. The comments in the user profiles can be divided into three categories: additional information about an individual or updates on the family or the individual, questions about an individual or searching for information, and information sharing (in several cases adding a URL). Most of the people in the user profiles are provided real names and 4 listed contact information.

We adapted the collaborative filtering mechanism, "link paths," used by Everything2 for the Polar Bear Expedition Digital Collections. Collaborative filtering is a means of automatically generating predictions (filtering) about the visitor preferences based the aggregation of previous usage data from all site visitors (collaborating) [12]. In this way, the link paths are a type of recommender system and literally relay feedback to visitors on how others reached a particular item or collection. The more people who use the site, the better the filtering mechanism will become. Amazon.com uses this type of algorithm when they offer book suggestions based on the purchasing patterns of customers. As usage of the site grows, we hope that new and unexpected relationships will emerge between subjects and collections to enable researchers to make unanticipated connections through the link paths. The link paths feature is one way that we adapted the "signs of use," such as margin comments or dog-eared pages, found in paper finding aids to the virtual environment.

There have been several problems with the link paths that have led us to fine tune this feature. Initially, we simply created a label titled "Link Paths" and provided a link to a help page for a description. The user surveys and interviews demonstrated that researchers were puzzled about the link paths. We thus redesigned this feature in a way we thought might be more familiar to visitors. We adapted the Amazon tag line for our site: 'Researchers who viewed this page also viewed…'

We also noticed that the link paths were not populating very rapidly. When they were populated, the homepage and help pages appeared. Consequently, we adjusted the link paths algorithm in February 2007. We relaxed our constraints for the number of times a particular link had to be made to populate the link paths. We have also eliminated several types of information from populating the link paths, such as help screens, about us, and contact us. Finally, we decided not to display user profile information in the link paths for privacy reasons. Now the link paths display collections, digital images, or information from the soldiers' database. We think this better supports the types of interrelations we were originally trying to foster.

Visitors can leave comments on any page in the Polar Bear Expedition Digital Collections. The team initially considered implementing a wiki-based annotation system which would have allowed users to directly edit collection and item descriptions. This was deemed too problematic in terms of maintaining authoritative descriptive metadata. We also thought about implementing a tagging function. In the end, we settled on a discussion-oriented commenting system that would allow users to contribute substantial pieces and interact, while retaining the archival voice intact. Comments become part of the overall system and are searchable along with all other text on the site.

Visitors have primarily used the comment feature in three ways: information sharing, question asking, and donation inquiries. Between January 2006 and January 2007 29 people posted 62 comments on the site. These are in addition to the comments and biographical information provided in the user profiles. Of those 29 people, 7 also provided information in the user profiles. This overlap indicates that there is a core community interested in the site; however, this community is small thus far and has not yet achieved critical mass. Many members of this community are also members of an active Polar Bear Expedition community that meets to commemorate the event yearly.

We have found that the comments elicit two types of information sharing: additional descriptive data and error correction. As in the user profiles, some visitors provide links to additional information or provide descriptive information about soldiers or images. Other comments have pointed out potential errors. We anticipated that this would happen and ask visitors for documentation. In many cases, we have changed the information in response to the evidence provided. The types of errors include omission of soldiers who served in the campaign, incorrect birth or death dates, and soldiers assigned to the wrong unit. Visitors have also used the comment feature to ask questions. Questions have been specific, asking for details about an individual, or diffuse, asking general historical or research methods inquiries. Finally, the third type of comments is donor inquiries. While we were prepared for correcting errors, we were not prepared for the amount of inquiries about donations or the desire of people to post their digital pictures on our site. Donations posed a problem since the Next Generation Finding Aids Research Project is not a collecting repository. We have referred all of these inquiries to the Bentley Library for them to make the decision about whether the items should be part of the online and/or physical collections.

The comment feature was popular with the survey respondents and interviewees. One noted, "Well I like the fact that I was able to post and say that I was looking for information on my grandfather … and given the engineering regiments 310th company A. That's sort of neat because obviously everyone on this webpage is going to be interested and there could be a great deal of networking that perhaps someone knows something. It's sort of a shot in the dark, if you will, but it's a nice feature to be able to take advantage of other people's expertise that's using the website." (Interview 3, section 224)

Discussion

We began this project very dissatisfied with the current systems that displayed EAD finding aids. In the process of reenvisioning online access tools, we have become more appreciative of the complexity involved in rethinking finding aids; however, we remain convinced that new approaches to visualization and interactive functionalities for researchers are needed. We have learned a substantial amount about manipulating and distilling EAD down to an informative and accessible chunk of information. We also found that creating a browsing structure that aligns with the way people want to use the information makes this a preferred means of navigation. This adds evidence to other research concerning the utilization of search and browse. Our project also provides some hope that finding aids systems without search tools can be made more accessible. For us, the cost of creating data to populate the browsing categories by recoding the EAD was high, but we have identified ways to streamline this process. Finally, we found that the most utilized social navigation feature on the Polar Bear site is the comments. We still believe that social navigation mechanisms do hold great promise for increasing access to archival materials; however, what we have seen is that some (e.g., comments) work better than others. We wonder whether the online community that is a part of the Polar Bear Expedition site has not yet reached a critical mass and therefore the effects of these features are blunted, which has been the case in other online communities [13].

