

# #MPLP Part 2: Replacing Item-Level Metadata with User-Generated Social Tags

Edward Benoit III

## ABSTRACT

This article explores the potential for integrating and/or supplementing archival description with user-generated tags. The study was a mixed-methods, quasi-experimental design using a sample collection of fifteen documents and fifteen photographs. Sixty participants, divided based on assessed prior domain knowledge, tagged the sample collection with minimal metadata. The generated tags were compared with real-world item-level metadata and query terms. The successful matching of participants' tags with both the unselected metadata and the query terms suggests social tags are an effective additional or supplemental access point to the digital archives.

© Edward Benoit III. 

## KEY WORDS

Social tagging, Participatory archives, MPLP

One of the more exciting aspects of the Web 2.0 movement is the growing popularity of crowdsourcing, or leveraging the wisdom of the crowd, to solve complex problems. Developed from the open source movement, software developers and scientists initially used crowdsourcing for commercial projects such as creating more efficient recommendation algorithms for Netflix and citizen scientist projects such as Galaxy Zoo.<sup>1</sup> Crowdsourcing evolved to include user-generated indexing and social tagging, allowing users to arrange, rearrange, and access information through more personal methods while providing additional access points for other users, and what David Weinberger calls the “third order of order.”<sup>2</sup> The inclusion of user participation within the creation and organization of knowledge alters the perception of professional knowledge and authority, while offering an engagement with users by addressing their personal needs.<sup>3</sup>

The archival community has faced a massive backlog problem over the past twenty years, to the extent that some archives housed more unprocessed, and therefore, inaccessible, collections than processed ones. In response, Mark Greene and Dennis Meissner proposed a drastic shift in both archival theory and practice toward the concept of “More Product, Less Process” or MPLP, and minimal processing.<sup>4</sup> Briefly, MPLP strives toward identifying and implementing a minimal standard level of processing across collections thereby simultaneously decreasing the time required for processing while increasing the number of collections available to users. Minimal processing expanded throughout archival practice, from its origins with arrangement and description to digital archives, resulting in an increase of available collections both physically and digitally.

In his expanded discussion of MPLP, Greene disputed arguments that both born-digital and digitized records require item-level description within their associated metadata.<sup>5</sup> Because users expect and demand more archival records to be digitally accessible, archivists must increase the number of digitized records by “abjuring item-level metadata” and archivists’ “fascination with individual documents.”<sup>6</sup> In rejecting item-level metadata, archivists and institutions reduce costs associated with digital archives creation, which in turn allows the digitization of additional collections. As one practitioner noted, “Every dollar spent to make [online] collections perfect is a dollar we’re not spending to get another collection online and to a larger potential audience.”<sup>7</sup>

A minimally processed digital archives, therefore, identifies the “golden minimum” metadata required to provide user access to the archival materials. This level remains flexible for an entire repository and may move from a series to a subseries to a folder level between collections depending on the collection. For example, folder-level metadata may be more suitable for a correspondence series containing several boxes and dozens of folders of correspondence;

whereas limiting metadata at the series level for a correspondence series containing three folders would still provide adequate access to the digitized records. Following these procedures replicates contemporary archival methods for analog records and thereby allows users an experience similar to physically visiting the archives.

Assuming repositories would apply labor savings from a minimal processing approach toward increasing the number of digitized collections, the MPLP model provides a workable solution for the stagnated and shrinking budgets of modern archives. Additionally, the newly digitized materials may be accessed and used remotely, thereby addressing the rising demands of the twenty-first-century patron. By itself, however, digital archivists' adoption of minimal processing does not take full advantage of content management systems such as OCLC's CONTENTdm, as it mitigates the benefits of increased access points provided through record-level metadata.

Interestingly, David Bearman and Margaret Hedstrom recognized the possibilities of minimal processing and electronic records early, stating:

In electronic records systems, metadata about the records and the configuration of permissions, views, and functions is created and controlled in the active data environment. In principle, this metadata if correctly specified could fully describe and document the records without post-hoc activity by the archivist.<sup>8</sup>

The abandonment of item-level description might better reflect the traditional approaches to description. Allen Benson discussed the nature of early online systems of archival photographs, stating, "Item-level records for the majority of archival photographic materials were not common in early card catalog systems, so consequently there were no item-level records being migrated into first-generation online catalog systems."<sup>9</sup> Several researchers echoed the MPLP approach without explicit mention. Jody DeRidder, Amanda Presnell, and Kevin Walker, for example, saw "human-created item-level metadata," as holding back the number of digitized materials.<sup>10</sup> An OCLC report similarly stated:

Vast quantities of digitized primary materials will trump a few superbly crafted special collections. Minimal description will not restrict use as much as limiting access to those who can show up in person. We must stop our slavish devotion to detail; the perfect has become the enemy of the possible.<sup>11</sup>

Although the MPLP approach to digital archives presents digital surrogates of archival materials in a similar fashion to their use in physical archives, many users (specifically, those without archival research experience) may have difficulties navigating the collection. Burt Altman and John Nemmers found users prefer item-level descriptions and have difficulty following online finding aids

(Christopher Prom provided similar results).<sup>12</sup> As DeRidder, Presnell, and Walker reflected on their decision to abandon item-level description, they stated, “A drawback, however, is that this method of Web delivery may currently be more suitable for scholars than for students.”<sup>13</sup> Furthermore, when looking at the use of archival resources, F. Gerald Ham et al. suggested, “Other user groups may frame questions different from those of historians.”<sup>14</sup>

The minimally processed digital archives could frustrate nontraditional archival users who approach digital archives similarly to other Web-based information retrieval systems. According to Iris Xie, most users “are only willing to devote a small amount of time to evaluate [search] results.”<sup>15</sup> In comparing search result lists and document evaluation, Iris Xie and Edward Benoit recommended providing additional information with search results to support users’ decision-making.<sup>16</sup> With only minimal metadata to guide their evaluations, however, users may either accidentally pass over relevant documents, or slowly evaluate each record regardless of metadata descriptions.

Social tagging within digital collections has gained interest in the past decade, and its inclusion could reintroduce some of the access points lost from a minimal processing approach.<sup>17</sup> Additionally, this framework will help archives deal with the inherent problems of description:

Classification systems, thesauri, and other metadata encoding schemes developed within one worldview do not include the concepts and terms needed to classify and name entities within another. Metadata standards built within continuum frameworks have been designed to support an enduring view of records and their contexts, capturing the dynamic and changing relationships between the multiple entities in the recordkeeping and archiving landscape.<sup>18</sup>

The high costs of creating and maintaining digital archives precluded many archives from providing users with digital content, or increasing the amount of digitized materials. As noted earlier, studies have shown users increasingly demand immediate online access to archival materials with detailed descriptions (access points). The adoption of minimal processing of digital archives limits the access points at the folder or series level rather than at the item-level description users’ desire. User-generated content such as tags could supplement the minimally processed metadata.

This is the second of two articles discussing the results of a mixed-methods, quasi-experimental research project focused on tag generation within a sample minimally processed digital archives.<sup>19</sup> While the first article focused on the potential use of prior domain knowledge as a tag quality assurance mechanism, this article compares the generated tags with item-level metadata and query logs. This article addresses the following research questions and hypotheses:

RQ 1 (a): In what ways do tags generated by expert and/or novice users in a minimally processed collection correspond with metadata in a traditionally processed digital archive?

RQ 1 (b): Does user knowledge affect the proportion of tags matching unselected metadata in a minimally processed digital archives?

H<sub>1</sub>: The proportion of tags matching unselected metadata is affected by the user's domain knowledge.

RQ 2 (a): In what ways do tags generated by expert and/or novice users in a minimally processed collection correspond with existing users' search terms in a digital archives?

RQ 2 (b): Does user knowledge affect the proportion of tags matching query terms in a minimally processed digital archives?

H<sub>2</sub>: The proportion of tag terms matching users' query log terms is affected by users' domain knowledge.

