SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di / Department of

Sociology and Social Research

Dottorato di Ricerca in / PhD program URBEUR_QUASI Ciclo / Cycle XXX

Curriculum in (se presente / if it is) The City and Society of Information

USERS' PARTICIPATION AND DISCOVERABILITY OF DIGITAL COLLECTIONS

Cognome / Surname Mets Nome / Name Õnne

Matricola / Registration number 798859

Tutore / Tutor: Dr. Marco Gui

Cotutore / Co-tutor: Prof. David Lamas (Tallinn University)
(se presente / if there is one)

Coordinatore / Coordinator: Prof. Lavinia Bifulco

ANNO ACCADEMICO / ACADEMIC YEAR 2016/2017

Table of Contents

List of Figures

List of Tables

Abbreviations

API – Application Programming Interface, a set of tools that developers can use to access structured data

BL – British Library

GLAM – Galleries, Libraries, Archives, Museums

ICT – Information and Communication Technology

JSON – JavaScript Object Notation

MARC – Machine-Readable Cataloguing Record

OCR – Optical Character Recognition, which facilitates full-text search

OPAC – On-line Public Access Catalogue

SNS – Social Network Site

TNA – The National Archives of the United Kingdom

UGC – User Generated Content

UX – User Experience

Abstract

The dissertation “Users' Participation and Discoverability of Digital Collections” aims to investigate and clarify the relationships between users' participation and discoverability of digital collections in the context of a library and an archive. The activity theoretical approach is followed in the project and considers technology as a mediator in the context of social interaction. The research project addresses the knowledge gap in social research in the field of digital libraries. The research questions explore what is the framework for user interaction and how users interact under set conditions with the archival and library collections in institutional catalogues and in Flickr. The British Library and the National Archives are studied on the basis of the document analysis of the 5 platforms in their use, statistical analysis of the attribution of over 744000 tags on those platforms, thematic analysis of 4 in-depth interviews with staff members, and 24 interviews with visitors of the respective organisations. The results show that social tagging contributes to discoverability in four modes: in invisible mode (tags are not searchable, only taggers can see their own tags); in individual mode (attributed tags are meaningful only for the taggers, including tags for institutional workflows); in restricted mode (tags are searchable for all, but authentication for adding tags is restricted); in public mode (tags are searchable for all and everyone can add tags upon signing up). The contributions of the project include suggestions for organisations to improve user interaction, a sizable collection of systematised quotes by the staff members relevant for practitioners and research, a methodological note on the importance of completing statistical user studies with detailed document analysis, and the proposal of an additional actor *Rescources* for the activity theoretical framework of activity system.

Keywords: social tagging, catalogues, Flickr, the British Library, the National Archives of the United Kingdom, activity theory.

Abstract (Italiano)

La ricerca “Users' Participation and Discoverability of Digital Collections” vuole indagare e chiarire le relazioni tra l’interazione sociale e la reperibilità delle raccolte digitali nel contesto di una biblioteca e di un archivio. Il progetto adotta l’approccio teorico della *activity theory*, assumendo la tecnologia come mediatore nel contesto dell’interazione sociale. Lo studio si pone l’obiettivo di colmare le lacune conoscitive della ricerca sociale sulle biblioteche digitali. Le domande di ricerca esplorano il quadro in cui si svolgono le interazioni degli utenti e come queste vengano siano soggette a condizionamenti definiti dalle raccolte di archivi e biblioteche in cataloghi istituzionali e in Flickr. La British Library e i National Archives sono studiati attraverso un’analisi documentale, a partire dalle 5 piattaforme utilizzate dalle due organizzazioni, un’analisi statistica dell’attribuzione di oltre 744000 tag presenti su tali piattaforme, e un’analisi tematica di quattro interviste in profondità con membri dello staff e di 24 interviste con visitatori delle rispettive organizzazioni. I risultati mostrano che l’uso sociale dei tag contribuisce alla reperibilità in quattro modi: in modo invisibile (se l’attribuzione del tag non è consultabile e solo l’autore può vedere le proprie attribuzioni); in modo individuale (se le attribuzioni sono significative solo per colui che le ha definite, compresi i tag funzionali all’organizzazione del lavoro); in modo limitato (se i tag sono consultabili da tutti ma vi sono restrizioni sulla registrazione per l’aggiunta di tag); in modo pubblico (se i tag sono consultabili da tutti e ciascuno può aggiungere tag dopo essersi registrato). Il progetto include contributi di vari tipi: considerazioni per le organizzazioni per migliorare l’interazione degli utenti, una considerevole raccolta di citazioni rilevanti per professionisti e per la ricerca, sistematizzate dai membri degli staff, una nota metodologica sull’importanza di integrare gli studi statistici sugli utenti con l’analisi documentale dettagliata, ed infine la proposta di un attore aggiuntivo del sistema dell’attività all’interno del quadro teorico dell’activity theory, definito *Risorse*.

