Defining relationships between Social Interaction and Discoverability of Digital Resources in Memory Institutions

Õnne Mets

University of Milano-Bicocca, Milan, Italy

o.mets@campus.unimib.it

Abstract. This paper gives an overview of a research project aimed to compare and clarify the relations between social interaction and discoverability of digital resources in three types of memory institutions: libraries, archives and museums. The main research questions are: how and to what extent social interaction is related to discoverability, what are the similarities and differences across the three types of institutions within this context, and what is the role of the academic community in the process? The multi-method empirical research enhances three case studies: a library, an archive and a museum from a single country context, United Kingdom. Each case study maps the phenomenon of crowdsourcing within the organization, then the social interaction emerged in Twitter will be described on national level, and the effect will be estimated and evaluated taking into account qualitative data. Comparative analysis between the organizations concludes the research cycle. The results are expected to contribute to effective design and maintenance of a discoverable digital library, sustainable user engagement, and collaboration between the organizations.

Keywords: digital libraries, digital repositories, libraries, museums, archives, discoverability, participation, user engagement, interaction, crowd sourcing.

1 Introduction

The current paper describes an interdisciplinary research project aimed to compare and clarify the relations between social interaction and discoverability of digital resources in three types of memory institutions: libraries, archives and museums.

The project includes social sciences into the digital library research at the time when digital libraries are more and more developed as socio-technical systems, and not merely seen as a new technology or organizational form but a change in the social and material bases of knowledge work and the relations among people who use and produce information artifacts and knowledge [1]. Research on social aspects of digital libraries is often in the context of user-centered design, work practice studies, the social web and other topics related to specific projects or programs instead of how users or other players are used in fulfilling the roles of digital libraries or how the

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

society can contribute to the social roles like fostering and enhancing collaboration and partnerships among and across individuals, institutions, groups, and domains of education, research or commerce [2].

The emergence of an increasing participatory culture is evident in which individuals contribute to the creation of information, knowledge and cultural artifacts through different modes of collaboration in the digital sphere [3:105; 4]. At the same time established institutions continue to try to find the best way to appropriate the technological affordances of the Internet [5]. Libraries, archives and museums overcome their limits and collaborate in making their collections jointly available online [6], see for example Europeana.eu [7]. Also the need for new methodologies is seen by which public institutions can access and maintain digital knowledge resources in manners that directly consult stakeholder communities [8].

In this context the paper discusses next the background research in Section 2. The research problem and questions are stated in Section 3. Section 4 explains the research design: sample, data collection techniques and methods of analysis. Section 5 states some preliminary findings from the first case study, British Library. Section 6 adds concluding remarks about the project.

2 Background research

2.1 Defining discoverability

According to the English language dictionary discoverability is "The quality of being discoverable; capability of being found out" [9:753]. Similarly "A SAGE White Paper on Collaborative Improvements in the Discoverability of Scholarly Content" [10:3] refers to the quality of "being found" by defining discoverability as "the description or measure of an item's level of successful integration into appropriate infrastructure maximizing its likelihood of being found by appropriate users".

First the definition from the white paper states the goal to find, not necessarily access the materials. Thus discoverability can be related to physical as well as online collections, but in this project we look at the digital resources only. Yet even in case of this limitation, the discovery may happen elsewhere, e.g. in Google Scholar. A great deal of information-seeking for academic content has moved to search engines, academic or discipline-specific databases and aggregations [2:132- 133].

Secondly, discoverability is related to the content, not to the collection exposed from the digital library. In specific, the definition puts discoverability in relation to an item. This in turn has detailed implications in terms of discovery tools. E.g. MARC records provide for the online discovery of archives and manuscript collections at the collection level, not on an item level [2:5-6].

Thirdly, by stating "...being found by appropriate users" the definition states the relation with users, and not with machines or robots.

Thus discovery is defined as "the process and infrastructure required for a user to find an appropriate item" [10:3]. The process of discovery may be supported by a discovery service, which is "a single interface, providing integrated access to the mul-

tiple information resources (catalogs, publishers' e-book and e-journal collections, subscription databases, archival collections) to which a library has rights". Discovery service uses consolidated subject indexing and metadata, and search results are generally deduped and relevance ranked, e.g. EBSCO Discovery Service [11].