## Literature Review

The participatory archives model engages community members during appraisal, arrangement, and description processes to provide a voice to marginalized communities and increase a sense of empowerment. This concept recently led to new theoretical models of interaction between users and archives. Scott Anderson and Robert Allen, for example, developed the framework for an archival commons, defined as “a space where cultural professionals, researchers, and interested members of the general public could contribute narrative and links among objects of interest held by archives, libraries, and/or museums and systematically reflect those activities within the primary repository itself.”<sup>20</sup>

Andrew Flinn, one of the leading advocates for participatory archives, argued that the interaction between user and record “affect[s] our understanding and knowledge of that archive.”<sup>21</sup> Additionally he argued, “Individual and collaborative scholarship and knowledge production are not completely separate modes of working or thinking; they can co-exist and even interact, informing and extending each other.”<sup>22</sup> Alexandra Eveleigh suggested the participatory archives, through engaging more users, could extend archival advocates essential in the current state of archives.<sup>23</sup> Isto Huvila viewed the participatory archives as a method of decentralizing the authority of archives because “Inclusion and greater participation are supposed to reveal a diversity of motivations, viewpoints, arguments and counterarguments, which become transparent when a critical mass is attained.”<sup>24</sup>

Kate Theimer, one of the leading advocates of technological integration, referred to the movement as Archives 2.0 (reflecting the ideas of Web 2.0 and Library 2.0).<sup>25</sup> By further expanding her discussion, Theimer reviewed the many

features of the 2.0 paradigm including the focus on innovation, flexibility, being technologically savvy, and not becoming obsessed with creating “perfect products.” The technology Theimer championed offers archivists increased engagement with both new and returning users through the use of a variety of Web 2.0 tools, including blogs, wikis, social media, social bookmarking, social tagging, and so on.

The motivation for technologically driven outreach includes an appreciation for the modern limitations of archivists. Max Evans highlighted the perilous modern archival situation of significantly increased collection acquisition combined with fiscal and temporal limitations, suggesting the leveraging of user knowledge through technology to ease the burden.<sup>26</sup> Eric Ketelaar argued for thinking of the archives as “a dynamic open-ended process,” and suggested that archivists must “connect the memories in our archives with the memories in people’s minds” to “make archives into people’s archives.”<sup>27</sup> James Gerencser viewed the interactive nature of Web 2.0 as a better method to reconnect and collaborate with users.<sup>28</sup>

Just as digital archives began altering the archivist/user relationship, Joy Palmer and Jane Stevenson argued Archives 2.0 further moves the relationship away from the traditional one-way toward a more dynamic user-driven approach because “attention is now more focused on direct engagement and active interaction with users in online spaces.”<sup>29</sup>

While many support the Archives 2.0 movement, others raise concerns over the loss of archival authority and the introduction of complexity. As Terry Baxter noted, “Allowing people to interact with information instead of just consuming it can enhance the process, bringing new value to individuals and networks, but it can also muddy the network, reducing authority and authenticity and, perhaps, value. It certainly introduces complexity.”<sup>30</sup> Elizabeth Yakel questioned the balance between user-generated information and the archival authority, and Randall Jimerson highlighted the need to think of “Web 2.0 technology [as] a tool, not a goal.”<sup>31</sup>

In spite of these concerns, Joy Palmer argued for more “risk-taking in respect of crowd-sourcing” and that “new trust metrics and heuristics will emerge.”<sup>32</sup> Furthermore, she called for additional research into the content created by users and how it could be integrated into or supplement archival description. Finally, Palmer stated, “Users should be treated as *peer collaborators*, intrinsic to the process of meaning-making, rather than outside interlopers (however welcome) who must be kept at arm’s length from the authoritative record.”<sup>33</sup> Flinn also defended the movement, arguing, “This need not be seen as an attack on professionalism or scholarship. Rather, non-professional participation in online archival activity provides an opportunity to re-think how future

professionalism and scholarship might be supported in a more collaborative, inclusive and democratic context.”<sup>34</sup>

Although the theoretical developments of Archives 2.0 and postmodernism, as well as their critics, will dictate the future directions of research, the majority of current literature on technology’s use within archival outreach remains within the applied research arena. Taken as both exploratory research and theoretical experimentation, the following case studies and aggregation of data represent the archival vanguard. The sheer breadth of applications indicates the young nature of the field and leaves room for additional research growth.

Two seminal works explore the potential of a variety of tools through a case study and survey of existing practice within repositories. Magia Ghetu Krause and Elizabeth Yakel investigated several Web 2.0 tools and their use with the Polar Bear Expedition Collections, providing users several tools for interacting with the collection, including a bookmarking system, user-generated comments, link paths, user profiles, and the traditional browsing and searching features of digital collections.<sup>35</sup> Krause and Yakel found the intractability of the finding aid, “transforms it from a static to a dynamic document, an ever-changing resource that provides multidirectional knowledge sharing.”<sup>36</sup>

Deborah Boyer, Robert Cheetham, and Mary Johnson discussed the management of the City Archives of Philadelphia’s photographic collection using GIS software.<sup>37</sup> Users can access and view photographs of the city on maps, compare the historic images with the modern street view (using Google Street View), comment on images, purchase an image, and notify the archives of potential errors.

Jodi Allison-Bunnell, Elizabeth Yakel, and Janet Hauck explored which specific metadata elements provide the most helpful information and are most important for researchers.<sup>38</sup> Additionally, the study investigated researchers’ opinions of Web 2.0 tools within digital archives. They found users “almost always wanted more information about collections and items,” and “they wanted as much detail as possible.”<sup>39</sup> This result held true for both textual and nontextual objects alike. Because archivists cannot feasibly describe all digital objects at the item level, “The crucial question becomes not what users want, but what they need.”<sup>40</sup> Regarding Web 2.0 tools, Allison-Bunnell, Yakel, and Hauck discovered that “participants were more interested in taking advantage of information left by other users than in contributing their own information to archival Web sites.”<sup>41</sup> At the same time, users thought the archival Web sites “tended to generate considerably more useful comments than general sites like Flickr or WorldCat,” because of their built-in, more dedicated communities.<sup>42</sup>

In another study, Mary Samoelian analyzed archival Web sites with digital collections and found a number of them relied on Web 2.0 technologies.<sup>43</sup>



Samouelian found from follow-up interviews that, “Participants were overwhelmingly positive about using a Web 2.0 application on their repository Web sites.”<sup>44</sup> The archivists suggested users were “the driving force behind the application” of Web 2.0 tools.<sup>45</sup> Based on her findings, Samouelian viewed Web 2.0 applications as having both strengths and weaknesses. On the one hand, the tools are great for institutional promotion and user engagement; however, the information generated may increase the heavy workload of archivists.<sup>46</sup>

While the Archives 2.0 movement offers significant potential benefits for both users and archivists, only recently have institutions begun integrating or experimenting with these systems. Yakel suggested many archivists remain reluctant to change the traditional model of user/archivist interaction and therefore approach Archives 2.0 with trepidation for its effect on archival work.<sup>47</sup> Research continues testing different approaches for adapting and utilizing Web 2.0 tools within the archives. For example, Michele Christian and Tanya Zanish-Belcher discussed the experience of Iowa State University’s use of YouTube,<sup>48</sup> while others highlighted applications of Flickr,<sup>49</sup> wikis,<sup>50</sup> Second Life,<sup>51</sup> and blogs.<sup>52</sup> Others explored the potential of social media in using primary sources in the classroom,<sup>53</sup> for National History Day research,<sup>54</sup> and for outreach.<sup>55</sup>

This research study is grounded in the minimal processing model and recognizes the contemporary necessity for a minimal approach. Furthermore, the study puts forth a potentially viable solution for the loss of access points within minimally processed digital archives, specifically, the supplementation of folder- or series-level metadata with domain expert user-generated tags. Through its application, this solution may begin moving minimally processed collections back toward the high number of access points previously available through traditional processing techniques.

The participatory archives and Archives 2.0 movements encourage the active role of users within archival description (either officially or supplementally). Allowing users to tag a digital collection enables them to provide their interpretation of archival records and provides additional contextualization for current and future researchers. Additionally, tagging is a dynamic process that develops and alters over time thereby reflecting the ever-changing interpretation of records.