Parole chiave: tag, cataloghi, Flickr, British Library, National Archives of the United Kingdom, activity theory.

1. Introduction

The current research project, “Users' Participation and Discoverability of Digital Collections”, studies human behaviour in a technological setting, specifically the platforms used for enabling interaction with digital collections. The digital collections included in this project are digital and/or digitised materials of memory institutions, i.e. libraries, archives and museums. Also a collection of metadata (records) can be a digital collection in the form of online catalogue.

Discoverability refers to the item's (e.g. an image, text, a record) ability to be found through appropriate infrastructure by appropriate users (Somerville and Conrad 2014, 3). It is different from the concept of visibility, which concerns placing information in locations where people may come across it (Somerville and Conrad 2013).

Social interaction occurs when at least two interacting agents share a common set of signs and a common set of rules (Hadnagy 2011, 47). In our context, the interaction is mediated: a person can take an action upon a collection item in a system, which is visible to and can be used and complemented by others.

The three concepts – digital collections, discoverability and users' participation – captured in the title of this study point towards two parallel processes explaining the research problem. First, people have become increasingly switched on, interacting as well as collaborating online (Tredinnick 2008). Secondly, the collections of memory institutions are becoming increasingly digital. This research project started from the consideration that if we knew more about the crossing points of these two processes, we could make useful suggestions to memory institutions about involving people, preferably across different types of memory institutions.

Under this assumption, Chapter 2 (State of the Art) first looks into the need for social research in the digital library context, then discusses the *social* in the context of *digital* and defines the term *discoverability* in the cultural heritage sector. Thirdly, participatory online audiences and crowdsourcing in memory institutions are examined. The reviews of case studies report findings based on a single institution or on collaborative projects between institutions of the same type. There is not enough

evidence about on-going collaboration between different types of memory institutions, which is integrated in their everyday processes and available to the end users. At the same time, online users are getting more used to integrated solutions, which augment different type of information in their interest from any source. This makes the case to compare the practices of memory institutions belonging to different types.

Collection items can be integrated into different systems, but they need to be described in a structured way in order to be found by users. This directs the focus of the project on to the concept of *discoverability*. Given that organisational capability is limited in describing millions of items in their collections, and that collaborative culture has become more popular among people, the main research question is formulated as follows: how are users' participation and the discoverability of digital collections related?

In order to find answers to the main research question, the following sub-questions were posed first. Where does user interaction take place? What do the institutions enable the users to do in those media and why? How do users interact under those conditions and why? Who are the users? And what kinds of relationships exist between the type of organisation and the platform?

The conceptual framework (Chapter 3) needed for such a setting should acknowledge the mediating tools, unlike a digital anthropological approach which places importance on the whole cultural process and sees digital as part of it. Activity theory enables the description the social context while considering the technological factors (Kaptelinin and Nardi 2006/2009). Therefore, activity theory is introduced in order to examine the relations between organisations, people, and the other actors concerning objectives, tools, community, rules, division of labour and outcomes. Additionally, participatory models (Simon 2010) are presented as part of the conceptual framework for enabling generalisation and categorisation of the findings about organisational practice.