2.2 Practice of organizations

Metadata is considered as a key feature to discoverability [12; 13]. The better the quality of metadata, the better the discoverability. But traditional indexing techniques are costly and labor-intensive and even practitioners are not sure whether they provide the only or best way to meet user needs [14] or are adequate for online resource discovery [15].

Libraries, archives and museums use the same tool, i.e. improving discoverability of items for attacking different problems.

Archivists organize records by collections, but an outreach strategy that limits visibility-raising efforts solely to the collection level is restricted in its ability to reach numerous potential digital patrons. Therefore engaging the users to add item-specific information would raise discoverability of the items [16].

Museum collections online have not become as engaging as they might for the general public and the reason might be the mismatch between keywords attributed to the items and those in use by public. Social tagging appeals to museums because it embodies self-directed learning philosophies by being in a dialog between the viewer and the work, and the viewer and the museum, and a user's assertion that a work of art is about something [17]. Additionally museums face a lot of non-textual items and acknowledge the potential of user-generated tagging in image indexing [14].

Libraries seem to have an opposite problem of having mostly textual works. Within the large amount of available content provided not only with metadata, but also with text recognition, the users might need human created links between the items in order to find what they are looking for amongst the masses of options. As libraries digitize massively, the increase of usage of these items would partly justify the costs and efforts for digitization [18].

In practice, in order to improve discoverability, often visibility of the items is being increased first. Visibility "involves placing information in locations where people will come across it in the work that they do" [19]. Embedding and linking from high-traffic sites like Flickr or Wikipedia can raise visibility as well as awareness and usage [2:165]. Also digital libraries that are crawled and indexed by common or academically oriented search engines are discoverable in search engine results as if they were aggregated [2:24].

A study of the users of digitized Hague Sheet Music shows that the users are often interested in specific songs and songwriters. Their discovery of assets in Hague Sheet Music collection via Wikipedia articles about specific songs, songwriters, and lyricists supports this characterization. When attempting to connect with potential digital patrons whose web searches are conducted at this level of specificity, archivists can achieve greater success in generating collection use by using Wikipedia to connect users with digital archival materials at the item level [16]. The University of Houston

was primarily interested in contributing visual images to Wikipedia to accompany already existing articles, and were successful in terms of referrals to them [18].

An Italy-focused study on Twitter found bigger museums to post more original content and applying more vertical communication model, whereas smaller museums are mostly re-tweeters, distribute not necessarily museum-related content, and are building wider networks around them [20]. A study on interactions between museums in Twitter found the main pattern of relationship being local, possibly mimicking offline relationships. The main criterion for explaining community structure was country, not language, leaving a secondary role of interaction patterns to topics [21].

3 The research problem and questions

We see in parallel two ongoing processes: 1) increased online interaction among people, and 2) collections of memory institutions becoming digital. In the first case the online communication and interaction, including the usage of social network sites has been in the focus of researchers. In the second case we witness that the triumph of online publishing of scholarly articles as well as digitization in the memory institutions has led to a continuous increase in the importance of digital libraries as a gateway and an archive. The more users come to digital repositories the more the systems are improved, which in turn aim to increase the usage again. More often the organizations recognize the need to engage users in this process.

Social interaction relates to the definition of communication, where at least two interacting agents share a common set of signs and a common set of rules [22:47]. Here 'social interaction' is used whenever users of digital repositories interact in a way, which is instantly or eventually (e.g. after administrator's approval) visible to other user(s).

Social interaction may emerge in metadata enrichment, collaborative indexing or tagging, donation of materials, recommendations, rankings, asking, sharing, commenting or other ways. These actions contribute to a variety of topics, like design and architecture, visibility, use and reuse of the content or data, usability, usefulness, discoverability, feedback, evaluation etc. [4; 23; 24; 25; 26, 27; 28]. Given that we face a broad area, and users help each other either directly or indirectly to discover digital content, this research is to study one aspect of the cultural phenomenon – the relations between social interaction and discoverability of digital resources.