## Social Tagging and Metadata

The internal organization of tags remains a highly debated topic with research indicating a chaotic environment desperately in need of control.<sup>56</sup> Other studies suggest user-generated tags conform to the standards of the



National Information Standards Organization.<sup>57</sup> The problems of using uncontrolled vocabulary remain among the central concerns with either integrating folksonomies into metadata or using them as outright indexes. Kristina Matusiak examined this issue from a practitioner's perspective and reiterated the unsolved access need for images in digital collections.<sup>58</sup> Through her comparison of images in a digital library and on the commercial site Flickr, Matusiak concluded social tagging is not "a simple or miraculous solution to many complex issues inherent in image description."<sup>59</sup> Rather than replacing traditional metadata descriptions of images, she recommended the use of tagging for supplemental descriptions. Agosti et al. explored the integration of user-generated information within a digital library interface as an enhancement of existing metadata.<sup>60</sup>

The growth of Flickr-based research increased tremendously following the 2008 Library of Congress Flickr project.<sup>61</sup> Besiki Stvilia and Corinne Jörgensen explored the use and nature of photosets on Flickr (not including the Commons).<sup>62</sup> Relating to tagging, "The study found that users did not usually tag individual photos and that the photoset or group metadata were often the only metadata associated with those photos."<sup>63</sup> Alternatively, EunKyung Chung and JungWon Yoon related user-generated tags with query terms used for image searches, finding differences within the specificity of tags versus the query terms.<sup>64</sup>

The Flickr-based research continued the trend toward exploration of the nature and similarities/differences between social tags and index terms. Abebe Rorissa, for example, compared tags from Flickr images to the index terms of the University of St. Andrews Library Photographic Archive.<sup>65</sup> He concluded the tags and index terms are significantly different, and should be used in collaboration for retrieval purposes. Oded Nov, Mor Naaman, and Chen Ye explored the nature of the users rather than the tags, finding the long-term users share fewer photos than new users, while providing more tags.<sup>66</sup>

Although the applications of social tagging within digital collections remains limited, the existing research indicates significant potential. Within a controlled context (applying some of the filtering mechanisms discussed earlier), tags give 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 (i.e., color, shape, etc.) or more thematic terms. Systems that allow users to sign in could provide personal tracking of interesting or relevant items within the collections.

## Methodology

As noted in the previous article, the research project relied on a mixed-methods, quasi-experimental research design with multiple data analysis

**Table 1. Research Questions and Associated Data and Analysis**

Research Question	Data Collected	Data Analysis
RQ1a: In what ways do tags generated by expert and/or novice users in a minimally processed collection correspond with metadata in a traditionally processed digital archive?	Tags generated by expert and novice users (at least one tag for 30 items per participant), unselected metadata from March on Milwaukee	Descriptive statistics
RQ1b: Does user knowledge affect the proportion of tags matching unselected metadata in a minimally processed digital archive?	Comparison of generated tags (by group) to unselected metadata (by record) tables, and comparison of generated tags (by group) to unselected metadata (all records) tables	Chi-square tests for association, phi, and Cramer's V
H <sub>1</sub> : The proportion of tags matching unselected metadata is affected by the user's domain knowledge.		
RQ2a: In what ways do tags generated by expert and/or novice users in a minimally processed collection correspond with existing users' search terms in a digital archive?	Tags generated by expert and novice users (at least one tag for 30 items per participant), March on Milwaukee query list extracted from server logs	Descriptive statistics
RQ2b: Does user knowledge affect the proportion of tags matching query terms in a minimally processed digital archive?	Comparison of users' query terms and sample collection metadata/tags table	Chi-square test for association, phi, and Cramer's V
H <sub>2</sub> : The proportion of tag terms matching users' query log terms is affected by users' domain knowledge.		

approaches.<sup>67</sup> Table 1 provides an overview of the data-collection methods and analysis for this article's research questions and hypotheses. This study utilizes a sample digital collection comprised of 15 photographs and 15 documents from the personal papers of James Groppi as included in the digital collection The March on Milwaukee Civil Rights History Project (hereafter called March on Milwaukee). While the existing collection contains item-level metadata, the study's sample collection only presented users with series-level metadata (see Table 2) thereby simulating a minimally processed digital collection. Thirty domain experts and thirty domain novice participants created at least one tag for each of the thirty items in the sample collection. Additionally, participants completed both pre- and postquestionnaires. The first article provides additional detail on the data collection procedures and the sample collection.

## DATA ANALYSIS

The sample digital archives contain a subset of the original metadata in the existing March on Milwaukee digital collection. Addressing RQ1 required a comparison of the generated tags from both experts and novices with a list of the metadata from the existing collection that was not shown to participants;

**Table 2. Sample Collection Minimal Metadata**

Title	Groppi Papers, Correspondence, Hate Mail	Groppi Papers, Correspondence, Support Mail	Groppi Papers, Correspondence, Criticism Mail	Groppi Papers, Photographs
Part of Collection	James Groppi Papers, 1956–1978			
Creator	Groppi, James, 1930–1985			
Type (DCMI)	Text			Image
Original Collection	James Groppi Papers, 1956–1978			
Original Item Location	Milwaukee Mss EX. Box 8, Folders 3–6	Milwaukee Mss EX. Box 1, Folders 1–6	Milwaukee Mss EX. Box 5, Folder 6	PH 4983
Original Item Type	Documents			Photographs
Finding Aid	<a href="http://digital.library.wisc.edu/1711.dl/wiarchives.uw-whs-mil000ex">http://digital.library.wisc.edu/1711.dl/wiarchives.uw-whs-mil000ex</a>			
Repository	Archives / Milwaukee Area Research Center. University of Wisconsin–Milwaukee Libraries			
Digital Publisher	University of Wisconsin–Milwaukee Libraries			
Date Digitized	2010			
Digital Format	image/jp2			
Digital Collection	March on Milwaukee—Civil Rights History Project			
Rights	The Wisconsin Historical Society			

this list is hereafter referred to as unselected metadata. A comparison group of unselected metadata was also generated for each sample record group (document and photograph) including the fields from the following Dublin Core elements: title, creator, subject, description, date, format, identifier, and language. The unselected metadata lists were filtered through a stop list prior to additional analysis as several fields included nondescriptive terms (such as articles). The comparison of unselected metadata and tags considered only exact matches rather than partial or matching word variations. The analysis generated descriptive statistics for each format grouping, highlighting the number and percentage of matching terms, and the number and percentage of new terms for both expert and novice groups.

Although the users' knowledge level was initially assessed on the questionnaire, this information was used only to put the participants into categorical groupings and not to differentiate knowledge levels within groupings during later analysis (e.g., participant 1 is more of an expert than participant 2). Because the independent variables (user knowledge) are, therefore, categorical (or nominal) rather than quantitative, a chi-square test best fit the needs of the research question. A  $2 \times 2$  table chi-square test for association based on the numerical values (number matching and number not matching) tested the following hypothesis:  $H_1$ : *The proportion of tags matching unselected metadata is affected by the user's domain knowledge.*

The researcher also calculated the phi and Cramer's V to analyze the strength of any potential relationships between group type and the number of matching terms. The strength of association test used will be phi since the  $X^2$  analysis was based on a  $2 \times 2$  table.

The data analysis addressing RQ2 followed a similar process to that of RQ1. Rather than looking at format-based groupings, however, this analysis focused on the entire sample collection. The query terms from actual users were parsed out of the existing server-log data and used as a comparison group. Parsing of the server logs resulted in 59,325 unique query terms used to search across all collections hosted by University of Wisconsin–Milwaukee Digital Collections (UWM–DC). Further reduction by collection-specific searches found 1,609 unique query terms used to search the March on Milwaukee collection alone. A list of unique tag terms created by each domain group (expert, novice) and a third list with all unique tag terms created were compared to both query term lists. Additionally, the unique unselected metadata terms were also compared to the March on Milwaukee query term list. The comparisons considered only exact matches rather than partial or matching word variations. The analysis generated descriptive statistics highlighting the number and percentage of matching terms, and the number and percentage of nonmatching terms for expert and novice tags, the combination of expert and novice tags, and the unselected metadata.

Research question 2(b) utilized chi-square tests for association to explore potential relationships between the independent variable and the proportion of tags matching user query terms, the dependent variable. Chi-square tests were selected because the dependent variables were nominal; specifically, matching or not-matching being the dichotomous categories. This analyzed the following hypothesis:  $H_2$ : *The proportion of tag terms matching users' query log terms is affected by users' domain knowledge.*

The researcher also calculated the phi and Cramer's V to analyze the strength of any possible relationships between group type and the number of matching terms. The strength of association test used was phi as the  $X^2$  analysis was based on a  $2 \times 2$  table.

## Results

The following section presents the results of the study pertaining to both research questions beginning with a comparison of the generated tags and the unselected metadata. While the initial section highlights user-generated tags' potential for replacing some item-level description, the second subsection compares the tags with real-world users' query terms. Both sections also discuss the similarities and differences between expert and novice tags.