Previously published studies do not provide the level of detail that would facilitate a comparative secondary study between different types of memory institutions. Therefore, the research design in Chapter 4 presents first, the approach for data collection, and then methods for analysis. The research design is aimed to be explorative and qualitative in order to describe both contexts carefully enough for comparison, which directs the research project to the case study approach. In order to

avoid cultural-political actors affecting the judgement of differences derived from the type of the institution, the case studies were selected from the same country. The British Library and the National Archives of the United Kingdom were selected as representing established organisations with initiatives to engage people online with collection items in various ways, including social tagging which is believed to improve discoverability. The preliminary document analysis reveals where this type of user interaction takes place. Both organisations enable social tagging in catalogues and on their Flickr pages. The following study focuses on 5 selected platforms: the main catalogue *Explore* and specialised catalogue *Archives and Manuscripts* of the British Library, the main catalogue *Discovery* of the National Archives, and the Flickr pages of both organisations. This facilitates a 3-dimensional comparison between the type of the organisation (library-archive), the type of the collection available for tagging (archival-library), and the type of the platform (catalogue-social network site).

The following two chapters focus on the results of three different methods applied to analysing the case studies. In Chapter 5, the document analysis of the interfaces and related help pages first reveals what the institutions have enabled the users to do in those mediums. Secondly, in the same chapter, social tagging data analysis clarifies how users interact under those conditions. Thirdly, in Chapter 6, thematic analysis of in-depth interviews with staff members contributes to answering why the institutions have decided on the approach described by document analysis. Thematic analysis of the interviews with visitors to both organisations clarifies who the users are and what explains their behaviour and preferences as taggers or non-taggers. Throughout the two chapters, evidence is provided for answering the last sub-question: what kind of relationship exists between the type of the organisation, collection or platform?

Chapter 7 summarises the findings and brings together the different contributions made by the current research project and systematically discusses them. By answering the research questions, it expects to provide a snapshot of the transformational stage of the analysed organisations, to contribute to academic literature in the field, and to provide suggestions for action in a consultancy perspective.

2. State of the Art

This chapter includes four sections. The need for social research in the context of digital collections is addressed in section 2.1. The emergence of the social in the context of the digital is discussed in section 2.2, and the term 'discoverability' is defined. The design of participation is characterised in section 2.3, and crowdsourcing in memory institutions in characterised with some detailed examples. The chapter results in presenting the research topic and the research questions in section 2.4.

2.1. Social Research

The internet is a social phenomenon with deep societal implications, e.g. on individual well-being, relations with others, and social capital building within communities, which highlight the importance of finding out about many aspects of individuals' behaviour in regard to the internet (Dutton 2013; Weinberger 2010; Haythornthwaite 2001). Established institutions also continue to try to find the best way to appropriate the technological affordances of the internet (Lee et al. 2013).

Van House et al. (2003) have noticed that digital libraries can be seen as not simply a new technology or organizational form, but as a change in the social and material bases of knowledge work and the relations among people who use and produce information artefacts and knowledge. In the context of constructing scholarly information infrastructure, Borgman (2010) enquires what to build when we know so little about how it will be used. This makes it important to include social sciences in digital library research.

Research on social aspects of digital libraries is often within the context of user-centred design, work practice studies, the social web, and other topics related to specific projects or programs. There is a knowledge gap about how users or other players are used in fulfilling the roles of digital libraries or how the society can contribute to the social roles of digital libraries, like fostering and enhancing collaboration and partnerships among and across individuals, institutions, groups, and domains of education, research, or commerce (Calhoun 2014).

The emergence of an increasingly participatory culture is evident, in which individuals contribute to the creation of information, knowledge, and cultural artefacts

through different modes of collaboration in the digital sphere (Tredinnick 2008, 105; Ridge 2017). Libraries, archives, and museums are also collaborating to make their cultural heritage collections available online (Verheul et al. 2010), but the practice of institutions often remains experimental or inconsistent. Europeana[1] for instance is a European Union-supported third-party initiative to aggregate cultural heritage data.

Public relations theories have placed an organization at the centre, pursuing a narrow understanding of relationships and neglecting a broader understanding of how discourse shapes meaning and relationships (Yang and Taylor 2015). The need for new methodologies by which public institutions can access and maintain digital knowledge resources in manners that directly consult stakeholder communities is clear (Boast et al. 2007). Following the social constructivist principle, knowledge creation is a shared, rather than individual experience (Aleksić-Maslać et al. 2009).