Conclusions

The Polar Bear Expedition Digital Collections is an ongoing project and further evaluation is planned. More long term experiments and projects that explore and evaluate other mechanisms for social interaction such as annotation, ranking, and tagging are needed. Each of the social navigation features described in this article creates a different set of affordances and precludes others. It will not be until we have experimented and studied the multiple options with different types of collections that we will be able to best represent and interpret primary sources to all of the potential audiences in virtual space.

Acknowledgements

The authors thank the staff of the Bentley Historical Library for making the digital images from the Polar Bear Expedition Collections available to us and for consulting with us throughout this project. We also want to acknowledge the contributions of former team members Christie Peterson, Dharma Akmon, Andrew Bangert, and Ricah Marquez.

References

[1] C.J. Prom. User Interactions with Electronic Finding Aids in a Controlled Setting. American Archivist 67/2. pp. 234-268. (2004).

[2] E. Yakel, (2004) "Encoded Archival Description: Are Finding Aids Boundary Spanners or Barriers for Users?" Journal of Archival Organization 2/1-2. pp. 63-77.

[3] Wendy Duff and Verne Harris. Stories and Names: Archival Description as Narrating Records and Constructing Meanings. Archival Science 2/3-4. pp. 263-285 (2002).

[4] M. Light and T. Hyry. Colophons and Annotations: New Directions for the Finding Aid. American Archivist. 65. pp. 216-230. (2002).

[5] Jim Coleman (n.d.) Final Report for the STIM Core Site. URL: http://www.stanford.edu/dept/HPS/sloanconference/CoreSiteReport.html (Last accessed 3 March 2007).

[6] J. Trant, Social Classification and Folksonomy in Art Museums: early data from the steve.museum tagger prototype. Paper presented at the ASIST SIG-CR workshop on Social Classification, 2006. URL: http://www.archimuse.com/papers/asist-CR-steve-0611.pdf (Last accessed 3 March 2007).

[7} I. Anderson. Are you being Served? Historians and the Search for Primary Sources, Archivaria 58. pp. 81-129. (2004).

[8] K. Höök, D. Benyon, & A. Munro. Editors' Introduction: Footprints in the Snow. In Designing Information Spaces: The Social Navigation Approach. K. Höök, D. Benyon, A. Munro eds. London: Springer-Verlag, 1999. pp. 1-16.

[9] C. Olston & E. Chi. ScentTrails: Integrating browsing and searching on the Web. ACM Transactions on Computer-Human Interaction (TOCHI) 10/3: 177 – 197. (2003).

[10] M.A. Katz & M.D. Byrne. Effects of scent and breadth on use of site-specific search on e-commerce Web sites. ACM Transactions on Computer-Human Interaction (TOCHI) 10/3. pp. 198 – 220. (2003).

[11] A. Dieberger, Social Connotations of Space in the Design for Virtual Communities and Social Navigation. Designing information spaces: the social navigation approach. A. Munro, K K. Höök, D. Benyon. eds. London: Springer-Verlag, (1999). pp. 293 - 313 .

[12] D. Goldberg, D. Nichols, B.M. Oki, & D. Terry. Using collaborative filtering to weave an information Tapestry. Communications of the ACM 35/12. pp. 61-70. (1992).

[13] M. Svensson, & K. Höök. Social Navigation of Food Recipes: Designing Kalas. In Designing Information Spaces: The Social Navigation Approach. K. Höök, D. Benyon, A. Munro eds. London: Springer: 201-222. . (2003).

Authors' Biographies

This paper was written by members of the Next Generation Finding Aids Research Team: Elizabeth Yakel is an associate professor at the University of Michigan School of Information (SI). Seth Shaw, site programmer, and Jeremy York are master's students and Magia Krause is a doctoral student at SI. Polly Reynolds, a graduate of SI, is an assistant archivist at the Bentley Historical Library, also on the University of Michigan campus. James Sweeney is an independent computer systems consultant.