## RESEARCH QUESTION 1

One of the goals of including user-generated tags as supplemental metadata within a minimally processed digital archives is the potential for replicating or replacing the detailed item-level metadata found in traditionally processed digital archives. The study explores this possibility using a test collection sampled from an existing collection, thereby allowing both the presentation of minimal metadata for the experiment and extracting the full item-level metadata for comparison with the user-generated tags. As noted earlier, the full item-level metadata not included in the minimally processed metadata seen by participants (unselected metadata) were aggregated into two lists (photographs and documents) for comparison with the participant-created tags. Although research question 1(b) tests for an association between prior domain knowledge and the proportion of tags that match the unselected metadata below, it is first important to highlight the ways in which tags generated by both experts and novices in a minimally processed collection correspond with the metadata of a traditional item-level processed digital archives.

The Dublin Core metadata standard remains a primary choice for digital collections due to its flexible interoperable nature. As such, it can also serve as a categorical structure for highlighting the similarities and differences between tags corresponding with existing metadata. The March on Milwaukee uses different combinations of the majority of the 15 Dublin Core elements within its metadata template. Within the Groppi Papers, the existing collection uses the following elements: title, creator, subject, description, publisher, date, type, format, identifier, language, relation, and rights. Table 3 displays the unique field names mapped to Dublin Core elements for both documents and photographs within the existing collection. Several of the fields were included within the minimal metadata provided to participants and are indicated with an asterisk (\*) in the table. Although the title field was included in the minimal metadata, the titles used in the experiment were generalized (e.g., Photograph 1, Support Mail 1, etc.), whereas the existing collection's titles were item-level specific (e.g., James Groppi and Vel Phillips on school bus, circa 1967–1968).

Additional aggregated lists of the so-called unselected metadata, that is the item-level metadata from the existing collection not included in the sample collection used in the experiment, were compiled for 6 Dublin Core elements: title, date, description, subject, identifier, and format. The lists were first made based on format (photograph, document) and then merged into a combined list for comparison with the user-generated tags. Table 4 lists the number of metadata terms within each format and element grouping. The documents did not contain any description or identifier metadata.

The unselected metadata terms were compared to the expert and novice tags initially by format and subsequently as complete sets. Table 5 reports the number

**Table 3. Existing Metadata Template for Groppi Papers**

Dublin Core Element	Unique Field Names	
	Photographs	Documents
Title	Title*	Title*
Creator	Creator*	Creator*
	Photographer	
Subject	Subject	Subject
	Topic	Topic
	Keywords	Keywords
	People	People
	Organization	Organization
	Event	Event
	Place	
Description	Description	
Publisher	Digital Publisher*	Digital Publisher*
Date	Date	Date
	Date Digitized*	Date Digitized*
Type	Type (DCMI)*	Type (DCMI)*
Format	Original Item Type*	Original Item Type*
	Original Item Format	Original Item Format
	Digital Format*	Genre
		Digital Format*
Identifier	Original Item ID	
	Digital ID	
	WHS Image ID	
Language		Language
Relation	Original Collection*	Original Collection*
	Repository*	Original Item Location*
	Digital Collection*	Folder Title
	Part of*	Repository*
		Digital Collection*
Rights	Rights*	Rights*

\* indicates field included in minimal metadata presented to participants

and percentage of matching terms for each format and element grouping. As a whole, the numbers suggest a high level of tags matched the unselected metadata for the title and subject elements, while metadata from the date and format fields did not usually match. Additionally, the identifier metadata never matched across the entire sample collection's tags, suggesting it would be a poor metadata field

**Table 4. Number of Unselected Metadata Terms by Dublin Core Element**

	Title	Date	Description	Subject	Identifier	Format
Photographs	61	7	165	68	38	2
Documents	37	12	0	50	0	5

**Table 5. Number and Percent of Unselected Metadata Terms Matching User-Generated Tags by Dublin Core Element**

	Photographs				Documents			
	Expert		Novice		Expert		Novice	
	#	%	#	%	#	%	#	%
Title	52	85.2%	34	55.7%	28	75.7%	20	54.1%
Date	1	14.3%	1	14.3%	3	25%	3	25%
Description	68	41.2%	44	26.7%	n/a		n/a	
Subject	43	63.2%	36	53%	35	70%	19	38%
Identifier	0	0%	0	0%	n/a		n/a	
Format	0	0%	0	0%	2	40%	2	40%

**Table 6. Percent of Tags Matching Unselected Metadata Unique by Dublin Core Element**

	Photographs		Documents	
	Expert %	Novice %	Expert %	Novice %
Title	36.5%	2.9%	28.6%	0%
Date	100%	100%	0%	0%
Description	39.7%	6.8%	n/a	n/a
Subject	25.6%	11.1%	14.3%	21.1%
Identifier	0%	0%	n/a	n/a
Format	0%	0%	0%	0%

to expect user-generated content to match. This is not surprising as the identifier is typically only known to the repository itself and not generally seen on the digital object. The description field, which only occurs for the photographs, was nearly twice more likely matched with an expert’s tag than with a novice’s.

Although the number of tags matching unselected metadata does illuminate some similarities and differences between expert and novice tags, further comparison requires focusing on the tags themselves. The following section discusses the matching tags for each element set unique to each domain group by format grouping. Table 6 summarizes the percentage of unique matching tags for each domain, format, and element grouping.



The photographs best highlight the difference between expert and novice unselected metadata matching tags. In 4 elements (title, date, description, and subject), both experts and novices provided at least one tag that matched the unselected metadata but was not included in their counterpart's tags. Although both domain groups (expert, novice) created these unique tags, the experts did so at a much higher rate. Within the title element metadata, for example, experts had 52 total tags match unselected metadata with 34 for the novice tags. Of these tags, 33 were duplicated by both experts and novices. The experts' tag set included 19 matching tags not in the novice set, while the novices only created a single additional unique tag. Focusing on the tags themselves, the unique expert tags provided specific information or identification of things within the images, such as St. Boniface, Vel Phillips, and Madison. It is also interesting to note the unselected metadata that was not replicated by any tags included general words, such as "back" or "between," which are difficult to include within tags unless using a compound, multiword, or phrase tag. The title nonreplicated unselected metadata also included date tags (1965, 1966, and 1968) that were difficult for participants to identify within a photograph, given no additional clues. This trend is duplicated with the date-element-specific metadata and the low matching rate. In fact, the 2 matching tags within the date element are the same 2 dates (1969 and 1967), which were unique matching tags within the title element for both experts and novices.

The final 2 elements with tags matching unselected metadata within the photographs, description and subject, offer similar similarities and differences as stated above. Within the description element, both domain groups shared 41 matching tags, with the experts providing 27 additional matching tags and the novices just 3. These unique tags included both specific terms, such as "1967" (novice) and "Wisconsin" (expert) as well as general terms, such as "small" (expert) and "people" (novice). The description element unselected metadata included 188 terms that did not match any tags. Although many of these metadata were again more general in nature, several provided specific information not recognized by the participants, including Bishop Athieliski, Harold Froehlich, and Howard Berliant.<sup>68</sup> Within the subject element, both domain groups shared 32 tags that matched unselected metadata, with novices creating an additional 4 and experts an additional 11 tags. The unique tags echo the previous discussion with specific and general terms. For the subject element, participant tags did not match 21 metadata terms; however, most were rather innocuous, and one could reasonably assume they might be replicated given enough tag development over time (e.g., "activists," "arrests," "courts," "law," etc.).

The trends noted within the photographs do not continue with the document tags. Unlike the photographs, the documents only had unique tags matching unselected metadata within the title and subject elements (all generated by experts). Furthermore, the unique document tags do not provide meaningful additional information. In the title element, for example, experts created eight unique tags

("1," "3," "5," "20," "26," "31," "6," and "June"). Although these look like simple numbers, they are parts of dates used within the titles for the letters. The experts tended to provide the full date ("June 4, 1969"), whereas novices usually provided an abbreviated date ("1969"). Within the subject element, the 5 additional expert tags matching metadata were active terms (e.g., "non-violence," "struggle," etc.), whereas the 4 unique novice tags were more passive descriptive terms (e.g., "whiteness," "relations," etc.). Although these minor differences exist, the participants primarily shared matching tag terms for documents across all elements with 40 title, 3 date, 60 description, and 2 format tags being shared.