2.2. Emergence of the Social in the Digital

The digital anthropological approach claims that the digital is not brought to culture, facilitating or changing it, but is a cultural object as well as a process. The development of media and communication technologies is not believed to be guided so much by the local appropriation of a technology, but more by the importance of listening to the differences in culture which determine what a particular technology becomes: "Studies of digital museums tend to be over-determined by the form of the digital and less descriptive of the intricate ways in which the digital can be embedded in pre-existing frames of being: of classification, epistemology and sociality" (Geismar 2012, 277–281).

'New Museology' has documented a shift of interest in museums away from objects and towards people, society, and experience. The museum database is not just a structure for storing information, but a symbolic form in which the interface and the object are the same thing. The social tags in a digital museum repository may be seen as representations of users as well as collections. The digital domain functions simultaneously as a representation of other sites and practices, and as a site and practice in itself. (Geismar 2012, 267–270)

Using the example of an archive (Geismar 2012, 272), it is illustrated that the collections are not always intended to be opened up promiscuously. Some collections need to be seen in an original context in order to avoid issues like recognized status

[1] europeana.eu
[2] flickr.com
[3] familysearch.org/volunteer/indexing

versus cultural identities. That leads to the inference that sometimes the preservation of not only the object but also the context might prevail over the ambiguity of openness.

One of the many social roles of digital libraries is the "free flow of ideas," where people can engage with crowdsourcing, annotations, tagging, ratings, recommendations, reviews, citation managers, alerts, social bookmarking, blogs, and wikis. The community benefits therein from a locus of shared work, providing virtual space for "rational and enlightened discourse," and facilitating interaction between content, creators, and the public (Calhoun 2014, 146). This approach brings the mediating tool into the picture.

Likewise, Somerville and Conrad (2013) believe that influence is migrating from organizations to networks and new "experts." People increasingly expect participatory information exchanges, and social networks are more influential in the process.

Social interaction may emerge in metadata enrichment, collaborative indexing or tagging, donation of materials, recommendations, rankings, asking, sharing, commenting, or in other ways. These actions contribute to a variety of aspects concerning digital collections, like design and architecture, visibility, use and reuse of the content or data, usability, usefulness, discoverability, feedback, evaluation, etc. (Ridge 2017; Hill et al. 2000, Kani-Zabihi et al. 2006; McMartin 2006, Mets et al. 2014; Recker et al. 2004).

Active participation may lead to information overload and even to noise, if the user-attributed information is perceived as useless. Weinberger (2010, 9) argues, that when the information overload was mainly seen as a psychological issue some decades ago, then these days we think of it more as a cultural condition and we do not worry that information overload "will cause us to have a mental breakdown but that we are not getting enough of the information we need". Internet scholar Clay Shirky follows the idea of cultural condition and claims that there is no information overload, but only filter failure. When people add tags in order to apply filters to the information that they retrieve then they add meaning, even scale meaning (Weinberger 2010).

Weinberger (2010, 9) finds that this is the reason why we adopt different methods to understand the world around us and make smarter decisions by being as smart as our new network medium allows. The technologies that we have rapidly evolved to help us fall into algorithmic and social categories, but mostly combine both.

“Algorithmic techniques use the vast memories and processing power of computers to manipulate swirling nebulae of data to find answers. The social tools help us find what’s interesting by using our friends’ choices as guides”. Social information may be controversial, but controversies create debates and dialogue about the content. This might complement the Weinberger's idea of the new knowledge being not what we all agree, but what we share.

Fullerton and Rarey (2012) claim that collaborative practices on social online platforms illustrate the interaction among users revealing a wish to build social capital and/or cultural capital. Matusiak (2006) adds that implementing social networking applications in digital collections can foster collaborative knowledge construction. Users can contribute to the depth of image description and enhance the intellectual content of digital collections. Expertise in local history and language is expected to be particularly valuable in cultural heritage collections, where users can help to identify images and enhance descriptions with their unique knowledge and perspectives. Users' comments can also be a source of evaluation data, indicating the relevance of collections to users' needs and providing direction for future development of digital image collections (Matusiak 2006).