The aim of the research is to compare and clarify the relation between social interaction and discoverability of digital resources. The core research questions are: How and to what extent are social interactions relevant to the discoverability of digital resources of memory institutions? What are the similarities and differences across the three types of memory institutions within this context? What is the role of the academic community in it?

The results of the research may help the organizations to become more efficient, and effective in designing and/or maintaining a discoverable digital library, sustainable user engagement, and collaboration between the organizations. This in turn may affect positively the closely connected areas, like visibility, use and reuse of the con-

tent. The findings of the research potentially contribute also to the knowledge on social interaction in a digital environment. Hypothetically the academic user community is partly overlapping across the three types of memory institutions, and expects the best practice examples to be replicated across the institutions. Presumably also the offline collaborations play a role in online occurrences.

4 The research design

The approach of the current project is based on three case studies, including a wider quantitative and a compact qualitative inquiry for substantial context. The sample is picked from a single country in order to be comparable by political and cultural context, and enabling an inference in the end.

The sample procedure of organizations includes one of each type of memory institution (a library, an archive and a museum), who is active in research and development of digital libraries and user engagement in the United Kingdom. The country has a high digital economic and social performance (according to DESI index [29]), rating high in social media usage by institutions as well as individuals.

The sample of users focuses to the academic community located at the same urban space as the organizations in order to detect possible offline impact. Members of the academic community have similar information behavior and can potentially be users of either three organizations. The sample group consists of 5 people per organization (may be overlapping), who are the users of at least one case study organization and volunteer to take part in the survey. Employed by a snowball technique, the sample will be biased, but in compliance with already mentioned distinctive characteristics.

Each case study goes through three phases: mapping the phenomenon, describing the process, and estimating the effect. The comparative analysis between the organizations throughout the three phases concludes the research design cycle.

4.1 Mapping the phenomenon

At first document analysis will be conducted for relevant documents and public interfaces of the case study organizations. By this we inquire about the context of the organizations; the environments and tools, how they have enabled the emergence of social interaction related to their collections.

Secondly, data analysis will be run for the environments (1-2 per a case), where the social interaction relates to discoverability. This insight inquires, if and how discoverability is improved quantitatively (participation patterns in total) and qualitatively (meaningfulness of contributions).

4.2 Describing the process

At first, we place the three case study organizations into the broader context of memory institutions in United Kingdom and their followers in Twitter. Twitter data analysis and application of social network analysis allow us to describe how these

institutions are related to each other, to their followers and what the connection among the followers is. We may be able to find, if the individuals or networks of individuals interacting with an organization(s) or subject collections, and which role the organizations play among themselves – the context which is missing while taking the insight by an organization only.

The social functions of Comment, Mention and Retweet are potentially the features enabling contributions. Therefor we can draw results similarly with the previous analysis in terms of total participation and meaningfulness of contributed information, and compare the outcomes.

Secondly, we ask the sample user group to keep diaries or the logs of their work with the digital resources of the case study organizations for one week. The task is to document every item that belongs to any of the repositories or external platforms of the organizations. The diary spreadsheet collects the following information: date; link to item that belongs to any repository of the three organizations (URL); work-related item (Yes/No); action that led to the item (searched how and where, returned to previously saved search/item, browsed how and where, recommended by who, how; found serendipitously where etc.); action with the item (looked up the metadata, opened online, skimmed/browsed, downloaded, saved, interacted: liked/favorited, commented, shared or recommended, metadata added etc.); comments.

4.3 Estimating the effect

In order to estimate the effect of social interaction to discoverability we use the results from quantitative data analysis as an input to interviews.

First, we hold one in-depth interview per each case study organization (max. 2 people) to inquire, what is the added value of the process, and increased visibility and discoverability for organizations? The interview reflects the quantitative findings and includes topics like: what are the aims for enabling user interactions offline and online; what are the quantitative and qualitative aims of the organization; what is the meaning of improved metadata, and engaged user communities for organizations; what are the experiences so far; what are the perceived profiles of the contributors.