The unselected metadata not replicated with the documents continues the trend of the photographs, with limited amounts of key information included within the nonreplicated terms. The format element metadata for both photographs and documents did not match well with participants' tags, with only 2 of a possible 7 terms matching. The lack of replication, in this case, is primarily due to the archival language used to describe formats. The 7 unselected metadata terms ("photographic," "prints," "letters," "manuscripts," "typescripts," "handwriting," "correspondence") were, in fact, all included within the participants' tags but with different expressions. While none of the participants used "typescripts," they did include "typewritten"; likewise for handwriting, where participants did include "handwritten."

As noted earlier, the study explores the potential for replicating/replacing the detailed metadata not included within a minimally processed collection by using a sample collection from an existing collection, thereby allowing a comparison of the users' tags and the unselected metadata. A compiled list of the full metadata for the sample items by format was compared to the minimally processed metadata provided to users. The results created 2 lists of unselected metadata, with the photograph list containing 278 terms and the document list containing 150 terms. The unselected metadata was compared to the lists of unique tags by domain and format, generating a table of matching and nonmatching counts (see Table 7); Figure 1 illustrates these differences.

For both the photographs and the documents, the experts' tags replicated the unselected metadata more than the novices' did. Not surprisingly, however, the highest matching rate for both formats occurred with the combination of experts' and novices' tags. A chi-square analysis of the data was conducted to test whether a statistically significant association existed between the number of matching tags and the user's domain knowledge based on  $H_1$ : *The proportion of tags matching unselected metadata is affected by the user's domain knowledge.*

Individual chi-square tests were run for the photograph and the document data. In both tests, all expected cell frequencies were greater than 5. The photograph test found a statistically significant association between the user's domain knowledge group (expert or novice) and the proportion of tags matching existing metadata,  $\chi^2(1) = 5.386$ ,  $p = .020$ . The association, however, is weak at best,  $\phi =$

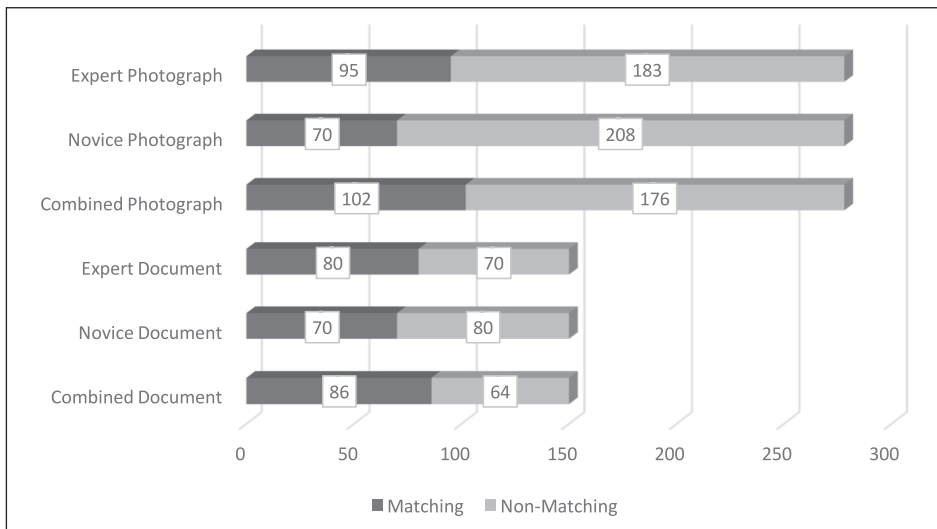
**Table 7. Proportion of Tags Matching Unselected Metadata**

		# Match	% Match	# Non-match	% Non-match
Photographs	Expert	95	34.17%	183	65.83%
	Novice	70	25.18%	208	74.82%
	Combined	102	36.69%	176	63.31%
Documents	Expert	80	53.33%	70	46.67%
	Novice	70	46.67%	80	53.33%
	Combined	86	57.33%	64	42.67%

0.098,  $p = .020$ . The document test, however, did not find a statistically significant association between the user’s domain knowledge group and the proportion of tags matching existing metadata,  $\chi^2(1) = 1.333, p = .248$ . Therefore, the hypothesis is rejected in the case of documents, but accepted for photographs, with the preface that the association is very weak. The weak association indicates that the difference between experts and novices remains quite close. Similar to previous weak associations, increasing the sample size might increase the associative strength.

RESEARCH QUESTION 2

Social tags cannot serve as useful tools if they do not assist with other users’ information retrieval. Similar to the previous research question, the use of a sample from an existing collection provides the necessary data for comparing tags with existing query terms. The Digital Collections at the UWM Libraries provided the query logs for the month of January 2014. Parsing of the server



**FIGURE 1.** Proportions of matching/nonmatching of tags to unselected metadata

**Table 8. Comparison of All Collection Query Terms and Tags**

	# Match	% Match	# Non-match	% Non-match
Expert	575	0.97%	58,750	99.03%
Novice	442	0.75%	58,883	99.25%
Combined	694	1.17%	58,631	98.83%

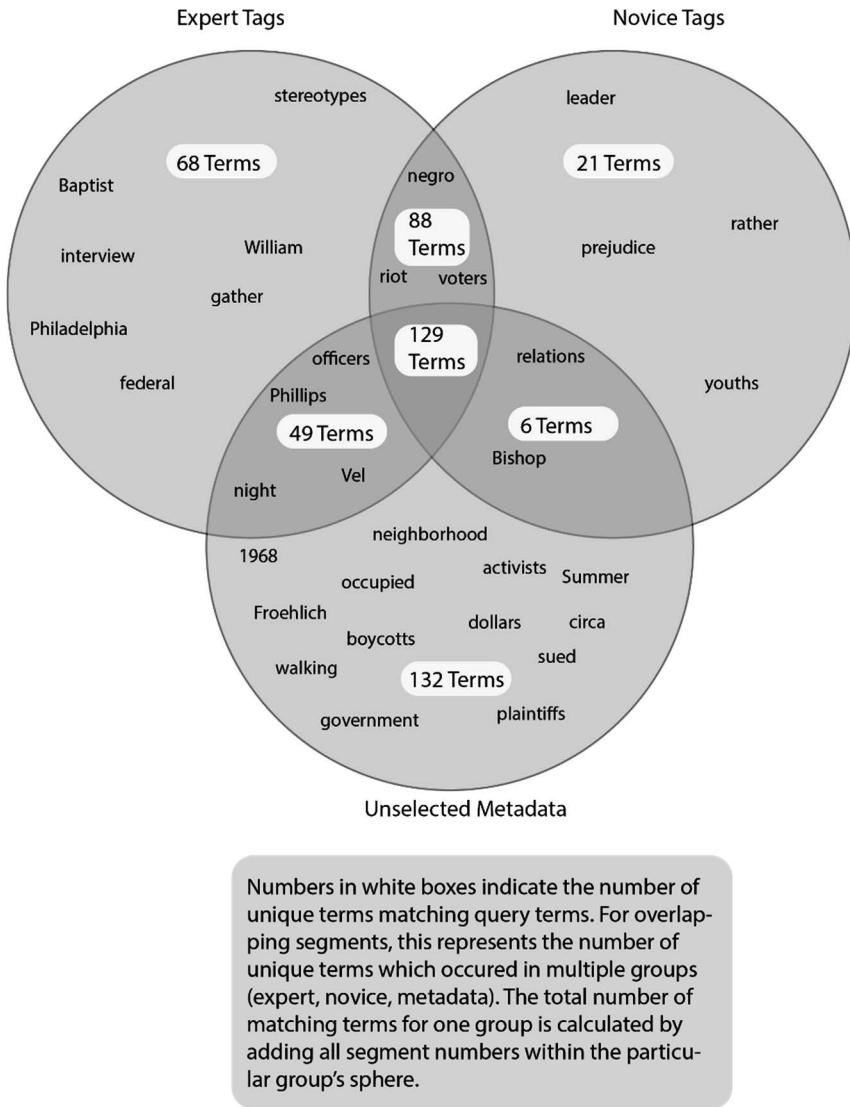
**Table 9. Comparison of March on Milwaukee Query Terms and Tags**

	# Match	% Match	# Non-match	% Non-match
Expert	333	20.70%	1,276	79.30%
Novice	243	15.10%	1,366	84.90%
Combined	360	22.37%	1,249	76.63%
Unselected Metadata	398	24.74%	1,211	75.26%

logs resulted in 59,325 unique query terms used to search across all collections hosted by UWM–DC. Further reduction by collection-specific searches found 1,609 unique query terms used to search the March on Milwaukee collection alone. Tables 8 and 9 display the results of comparisons for both query lists to the unique tag terms created by experts, novices, and both groups combined. Table 9 also includes a comparison with the unselected metadata for both photographs and documents compiled for the previous research question.