A popular service for tagging images is Flickr[2] (Daly and Ballantyne 2009), which supports many social behaviours that are not available in museums and galleries. It can be more appealing to see photographs in a gallery, but on the other hand, providing social functions around objects promotes other kinds of user experiences. E.g. the Library of Congress offered an image “Workers leaving Pennsylvania shipyards, Beaumont, Texas” taken by John Vachon in 1943 to Flickr Commons, an area of Flickr reserved for images in the public domain. Visitors could have hunted it in an online database before, but never seen it exhibited. Via Flickr it was shared by thematic blogs or personal e-mails from the site, annotated with socio-political commentary, and links were provided to similar materials or to personal stories related to the event or the place etc. It depends on institutional goals and priorities whether this social value is worth the aesthetic trade-offs. But if the goal is to encourage visitors to engage with each other about the stories and information at hand, then Flickr is the ideal choice (Simon 2010, 137).

[2] flickr.com

2.2.1. Discoverability

According to the English language dictionary discoverability is "the quality of being discoverable; capability of being found out" ("Discoverability" 1989, 753).

Similarly "A SAGE White Paper on Collaborative Improvements in the Discoverability of Scholarly Content" (Somerville and Conrad, 2014, p. 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."

Firstly, the definition states the goal is to *find*, although not necessarily access, the materials. Discoverability can be related to physical as well as online collections, available to all or with restricted access. Discoverability may also happen elsewhere, e.g. through Google Scholar. A great deal of information-seeking for academic content has moved to public search engines, academic search engines, or discipline-specific databases and aggregations (Calhoun, 2014, p. 132-133).

Secondly, discoverability is related to the content, not to a collection or a system as a whole. Specifically, the definition places discoverability in relation to an item. This in turn has detailed implications in terms of discovery tools: e.g. MARC records provide online discoverability for archives and manuscript collections at the collection level, not at an item level (Calhoun, 2014, p. 5-6).

Thirdly, the definition's inclusion of "being found by appropriate users" reinforces the relation with users, presumably with human beings and not with machines or robots.

Discovery is defined as "the process and infrastructure required for a user to find an appropriate item" (Somerville and Conrad 2014, p. 3). The process of discovery may be supported by a discovery service, which is "a single interface, providing integrated access to the multiple information resources (catalogues, publishers' e-book and e-journal collections, subscription databases, archival collections) to which a library has rights." A discovery service uses consolidated subject indexing and metadata, and search results are generally deduped and relevance ranked: e.g. EBSCO Discovery Service ("Discovery service" n.d.).

Occasionally, the literature reveals the use of the term 'discoverability' to mean 'visibility,' which "involves placing information in locations where people will come across it in the work that they do" (Somerville and Conrad 2013).

Metadata is considered as a key feature to discoverability (Higgins 2011; Westbrook et al. 2012). The better the quality of the metadata, the better the discoverability.

It is especially true in case of images, which is difficult to access without prior indexing. Human indexers can interpret the meaning of the picture, assign subject headings, and transcribe image captions and textual annotations in order to expose the items for different users from different disciplines and with different interests and personal and professional backgrounds. In practice, digital librarians struggle with an increasing number of digital images that need to be indexed for online delivery. Traditional indexing techniques are costly and labour-intensive, and even practitioners are not sure whether they represent the only or the best way to meet user needs (Matusiak 2006) or are adequate for online resource discovery (Macgregor and McCulloch 2006).

Social classification – also referred to as distributed classification, social tagging, ethnoclassification, and folksonomy – represents organising content in the web environment where users create their own textual descriptors using natural language terms (tags) and share them with a community of users. It offers an opportunity for greater user engagement and help in building virtual communities. An interlinked system of tags supports browsing activities and the serendipitous discovery of images in the digital environment (Matusiak 2006).

The most important strength of social tagging is its close connection with users and their language. User-generated metadata reflects an increasingly multilingual and multicultural web audience (Matusiak 2006). Daly and Ballantyne (2009) agree that tagging reflects the language of users which might also be used for searching in the future, whereas controlled vocabularies may become distanced from the real usage.

If controlled vocabularies and standards enable uniform access and interoperability, then social classification brings in user language, perspective, and expertise and may eventually lead towards more user-oriented indexing. Yet social classification does not have to be seen as an alternative or replacement for traditional indexing, but rather as an enhancement (Matusiak 2006).