Secondly, semi-structured interviews will be held with every user, who completed the diary. The aim is to inquire, what is the added value of the process and increased visibility and discoverability for academic community? The interview reflects on diaries, asks additional information if needed and covers topics like: what is the meaning of participation and contribution; offline activities of the users related to the organization and/or the user community; the urban factor in the process; previous experiences in user participation; experiences with other organizations.

The final question that concludes the research design is: How do library, archive and museum differ in the problem statement, process and results regarding enabling social interaction for improving discoverability of items? We will answer it through a comparative analysis between the three case study organizations throughout the three phases of the research.

5 Preliminary findings

The British Library Flickr account (http://www.flickr.com/people/britishlibrary) was established in August 2007 with the purpose to engage wider audiences with the Library's collections and enabling them to contribute tags to the images. The short-term goal was to increase the visibility and findability of the images. The long-term task is to improve discoverability of the images by finding the way, how to incorporate the contributed tags into the digital library. For the time being Flickr was chosen for the short term task, because that platform is designed for social interaction unlike the digital library system [30].

In the beginning of 2014, 1 021 040 cropped images from public domain books were published on the Flickr page. The library had metadata of the books, but nothing attached to the images. Thus these images were not searchable for users.

The aim of the Flickr data analysis is to detect main trends of user behavior while tagging images within one year, i.e. 2014. The main parameters are: 1) quantitative (how many people involved/tags contributed) - for mapping the phenomenon within the institutional context, 2) qualitative (how meaningful the tags are) - for estimating the potentiality of improving discoverability in practice.

Methods. The initial data file delivered by the British Library contained information about contributions of tags within the year 2014. The .tsv file was imported into R [31] and cleaned according to the following criteria: (1) The main three variables were selected for analysis: 'flickrid' (refers to the image that was available for tagging), 'author' (flickr user id), and 'tag' (entered text without spaces, punctuation and upper cases); (2) After description of the overall statistics and right before analyzing the most given tags, the column 'tag' was cleaned from numerical data in order to count tags like 'sysnum[n]', 'page[n]', 'date[n]', 'vol[n]' etc. as one unique tag for seeing better the diversity of tags. Additionally, for the same reason all occurrences of one of the most common tag 'pubplace[town or other place]' (presumably referring to the place of publication) were replaced with a single tag 'pubplaceX'.

Results. During 2014, 332 users attributed 12 178 295 tags (incl. 164 270 unique tags) to 1 021 040 images. Notably, one user contributed 12 034 916 tags, i.e. 98,8% (including 126 000 unique tags), which is possible only by attributing already existing relevant metadata of books to the images in a computerized way. The next two greater contributors gave accordingly 35 568 and 34 415 tags, which is 0,3% of total tags per each. The most dedicated group of contributors formed 6% of total users (n=20), attributing >1026 tags. The median 8 and mode 1 of attributed tags per user do not vary, while including or excluding the top 3 contributors. 56% of users attributed 1-10 tags.

After modification of tags (eliminating numerical data and counting most occurrences of 'pubplace[...]' as on tag), 46 956 unique modified tags remained, including 20 808 tags given by the most active contributor. Given that 55 496 unique modified tags were attributed by users in total, we can assume that 8540 tags were in use by different contributors.

The biggest corpus of tags is composed presumably using the information about the author of the publication (not the image), which may not be relevant for the in-

formation retrieval of images. The content analysis of 1200 most given unique modified tags reveals that 56 tags (4,6%) refer to the theme or format of the image and thus can potentially increase the discoverability of the item. Examples include (in the sequence of most attributed, underscores added for better readability): 'small' (attributed 417348 times), 'large' (385460), 'medium' (217128), 'map' (44537), 'portrait' (5309), 'music' (2337), 'architecture' (1965), 'decoration' (1838), 'initial' (1806), 'coat_of_arms' (1326), 'heraldry' (1259), 'people' (1133), 'split' (831), 'cover' (798), 'world_war' (674).