An examination of all of the matching tags/metadata terms highlights the relationship between expert tags, novice tags, and metadata terms. Figure 2 illustrates the relationships in a Venn diagram with the number of unique matching terms indicated for each segment and examples of terms found in each segment. The unselected metadata segment of the diagram is used for the unselected metadata grouping; for example, the Venn diagram segment overlapping expert and unselected metadata shows 49 unique terms that matched the query term list occurred within both the expert and unselected metadata lists.

As noted in the middle of the diagram, 129 terms were included in all 3 groups (expert, novice, and metadata). The diagram did not provide enough room for examples of this particular subgrouping. Many of the terms included in all 3 groups describe major themes of the collection as well as key persons or places from the collection. Examples of theme-related terms include: “black,” “bus” or “busing,” “colored,” “demonstration(s),” “housing,” “march” or “marching,” “protest,” “power,” “integration,” “segregation,” “school(s),” and “youth.” Other terms highlight important elements or icons of the photographs, such as “burning” for the image of the Freedom House burning, “fist” for the image of Groppi’s raised fist of resistance, and “wagon” for the image of an arrested Fr. Groppi sitting in a police wagon. Several dates, or parts of dates, appeared in the shared list as



**FIGURE 2.** Tags and unselected metadata matching user query terms

well, including “1966,” “1967,” “December,” “February,” “March,” “May,” “July,” “August,” and “September.” A final characteristic of this subgrouping of terms is the inclusion of key people or places in the photographs and documents. Examples include groups like the “Commandos” and the “NAACP,” important places, such as “Milwaukee” and “Wisconsin,” and authors or subjects of the letters and photographs, such as “Groppi” himself, “LaValle,” “Crooms,” “McKissick,” “Waiss,” and “Waverly.” The inclusion of all of the subgroupings’ terms by experts, novices, and the unselected metadata indicate their importance to both the collection and users’ perception of the collection.

An analysis of the participant-exclusive tags matching user query terms also notes some important themes and potential causality (looking at expert only, novice only, and expert and novice subgroupings combined). Many of the tags are different forms, versions, or conjugations of words found within the metadata terms. Often, it is simply a plural version, such as “newspaper” appearing in the metadata, expert, and novice subgrouping, while “newspapers” is only in the expert subgrouping (additional examples will include associated subgroupings in parenthesis). Additional examples are “youth” (metadata, expert, and novice) and “youths” (novice only), and “group” (metadata, expert, and novice) and “groups” (expert only). More often, however, the tag is a different version, such as “desegregation” (expert and novice) versus “de-segregation” (metadata, expert, and novice). In addition, taking the alterations yet further, some of the participants’ tags conjugate the term “to desegregate” (novice only), creating another variation. Finally, the tags offer abbreviations for terms or phrases, such as “Rev” for Reverend, “feb” for February, or “photos” for photographs.

Although the differences between these tags and the metadata terms appear minor, the matching between user search terms and the alternative variations raises their importance and significance. Modern users have become accustomed to the Google-style search that automatically corrects misspellings and searches multiple tenses, cases, and even derivations of the words, whereas most content management systems for digital collections, such as CONTENTdm, do adjust search terms. The inclusion of the term variations within the query log indicates users are still searching with vernacular, and the participants’ tags also containing similar variations allow for successful matching between tag and query terms.

Additional analysis of the participants’ matching tags not included within the metadata reveals another trend: the importance and/or usefulness of transcription of documents. The vast majority of these tags come from the document tags rather than the photographic tags. Specifically, 102 tags occurred only within the document tag sets and an additional 36 tags occurred within both the photograph and document sets. This represents a combined 78% of the 177 tags that match user query terms but do not match unselected metadata (or 57.6% if excluding the tags also occurring within the photograph sets). When looked at by domain knowledge group, the unique tags created by experts alone or novices alone are consistent with 67.6% and 66.7% respectively (unique tags occurring in both expert and novice groups raises the percentage to 88.6%). Because the document unselected metadata does not include the description Dublin Core element, it also does not contain transcribed information from the documents themselves. The tags, on the other hand, often did come from the document contents, and the above analysis suggests a strong connection between the tags and user search terms.

Expert users' tags match the two query term lists in higher proportions than the novices'; however, the combination of tags outperformed both individual groupings. Chi-square analysis of the data was performed to test for a statistically significant association between users' domain knowledge grouping (expert, novice) and the proportion of tag terms that matched both query-log term lists based on  $H_2$ : *The proportion of tag terms matching users' query log terms is affected by users' domain knowledge.*

Individual chi-square tests were run for the all-collections query list and the March on Milwaukee-specific query list. In both tests, all expected cell frequencies were greater than 5. The all-collections test found a statistically significant association between the user's domain knowledge group and the proportion of tags matching query terms,  $\chi^2(1) = 17.826$ ,  $p < .0005$ .<sup>69</sup> The association, however, is weak at best,  $\phi = -0.012$ ,  $p < .0005$ . The March on Milwaukee-specific test found a statistically significant association between the user's domain knowledge group and the proportion of tags matching query terms,  $\chi^2(1) = 17.128$ ,  $p < .0005$ .<sup>70</sup> The association, however, is weak at best,  $\phi = 0.073$ ,  $p < .0005$ . Both weak association findings replicate issues noted with earlier statistical tests. Although statistical differences exist between experts and novices, the differences are minor with the groups performing close to each other. Increasing the sample size could increase the difference between experts and novices, thereby strengthening the statistical associations.

## Discussion and Conclusion

The study's findings answer calls for additional research into how user-generated content could be integrated into or supplement archival description.<sup>71</sup> The successful matching of participants' tags with both the unselected metadata and the query terms suggests social tags are effective additional or supplemental access points to digital archives. The study also provides theoretical implications based on previous research into social tagging in general and social tagging within archives specifically. The comparison of participants' tags with the unselected metadata and the high degree of successful matches replicate the previous findings of Kipp and Campbell, who found tags often develop the same concepts as traditional indexing, although, in this case, through metadata rather than index terms.<sup>72</sup> The matching of tags with unselected metadata and query-long terms should further alleviate archivists' concern over user reliability.

The comparison of generated tags with the unselected metadata and query terms demonstrates the benefits of including both expert and novice tags. The proportion of unselected metadata and query terms matching expert tags was higher than that matching novice tags. The combination of experts and



novices, however, provided an even higher percentage, thereby demonstrating the strength of incorporating both sets of tags into a collection. Additionally, since the study did not include intermediate users' tags (as is discussed later), the combination of all three might be even higher.

Another anticipated benefit is the potential for tags to replicate the unselected portions of traditional item-level metadata. The findings do not indicate a high level of replication of unselected metadata from either experts or novices. Even the combination of experts and novices did not produce more than 57% replication. This suggests the integration of tagging and minimal processing cannot completely replace the traditional item-level description/metadata of digital archives. In practical terms, repositories considering allowing user tagging must be clear with their expectations and understand that tagging results in a different type of description.

Although the tags do not replicate the unselected metadata, they do serve as access points to the collection. Similar to previous points, the experts' tags again scored higher than novices', with the combination of both groups exceeding the individual groupings. A comparison of the proportion of March on Milwaukee query terms that match generated tags with those matching the unselected metadata shows a similar level (22.37% for tags, 24.74% for traditional metadata). This suggests the lack of matching unselected metadata is not as important when considering the terms users actually use for searching the collection. In this case, the tags provide similar access to the collection as that provided by their traditional metadata counterpart. Additionally, while the metadata terms in a collection are static, the number of unique tags would likely grow over time, thereby increasing the likelihood of query terms matching tags to overtake the full metadata rates.

The study's findings provide practical implications for metadata creation, specifically by increasing the quality and breadth of metadata in a collection. Participants created many tags that matched the real-world user query terms, but did not match the unselected metadata. This implies users are searching for terms not included within the standard metadata corpus. Although users will always search for terms not found within a collection, the matching tags indicate the need to increase access points to the collection to serve users' searching behavior. The documents in particular would benefit from additional content-driven or transcription-like metadata as those types of tags comprised the largest portion of the additional tags matching the query-log terms.