2.3. Design of Participation

There are several types of participatory social media online audiences and many people fall into several categories. The divisions may vary by country, age group or

other variables, but one thing that stays constant is that the creators are a small part of the landscape (Simon 2010, 8):

1. Creators 24% – produce content, upload videos, write blogs
2. Critics 37% – submit reviews, rate content, and comment on social media sites
3. Collectors 21% – organize links and aggregate content for personal or social consumption
4. Joiners 51% – maintain accounts on social networking sites like Facebook or LinkedIn
5. Spectators 73% – read blogs, watch YouTube videos, visit social sites
6. Inactives 18% – do not visit social sites.

Sometimes, just reading or watching makes someone an important participant: e.g. participation as a viewer affects the status of each video in the YouTube system, Google gets instantly smarter every time someone clicks an ad, and Netflix gets smarter when someone recommends a video (Simon 2010, 10, 85).

According to Nielsen (2006 in Simon 2010), this participation inequality relates to the "90-9-1" principle: in most online communities, 90% of users are lurkers, who never contribute; 9% of users contribute a little; and 1% account for almost all the action. Furthermore, only 0.16% of visitors to YouTube will ever upload a video and only 0.2% of visitors to Flickr will ever post a photo. But the aim is not to increase these percentages. The platform's value is more dependent on the number of active critics, collectors, and joiners than the number of creators. The overall YouTube experience would likely be worse for spectators if the service was glutted with millions more low-quality videos. But the more interpretation, prioritization, and discussion there is around the content, the more people can access the videos (and the conversations) that are most valuable to them. (Simon 2010, 9–10).

The best participatory projects create new value for the institution, participants, and non-participating audience members. Projects suffer when visitors perceive that the staff are pandering to them or wasting their time with trivialities. The work participants do should ultimately be of value to the institution, and the staff members need to offer participants something fundamental: personal fulfilment (Simon 2010, 17–18).

According to Simon (2010, 22–25), there are two counter-intuitive design principles at the heart of successful participatory projects, which make the visitors to feel confident participating in creative work with strangers:

1. Participants thrive on constraints, not open-ended opportunities for self-expression. They need scaffolding for experiences that put their contributions to meaningful use.
2. To collaborate confidently with strangers, participants need to engage through personal, not social entry points.

But, simultaneously, participatory projects do not have to be fully designed before launch. Adaptive evaluation techniques are particularly natural, because collecting data about user behaviour is fairly easy with analytical tools, and most web designers, particularly those working on social websites, expect their work to evolve over time. Most Web 2.0 sites are in "perpetual beta," meaning that they are released before completion and remain a work in progress to respond to observed user behaviours.

Participatory techniques can help visitors to develop specific skills, referred to as "21st century skills," "innovative skills," or "new media literacies," related to creativity, collaboration, and innovation. These skills enable people to collaborate and interact with others from diverse backgrounds; generate creative ideas both alone and with others; access, evaluate, and interpret different information sources; analyse, adapt, and create media products; be self-directed learners; adapt to varied roles, job responsibilities, schedules, and contexts; and act responsibly with the interests of the larger community in mind. Some institutions have adopted participatory learning skills as part of their commitment to overall visitor learning, which may result in tension with content learning (Simon 2010, 193-195).

Power. To be successful leaders in a socially networked world, cultural institutions must feel comfortable managing platforms as well as providing content. One of the primary fears is the fear of losing control. But intervention by users does not mean giving all the power to visitors, because institutional control remains in platform management. Platform managers have the power to define the types of interaction available to users, set the rules of behaviour, preserve and exploit user-generated content, promote and feature preferred content (Simon 2010, 121). The power structure depends on the nature of the project as well as the institutional culture. Working toward participation leads to examination of the institution's mission statement, which helps staff members and stakeholders to understand the value of participatory projects and paves the way for experiments and innovation (Simon 2010, 188–193).

Memory institutions tend to be protective of their collections and restrict making objects social by being shared. But the emergence of the social web is changing the way professionals think about sharing in cultural institutions. By sharing, people explain the world around them and help others to learn. Some share collection data and images openly on third-party social websites like Flickr or Wikipedia; others build their own online platforms with custom functions and design that allow visitors to remix objects and spread them with social web sharing tools. Simon considers it particularly radical when museums share their digital collection content and software coding openly with external programmers, who can then develop their own platforms and experiences around the digital media; e.g. the Brooklyn Museum and the Victoria & Albert Museum have made their collection databases openly available to outside programmers, who have used them to create their own online and mobile phone applications (Simon 2010, 172–174).

