INTRODUCTION

The archival community faced a massive back log problem during the past twenty years, to the extent that some archives house more unprocessed (and therefore inaccessible) collections than processed ones. In response, Greene and Messner (2005) proposed a drastic shift in how archival theory and practice toward the concept of “More Product, Less Process,” and minimal processing. I propose extending the MCLP concept to digital libraries and archives as a cost reducing mechanism. This technique prioritizes the collection as a whole or over individual items, specifically regarding metadata. The lack of sufficient metadata for each item limits user access, however incorporating a second archival concept, the participatory archives, could help build a robust lesson for the collections. The participatory archives theory incorporates community building of archives through Web 2.0, such as social tagging.

Social tagging within digital collections without some measure of control could generate too many useless terms, thereby hindering access rather than increasing it. While some suggest digital libraries and archivists could simply approve or disapprove each tag, such a system requires too much oversight. I propose categorizing the users rather than the tags, specifically permitting users who self-identify as subject area experts to tag the collections. The expert users provide more reliable tags, helping the needs of institutions and increasing access to the collections. The dissertation project focuses on the following research questions:

• What are the similarities and differences between tags generated by expert users and novice users in a digital archive?
• How do tags generated by expert and novice users compare with existing metadata in a digital archive?
• How do tags generated by expert and novice users compare with users’ search terms in a digital archive’s query log?

LITERATURE

Since the 1970s, archives saw a tremendous increase in the number of accessioned collections due to a combination of factors including: a rise in the number of collecting repositories; postmodern approaches increasing the breadth of collections; inclusion of manuscript collections within traditional archives; and an increase in both format and amount of records created. While accessions increased, the staff and budgets of most repositories stagnated or began shrinking, thus dramatically slowing the amount of time between accessioning and a fully processed collection. The backlog remained publically inaccessible, and although many varied voices concern over the time, few suggested solutions to the lags. However, in 1989, McCarthy & Lynch (1989) saw the problem remained ignored by most until the introduction of the More Product, Less Process (MCLP) processing model (Greene & Messner, 2005).

MCLP responded a “golden minimum” solution through one simple question: “What is the least we can do to get the job done in a way that is adequate to user needs, now and in the future?” (Greene & Messner, 2005, p. 240). Several universities quickly test the MCLP model through an initial digitizing projects (McCreed, 2006; Mercer Sabre & Hamburger, 2008; Strom, 2005; Weideman, 2006). More recently, some dispute MCLP’s validity (Cox, 2010; Van Ness, 2010) and report on its continued divisive nature within archival circles (Crowe & Spivack, 2010; Trueman, 2010). Although interested in an arrangement and description techniques, Greene (2010) dismisses any concern over MCLP’s application to electronic records, specifically referring to Johnson’s (2007) discussion. Regarding digitalization efforts, Greene finds no justification for an item-level metadata only approach, citing the work at the University of Wisconsin-Oshkosh and the Smithsonian as examples of digital collections with folder-level or series-level metadata.

Web 2.0, Archives, & Digital Collections

The technology Theimer (2011a, 2011b) champions offers archivists increased user engagement. Likewise, Evans (2007) highlighted the perils modern archival situation of significantly increased collection acquisition combined with limited resources and time limitations: suggesting the leveraging of user knowledge through technology to ease the burden. Anderson and Allen (2009) developed the framework for an archival commons which develops additional contextual information through user-generated links, both intra-repository and inter-repository. The links allow users to connect their material needs through virtually rearranging materials, be it chronologically, thematically, or otherwise. The “new” arrangements and links remain publically accessible, assisting other researchers interested in similar topics.

Two seminars work the potential of a wide variety of Web 2.0 tools through a case study and a survey of existing practice within repositories. Krause and Yaked (2007) investigated several Web 2.0 tools used within the Polar Bear Digital Repository and Likewix. Likewise, Boyer, Cheetham, and Johnson (2002) examine using GIS software to manage the City Archives of Philadelphia’s photographic collection. Users can access and view photographs of the city maps, compare the historic images with the modern street view, comment on images, purchase an image, and notify the archives of potential errors. The most direct Web 2.0 style applications within archives remain social tagging and social media, with the former more controversial. While tagging offers both personal user engagement and additional access points, it may also jeopardize the reliability of metadata without authority control.

The minority of in-depth digital collection tagging studies include two major projects: the steve.museum project lead by the Metropolitan Museum of Art (Bearman & Tront, 2006; Tront 2006, 2009a, 2009b & 2009c) and the Library of Congress Flickr project (Springer et al., 2008).

A significant corpus of literature regarding the use of Flickr began developing following the Library of Congress Flickr project. These studies continued exploring the nature of tags (Stevva & Frangman, 2009 & 2010; Chung & Yoon, 2009; Retona, 2010; and Nov Naaman & Ye, 2010). Introduced methodological metrics (Cox, Clough & Siersdorfer, 2011), highlighted case-studies (Ghan, 2010, and Garvin, 2009), explored the experiences of The Commons’ participating institutions (Vaughan, 2010), and compared the tags of the Library of Congress with other Flickr based institutions (Benoit, 2012).

While major tagging projects exist within both the library and museum worlds, the archival community has not produced a similar study. Even at a small scale, only limited literature currently exists. One such study of the Oregon State University Archives on Flickr merely indicates the quantitative information, and does not engage the user’s experience or a linguistically analyze the tags produced (Edmundson-Morton, 2009). Despite the limited applications of social tagging, the existing literature indicates significant potential. Within a controlled context, tags provide users additional access points to the collections. These new access points typically offer perspectives on items not typically included within official metadata, such as general descriptors or more thematic term.

METHODOLOGY

The dissertation project will utilize a sample of materials from the March on Milwaukee Civil Rights History Project at UWM. The sample collection contains four sets of digital objects taken from the Groppi Papers: five anonymous hate mail letters; five anonymous letters critical of the movement; five letters of support; and 15 black and white images. Each set will be created in a component object in CONTENTdm, and contain the following minimal metadata fields: original collection, original location, folder title, finding aid, repository, rights, digital publisher, data digitized, and digital collection.

Two groups of 30 participants (self-identified domain experts and self-identified domain novices) will generate tags for several items and assigns a confidence ranking to each tag (on a scale of 1-5). Additionally, each participant will complete a pre- and post-study questionnaire. The pre-study questionnaire will gather demographic information, self-identification of domain knowledge, and the participants’ prior tagging experience.

The resulting tags will be analyzed through both open coding and a comparison to each item’s existing metadata within the existing digital collection. Additionally, the generated tags will be compared with real users’ query terms from within the query logs of the digital collection. Finally, the tags will be analyzed with both descriptive and inferential statistics (t-tests) to illuminate the variance between sample populations.

CONTACT INFORMATION

School of Information Studies
University of Wisconsin-Milwaukee
Northenry Road, Building B
P.O. Box 413
Milwaukee, WI 53201-0413
414-229-3852 work phone
414-229-6690 fax
www.Edlibonet.com
edlibonet@uwm.edu