As noted earlier, the real-world metadata for the documents did not include the description Dublin Core element thereby leaving a significant deficiency within the item-level metadata. The tags matching query terms but not the unselected metadata would fill the description element well. A repository could use tag and query analyses to identify metadata gaps in both minimally

processed and traditionally processed collections and develop new targeted strategies for filling the gaps. Finally, while tags may not entirely replace item-level metadata, they do provide enough coverage to question the need for the labor-intensive practice of item-level description.

## LIMITATIONS AND FUTURE DIRECTIONS

Research studies inherently contain limitations through their design or analysis. While participants interacted with the sample digital collection in near-real-world conditions, the variable controls necessary for a quasi-experimental design precluded using a completely natural environment. Additional research following a more natural approach in a longitudinal study remains necessary for confirmation of the current study's findings. Likewise, future research should expand the scope of formats used in participatory archival research to include recorded audio and moving image materials. Finally, tagging enticement remains another outstanding issue with the use of social tagging in archives. While this study paid participants, monetary enticement is not a sustainable direction for participatory archives. Future studies should explore nonmonetary approaches to increase participation in tagging, commenting, and other user-based description techniques.

## NOTES

- <sup>1</sup> David Weinberger, *Too Big to Know: Rethinking Knowledge Now that the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room* (New York: Basic Books, 2011).
- <sup>2</sup> David Weinberger, *Everything Is Miscellaneous: The Power of the New Digital Disorder* (New York: Times Books, 2007).
- <sup>3</sup> Charles Leadbeater and Debbie Powell, *We-Think: Mass Innovation, Not Mass Production* (London: Profile Books, 2008); Michael Lewis, *Next: The Future Just Happened* (New York: W. W. Norton & Company, 2001); Mirko Tobias Schäfer, *Bastard Culture! User Participation and the Extension of Cultural Industries* (Amsterdam: Amsterdam University Press, 2011); Clay Shirky, *Cognitive Surplus: Creativity and Generosity in a Connected Age* (New York: Penguin Press, 2010); Weinberger, *Everything Is Miscellaneous*; Weinberger, *Too Big to Know*.
- <sup>4</sup> Mark A. Greene and Dennis Meissner, "More Product, Less Process: Revamping Traditional Archival Processing," *The American Archivist* 68, no. 2 (2005): 208–63.
- <sup>5</sup> Mark A. Greene, "MPLP: It's Not Just for Processing Anymore." *The American Archivist* 73, no. 1 (2010): 192.
- <sup>6</sup> Mark A. Greene, "Doing Less Before It's Done Unto You: Reshaping Workflows for Efficiency before the Wolf Is at the Door," *RBM: A Journal of Rare Books, Manuscripts and Cultural Heritage* 12, no. 2 (2011): 101.
- <sup>7</sup> Joshua Ranger, "More Bytes, Less Bite: Cutting Corners in Digitization," 2008, <http://www.archivists.org/conference/sanfrancisco2008/docs/session701-ranger.pdf>.
- <sup>8</sup> David Bearman and Margaret Hedstrom, "Commentary Reinventing Archives for Electronic Records: Alternative Service Delivery Options," in *Electronic Records Management Program Strategies*, ed. Margaret Hedstrom (Pittsburgh: Archives & Museum Informatics, 1993), 87.

- <sup>9</sup> Allen C. Benson, "The Archival Photograph and Its Meaning: Formalisms for Modeling Images," *Journal of Archival Organization* 7, no. 4 (2009): 169.
- <sup>10</sup> Jody DeRidder, Amanda Presnell, and Kevin Walker, "Leveraging Encoded Archival Description for Access to Digital Content: A Cost and Usability Analysis," *The American Archivist* 75, no. 1 (2012): 144.
- <sup>11</sup> Ricky Erway and Jennifer Schaffner, *Shifting Gears: Gearing Up to Get into the Flow*, 2nd ed. (Dublin, Ohio: OCLC Programs and Research, 2017), doi:10.25333/C3159X.
- <sup>12</sup> Burt Altman and John R. Nemmers, "The Usability of On-Line Archival Resources: The Polaris Project Finding Aid," *The American Archivist* 64, no. 1 (2001): 121–31; Christopher J. Prom, "User Interactions with Electronic Finding Aids in a Controlled Setting," *The American Archivist* 67, no. 2 (2004): 234–68.
- <sup>13</sup> DeRidder, Presnell, and Walker, "Leveraging Encoded Archival Description for Access to Digital Content."
- <sup>14</sup> F. Gerald Ham et al., "Is the Past Still Prologue?: History and Archival Education," *The American Archivist* 56, no. 4 (1993): 718–29.
- <sup>15</sup> Iris Xie, *Interactive Information Retrieval in Digital Environments* (Hershey, Penn.: IGI Pub., 2008).
- <sup>16</sup> Iris Xie and Edward Benoit III, "Search Result List Evaluation versus Document Evaluation: Similarities and Differences," *Journal of Documentation* 69, no. 1 (2013): 49–80.
- <sup>17</sup> Maristella Agosti et al., "Annotation as a Support to User Interaction for Content Enhancement in Digital Libraries," in *Proceedings of the Working Conference on Advanced Visual Interfaces, AVI '06* (New York: ACM, 2006), 151–54, <http://doi.acm.org/10.1145/1133265.1133296>; David Bearman and Jennifer Trant, "Social Terminology Enhancement through Vernacular Engagement: Exploring Collaborative Annotation to Encourage Interaction with Museum Collections," *D-Lib Magazine* 11, no. 9 (2005), <http://www.dlib.org/dlib/september05/bearman/09bearman.html>; Krystyna K. Matusiak, "Towards User-Centered Indexing in Digital Image Collections," *OCLC Systems & Services: International Digital Library Perspectives* 22, no. 4 (2006): 283–98; Michelle Springer et al., *For the Common Good: The Library of Congress Flickr Pilot Project* (Washington, D.C.: The Library of Congress, 2008), [http://www.loc.gov/rr/print/flickr\\_report\\_final.pdf](http://www.loc.gov/rr/print/flickr_report_final.pdf); Jennifer Trant, "Exploring the Potential for Social Tagging and Folksonomy in Art Museums: Proof of Concept," *Art Museums: Proof of Concept. New Review of Hypermedia and Multimedia* 12, no. 1 (2006): 83–105; Helena Zinkham and Michelle Springer, "Taking Photographs to the People: The Flickr Commons Project and the Library of Congress," in *A Different Kind of Web: New Connections between Archives and Our Users*, ed. Kate Theimer (Chicago: Society of American Archivists, 2011), 102–15.
- <sup>18</sup> Frank Upward, Sue McKemmish, and Barbara Reed, "Archivists and Changing Social and Information Spaces: A Continuum Approach to Recordkeeping and Archiving in Online Cultures," *Archivaria* 72 (2011): 230.
- <sup>19</sup> Edward Benoit III, "#MPLP Part 1: Comparing Domain Expert and Novice Social Tags in a Minimally Processed Digital Archives," *The American Archivist* 80, no. 2 (2017): 145–76.
- <sup>20</sup> Scott R. Anderson and Robert B. Allen, "Envisioning the Archival Commons," *The American Archivist* 72, no. 2 (2009): 383. Emphasis original.
- <sup>21</sup> Andrew Flinn, "An Attack on Professionalism and Scholarship? Democratising Archives and the Production of Knowledge," *Ariadne* 62 (2010), <http://www.ariadne.ac.uk/issue62/flinn>.
- <sup>22</sup> Flinn, "An Attack on Professionalism and Scholarship?"
- <sup>23</sup> Alexandra Eveleigh, "Welcoming the World: An Exploration of Participatory Archives" (presented at the International Conference on Archives, Brisbane, Australia, 2012), <http://ica2012.ica.org/files/pdf/Full%20papers%20upload/ica12Final00128.pdf>.
- <sup>24</sup> Isto Huvila, "Participatory Archive: Towards Decentralised Curation, Radical User Orientation, and Broader Contextualisation of Records Management," *Archival Science* 8, no. 1 (2008): 25.
- <sup>25</sup> Kate Theimer, ed., *A Different Kind of Web: New Connections between Archives and Our Users* (Chicago: Society of American Archivists, 2011).
- <sup>26</sup> Max J. Evans, "Archives of the People, by the People, for the People," *The American Archivist* 70, no. 2 (2007): 397.
- <sup>27</sup> Eric Ketelaar, "Cultivating Archives: Meanings and Identities," *Archival Science* 12, no. 1 (2012): 29; and Eric Ketelaar, "Being Digital in People's Archives," *Archives & Manuscripts* 31, no. 2 (2003): 12–13.