1 021 040 images received in average 12 tags in total. The 5 most tagged images got 36 to 53 tags. In average there were 3073 items available for tagging per one contributor who took part.

The further analysis inquire, what is the interval of tagging per a contributor (does the user try tagging out once or keeps returning); is there a correlation between authors and tags (do the different users tend to give similar tags; how much do tags differ per a contributor); do the tags match with the metadata given to books; is there a correlation between contributors and tagged items (do the same/different users tend to tag the same objects); how do these results differ from the next year, 2015; is there a correlation between time series of tagging and of visits to Flickr page?

Limitations. The data about users is based on unique Flickr ids. Therefore it cannot be confirmed that some users have created several user accounts and thus appear in the database more than once. Also due to the volume of the dataset we rely on the automatized results while counting the unique tags. Thus some obviously similar tags are counted many times, if they contain any distinctive characters. Yet the error margin for the dataset with nearly 13 million items is expected to be low.

6 Concluding remarks

Social interaction related to digital libraries is only an emerging trend and there is a need to carry out additional comparative study in another European country in order to claim that the results are consistent in Europe.

The author has not recognized similar research carried out previously. On one hand it justifies the variety of methods included in this research project. On the other hand, it gives grounds to believe, that the conclusions will suggest methodical adjustments for the further studies.

Acknowledgements. This research is pursued at the doctoral program "The City and the Society of Information", supervised by Dr. Marco Gui, Department of Sociology and Social Research, University of Milano-Bicocca, and Prof. David Lamas, Area of Human-Computer Interaction, Tallinn University. The author acknowledges the data received from the British Library. However the publication reflects only the views of the author, and the Library cannot be held responsible for any use of the data.

References

1. Van House, Nancy A., Bishop, Ann Peterson and Buttenfield, Barbara P. (2003) Introduction: Digital Libraries as Sociotechnical Systems, pp. 1–21. In: Buttenfield, B. P., Van House, N. A., & Bishop, A. P. Digital Library Use: Social Practice in Design and Evaluation. Cambridge, Mass: The MIT Press.
2. Calhoun, K. (2014). Exploring Digital Libraries. Foundations, practice, prospects. London: Facet Publishing.
3. Tredinnick, Luke (2008) Digital information culture: the individual and society in the digital age. Oxford, Chandos.
4. Ridge, Mia. Crowdsourcing our Cultural Heritage. (2014). Farnham, Surrey, England: Ashgate.
5. Lee, Francis L. F; Leung, Louis; Qiu, Jack L.; Chu, Donna S. C. (2013) Introduction: Challenges for New Media Research. Frontiers in New Media Research. New York: Routledge, pp. 6-14.
6. Verheul, Ingeborg; Tammaro, Anna Maria; Witt, Steve (2010) Foreword. Digital Library Futures: User perspectives and institutional strategies. IFLA Publications Series, vol. 146. Berlin/Munich, De Gruyter Saur.
7. Europeana. `http://europeana.eu` (accessed Dec 2015)
8. Boast, Robin; Bravo, Michael; Srinivasan, Ramesh. (2007) Return to Babel: Emergent Diversity, Digital Resources, and Local Knowledge. The Information Society: An International Journal, vol. 23, iss. 5, pp. 395-403.
9. Discoverability. (1989). In J. A. Simpson, & E.S.C. Weiner (Eds.), The Oxford English dictionary (2nd ed.). Oxford: Clarendon Press; Oxford, New York: Oxford University Press.
10. Somerville, M. M., & Conrad, L. Y. (2014). Collaborative Improvements in the Discoverability of Scholarly Content: Accomplishments, Aspirations, and Opportunities. A SAGE White Paper. Los Angeles, CA: SAGE. doi: 10.4135/wp140116.
11. Discovery service. (n.d.). In Reitz, J. M. ODLIS. Online Dictionary for Library and Information Science. Retrieved from `http://www.abc-clio.com/ODLIS/odlis_d.aspx`
12. Higgins, Sarah. Digital Curation: The Emergence of a New Discipline. 2011, Vol. 6, No. 2, pp. 78-88.
13. Westbrook, R. N., Johnson, D., Carter, K., & Lockwood, A. (2012). Metadata Clean Sweep: A Digital Library Audit Project. D-Lib Magazine, 18(5-6).
14. Matusiak, K. K. (2006). Towards user-centered indexing in digital image collections. OCLC Systems & Services: International digital library perspectives, 22(4), 283-298. doi:http://dx.doi.org.proxy.unimib.it/10.1108/10650750610706998
15. Macgregor, G., & McCulloch, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. Library Review, 55(5), 291-300.
16. Szajewski, M. (2013). Using Wikipedia to Enhance the Visibility of Digitized Archival Assets. D-Lib Magazine, 19(3/4). doi:10.1045/march2013-szajewski.
17. Trant, J., & Wyman, B. (2006, May). Investigating social tagging and folksonomy in art museums with steve. museum. In Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland.
18. Galloway, E., & DellaCorte, C. (2014). Increasing the Discoverability of Digital Collections Using Wikipedia: The Pitt Experience. Pennsylvania Libraries: Research & Practice, 2(1), 84-96. doi:http://dx.doi.org/10.5195/palrap.2014.60