Tagging. New social networking paradigms of user-contributed metadata are emerging, and users can contribute via textual or visual annotations (Kowalczyk and Shankar 2011). Tagging is claimed to work best with smaller communities of like-minded people who share knowledge bases, interests and skill sets (Geismar 2012, 270).

Some users may have spare time to read materials in their topic of interest and would not mind correcting erased letters, for instance, or index materials they know, like family history records. For example, FamilySearch Indexing[3] is an initiative that crowdsources the indexing of family history records and makes the results freely available. One of the earliest proof-reading projects, dating from 2000, is Distributed Proofreaders[4], which was inspired by Project Gutenberg (Calhoun 2014, 253).

In the context of art museums, tagging is a way for people to connect directly with works of art, to own them by labelling or naming them, and to assert personal perspectives and associations between objects – for self-presentation. It may help to foster and maintain links with specialized groups like volunteers and docents, or to support the work of teachers and students. Thus social altruism prevails over the motivation of personal gain (Trant and Wyman 2006, Geismar 2012, 271).

Users are sometimes also authors (Kowalczyk and Shankar 2011). In that case, they contribute before the work is made visible through the digital library by

[3] familysearch.org/volunteer/indexing

[4] pgdp.net/c

providing metadata through the publisher. This kind of metadata is shown as 'institutional metadata.'

As a negative result of social classification, misspellings, badly encoded word groupings, singular and plural forms, personal tags, and single use tags may occur. But again, making the presence of a librarian evident by improving "sloppy tags" may discourage users (Matusiak 2006).

2.3.1. Crowdsourcing in Memory Institutions

There is a long tradition of volunteer augmentations of GLAM collections through public participation in collection, research, and observation which pre-dates technology-enabled crowdsourcing as we know it (Ridge 2017).

Crowdsourcing, originally described as the act of taking work once performed within an organisation and outsourcing it to the general public through an open call for participants (Howe 2006 in Ridge 2017), is becoming increasingly common in museum, libraries, archives, and the humanities as a tool for digitising or computing vast amounts of data. Ridge sees crowdsourcing in cultural heritage as more than a framework for creating content: crowdsourcing is a form of engagement with the collections and research of memory institutions, benefitting both audiences and institutions.

Cultural institutions are progressively exploring crowdsourcing, and projects-related research is increasing, but little has been written more generally about crowdsourcing practices promoted by galleries, libraries, archives, museums, and education institutions (Carletti et al. 2013). The study by Carletti et al. suggests the value of investigating sociotechnological systems that can support the collaboration between cultural institutions and their public, as well as efficiently combining multisource content. Rethinking the relationship between official and unofficial knowledge is expected to be the main challenge that cultural institutions have to face when undertaking a crowdsourcing process.

There have been concerns as to whether digital libraries will succeed in the user engagement (Kreijns et al. 2003), as Wikipedia has done, unless they implement social networking applications on a larger scale and create encouraging environments for people to contribute their expertise. User language can be incorporated into digital collections by allowing users to add their own tags to the metadata in the records; to

provide feedback on the terms assigned by indexers; or developing a controlled vocabulary through the use of user-supplied tags (Matusiak 2006).

In practice, the crowd-driven annotation technique has been considered successful, not only in a variety of ways for accessing content but also for engaging users with online collections (Van Hooland 2011). The cultural heritage sector has embraced this practice and is progressively incorporating it, together with other crowdsourcing initiatives, as part of their workflows (Oomen and Aroyo 2011, 138–149). Ridge (2017) notes that while the potential savings in staff resources and enhancements to collections are the most obvious benefits of cultural heritage crowdsourcing, deepening relationships with new and pre-existing communities have been important to many organisations.

Cultural heritage crowdsourcing projects ask the public to undertake tasks that cannot be done automatically, in an environment where the activities or goals (or both) provide inherent rewards for participation, and where participation contributes to a shared, significant goal or research area (Ridge 2017). Ridge has noticed growing evidence that typical GLAM crowdsourcing activities encourage skills development and deeper engagement with cultural heritage and related disciplines.