- <sup>28</sup> James Gerencser, "New Tools Equal New Opportunities: Using Social Media to Achieve Archival Management Goals," in *A Different Kind of Web*.
- <sup>29</sup> Joy Palmer and Jane Stevenson, "Something Worth Sitting for? Some Implications of Web 2.0 for Outreach," in *A Different Kind of Web*, 6.
- <sup>30</sup> Terry D. Baxter, "Going to See the Elephant: Archives, Diversity, and the Social Web," in *A Different Kind of Web*, 286.
- <sup>31</sup> Elizabeth Yakel, "Balancing Archival Authority with Encouraging Authentic Voices to Engage with Records," in *A Different Kind of Web*; and Randall C. Jimerson, "Archives 101 in a 2.0 World: The Continuing Need for Parallel Systems," in *A Different Kind of Web*, 305.
- <sup>32</sup> Joy Palmer, "Archives 2.0: If We Build It, Will They Come?," *Ariadne* 60 (2009), <http://www.ariadne.ac.uk/issue60/palmer>.
- <sup>33</sup> Palmer, "Archives 2.0: If We Build It, Will They Come?" Emphasis original.
- <sup>34</sup> Flinn, "An Attack on Professionalism and Scholarship?"
- <sup>35</sup> Magia Ghetu Krause and Elizabeth Yakel, "Interaction in Virtual Archives: The Polar Bear Expedition Digital Collections Next Generation Finding Aid," *The American Archivist* 70, no. 2 (2007), 282–314.
- <sup>36</sup> Krause and Yakel, "Interaction in Virtual Archives," 308.
- <sup>37</sup> Deborah Boyer, Robert Cheetham, and Mary Johnson, "Using GIS to Manage Philadelphia's Archival Photographs," *The American Archivist* 74, no. 2 (2011), 652–63.
- <sup>38</sup> Jodi Allison-Bunnell, Elizabeth Yakel, and Janet Hauck, "Researchers at Work: Assessing Needs for Content and Presentation of Archival Materials," *Journal of Archival Organization* 9, no. 2 (2011), 67–104.
- <sup>39</sup> Allison-Bunnell, Yakel, and Hauck, "Researchers at Work," 86.
- <sup>40</sup> Allison-Bunnell, Yakel, and Hauck, "Researchers at Work," 87.
- <sup>41</sup> Allison-Bunnell, Yakel, and Hauck, "Researchers at Work," 92.
- <sup>42</sup> Allison-Bunnell, Yakel, and Hauck, "Researchers at Work," 93.
- <sup>43</sup> Mary Samouelian, "Embracing Web 2.0: Archives and the Newest Generation of Web Applications," *The American Archivist* 72, no. 1 (2009): 58.
- <sup>44</sup> Samouelian, "Embracing Web 2.0," 62.
- <sup>45</sup> Samouelian, "Embracing Web 2.0," 63.
- <sup>46</sup> Samouelian, "Embracing Web 2.0," 65.
- <sup>47</sup> Elizabeth Yakel, "Inviting the User into the Virtual Archives," *OCLC Systems & Services: International Digital Library Perspectives* 22, no. 3 (2006): 159.
- <sup>48</sup> Michele Christian and Tanya Zanish-Belcher, "'Broadcast Yourself': Putting Iowa State University's History on YouTube," in *A Different Kind of Web*, 33–41, <http://works.bepress.com/mchristian/10>.
- <sup>49</sup> Patrick Peccatte, "Liberating Archival Images: The PhotosNormandie Project on Flickr," in *A Different Kind of Web*, 148–58.
- <sup>50</sup> Michele Combs, "Wikipedia as an Access Point for Manuscript Collections," in *A Different Kind of Web*, 139–47; Guy Grannum, "Harnessing User Knowledge: The National Archives' Your Archives Wiki," in *A Different Kind of Web*, 116–27; Amy Schindler, "A New Look for Old Information: Creating a Wiki to Share Campus History," in *A Different Kind of Web*, 191–202.
- <sup>51</sup> Mattie Taormina, "The Virtual Archives: Using Second Life to Facilitate Browsing and Archival Literacy," in *A Different Kind of Web*.
- <sup>52</sup> Malinda Triller, "Double-Duty Blogging: A Reference Blog for Management and Outreach," in *A Different Kind of Web*, 203–12.
- <sup>53</sup> Jeffrey W. McClurken, "Waiting for Web 2.0: Archives and Teaching Undergraduates in a Digital Age," in *A Different Kind of Web*, 243–54.
- <sup>54</sup> Tobi Voigt, "Is National History Day Ready for Web 2.0?," in *A Different Kind of Web*, 233–42.
- <sup>55</sup> Jessica Lacher-Feldman, "Making Friends and Fans: Using Facebook for Special Collections Outreach," in *A Different Kind of Web*, 54–64; Medina-Smith, "Going Where the Users Are: The Jewish Women's Archive and Its Use of Twitter," in *A Different Kind of Web*.
- <sup>56</sup> Pauline Rafferty and Rob Hilderley, "Flickr and Democratic Indexing: Dialogic Approaches to Indexing," *Aslib Proceedings* 59, nos. 4–5 (2007): 397–410.

- <sup>57</sup> Louise F. Spiteri, "The Structure and Form of Folksonomy Tags: The Road to the Public Library Catalog," *Information Technology and Libraries* 26, no. 3 (2013): 13–25.
- <sup>58</sup> Matusiak, "Towards User-Centered Indexing in Digital Image Collections," 286.
- <sup>59</sup> Matusiak, "Towards User-Centered Indexing in Digital Image Collections," 294.
- <sup>60</sup> Agosti et al., "Annotation as a Support to User Interaction for Content Enhancement in Digital Libraries."
- <sup>61</sup> Springer et al., *For the Common Good: The Library of Congress Flickr Pilot Project*; Zinkham and Springer, "Taking Photographs to the People: The Flickr Commons Project and the Library of Congress."
- <sup>62</sup> Besiki Stvilia and Corinne Jörgensen, "User-Generated Collection-Level Metadata in an Online Photo-Sharing System," *Library & Information Science Research* 31, no. 1 (2009), 54–65.
- <sup>63</sup> Stvilia and Jörgensen, "User-Generated Collection-Level Metadata in an Online Photo-Sharing System," 64.
- <sup>64</sup> EunKyung Chung and JungWon Yoon, "Categorical and Specificity Differences between User-Supplied Tags and Search Query Terms for Images," *Information Research* 14, no. 3 (2009), <http://www.informationr.net/ir/14-3/paper408.html>.
- <sup>65</sup> Abebe Rorissa, "A Comparative Study of Flickr Tags and Index Terms in a General Image Collection," *Journal of the American Society for Information Science and Technology* 61, no. 11 (2010), 2230–42.
- <sup>66</sup> Oded Nov, Mor Naaman, and Chen Ye, "Analysis of Participation in an Online Photo-Sharing Community: A Multidimensional Perspective," *Journal of the American Society for Information Science and Technology* 61, no. 3 (2010): 555–66.
- <sup>67</sup> Benoit, "#MPLP Part 1."
- <sup>68</sup> Harold Froehlich was a state representative standing next to Fr. Groppi in a photograph, and Howard Berliant was a photographer of one of the images in the sample collection.
- <sup>69</sup> The p-value is 0.000024.
- <sup>70</sup> The p-value is 0.000035.
- <sup>71</sup> Jimerson, "Archives 101 in a 2.0 World: The Continuing Need for Parallel Systems."
- <sup>72</sup> Margaret E. I. Kipp and D. Grant Campbell, "Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices," in *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*, 2006, <http://eprints.rclis.org/archive/00008315/>.

## ABOUT THE AUTHOR

---



Edward Benoit III is assistant professor at the School of Library and Information Science at Louisiana State University. He is the coordinator of both the archival studies and cultural heritage resource management MLIS specializations. He received an MA in history, an MLIS, and a PhD in information studies from the University of Wisconsin–Milwaukee. His research focuses on participatory and community archives, nontraditional archival materials, and archival education. He is the founder and director of the Virtual Footlocker Project that examines the personal archiving habits of twenty-first-century soldiers in an effort to develop new digital capture and preservation technologies to support their needs.