19. Somerville, M. M., & Conrad, L. Y. (2013). Discoverability Challenges and Collaboration Opportunities within the Scholarly Communications Ecosystem: A SAGE White Paper Update. Collaborative Librarianship, 5(1). Retrieved from http://collaborativelibrarianship.org/index.php/jocl/article/view/240/181.
20. Sturiale, Valentina. (13.04.2013). Museum Week 2016: La Vittoria Dei Piccoli Musei Su Twitter. Retrieved from http://www.viralbeat.com/blog/museum-week-2016-vittoria-piccoli-musei-twitter.
21. Espinós, A. (2014). Do Museums Worldwide form a true Community on Twitter? Some insights on the museum Twitter ecosystem through Social Network Analysis and Network Science. Retrieved from http://www.lamagnetica.com/twitter-museum-study.
22. Hadnagy, Christopher. (2011). Social engineering: the art of human hacking. Indianapolis (Ind.): Wiley.
23. Hill, L. L., Carver, L., Larsgaard, M., Dolin, R., Smith, T. R., Frew, J. and Rae, M.-A. (2000). Alexandria digital library: user evaluation studies and system design. Journal of the American Society for Information Science. Vol. 51, iss. 3, pp. 246–259. doi: 10.1002/(SICI)1097-4571(2000)51:3<246::AID-ASI4>3.0.CO;2-6
24. Kani-Zabihi, Elahe; Ghinea, Gheorghita; Chen, Sherry Y. (2006). Digital libraries: what do users want? Online Information Review, Vol. 30 Iss: 4, pp. 395–412. doi: http://dx.doi.org/10.1108/14684520610686292
25. Krystyna K. Matusiak, (2006). Towards user-centered indexing in digital image collections, OCLC Systems & Services: International digital library perspectives, Vol. 22 Iss: 4, pp.283 – 298. doi:http://dx.doi.org/10.1108/10650750610706998.
26. McMartin, F. (2006). MERLOT: A Model for User Involvement in Digital Library Design and Implementation. Journal Of Digital Information, 5(3). https://journals.tdl.org/jodi/index.php/jodi/article/view/143 (accessed 7.12.2015).
27. Mets, Õnne; Gstrein, Silvia; Gründhammer, Veronika (2014). Increasing the visibility of library records via consortial search engine. IEEE/ACM Joint Conference on Digital Libraries. 8-12.09.2014, London, IEEE, pp. 169-172.
28. Recker, Mimi M.; Dorward, James, & Nelson, Laurie Miller. (2004). Discovery and Use of Online Learning Resources: Case Study Findings. Journal of Educational Technology & Society, 7(2), 93–104.
29. European Commission. (24.06.2015). The Digital Economy and Society Index (DESI). https://ec.europa.eu/digital-agenda/en/desi (accessed Dec 2015)
30. Mahey, Mahendra. (9.06.2016). Consultation via Skype.
31. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org.