Crowdsourcing in cultural heritage benefits from its ability to draw upon the notion of the ‘greater good’ in invitations to participate, and this may explain why projects generally follow collaborative and cooperative, rather than competitive, models. Tensions are in the role of expertise and disruption of professional status, or lines of resistance to the dissolving of professional boundaries, and in validation of contributions (Ridge 2017).

Ridge (2017) agrees with Simon (2010) and many others on the small numbers of active contributors and points to the variety of terms used instead of 'crowd', like ‘community-sourcing,’ ‘targeted crowdsourcing,’ or ‘microvolunteering’, which acknowledge that often the 'crowd' is neither large nor truly anonymous. These terms additionally reflect the fact that while some cultural heritage crowdsourcing projects are inspired by a desire for greater public engagement, the more specialised the skills, knowledge, or equipment required, the more individuals may fall out from the pool of potential participants as being unable to acquire the necessary attributes.

Ridge generalises that the tasks performed by participants in cultural heritage crowdsourcing involve transforming content from one format to another (for example, transcribing text or musical notation), describing artefacts (through tags,

classifications, structured annotations or free text), synthesising new knowledge, or producing creative artefacts (such as photography or design). Due to the variability of materials in cultural heritage collections, tasks could be quick and uncomplicated or could require subjective judgement to accomplish (e.g. adding structured mark-up or choose between hierarchical subject terms).

The conclusions of a transcription project Transcribe Bentham (Causer and Terras in Ridge 2017, Chapter 3) illustrate the overall experience: the majority of work is done by minority of users; volunteers have an interest in the subject, crowdsourcing or the technology and sense of altruism; lack of time and issues with technology might limit participation whereas media attention increases it; the project results in increasing the digital literacy skills of participants, contributing to scholarship and widening access to the material, adjusting workflows, and exploiting investments for digitization, software development, and staff salaries.

Transcribe Bentham was launched to the public in September 2010. In less than four months, 350 users had partially or fully transcribed 439 manuscripts; however, only one volunteer regularly participated. But an article in New York Times had a transformational effect, more than doubling the amount of (partially) transcribed manuscripts. An overview of almost 3 years of the Transcribe Bentham project reveals that 17 people transcribed 0.6-20.7% of manuscripts, 10 continued participating, 7 were from UK – as was the organisation behind the project, University College London. (Causer and Terras in Ridge 2017, Chapter 3)

Memory institutions are located in different contexts from a collections point of view. Archivists organize records by collections, but an outreach strategy that limits visibility-raising efforts solely to the collection level is limited in its ability to reach numerous potential digital patrons. Therefore, engaging the users to add item-specific information would raise discoverability of the items (Szajewski 2013).

Libraries participate in numerous digitization projects (Galloway and DellaCorte 2014), and digitize continuously. Increasing the usage of these items would partly justify the cost and effort of digitization. Ridge (2017) mentions the sheer quantity of archival material and a desire to make better use of collections in the face of reduced funding for digitisation and other collections work as some of the institutional drivers behind the popularity of crowdsourcing.

The Library of Congress in the USA was a pioneer in the use of Flickr for crowdsourcing among memory institutions. Their collaboration led to development of

the Flickr Commons. In January 2008, the Library of Congress launched a project aimed to increase awareness by sharing photographs from the Library's collections with people who enjoy images but might not visit the Library's own Web site; to gain a better understanding of how social tagging and community input could benefit both the Library and users of the collections; and to gain experience participating in the emergent web communities that would be interested in the kinds of materials in the Library's collections. In less than 10 months, 67,176 tags were added by 2,518 unique Flickr accounts, and 4,548 of the 4,615 photos have at least one community-provided tag (Springer et al. 2008, Oomen and Aroyo 2011).

Museum collections online have not proved to be as engaging as they might be for the general public, and one reason might be the mismatch between the keywords attributed to the items and those in use by the general public (Trant and Wyman 2006). Museums have a lot of untextual items and acknowledge the potential of user-generated tagging in image indexing (Matusiak 2006). A more open classificatory system enables the 'networked object' to become part of the public culture, to be combined with other cultural forms, and for its meaning to be disputed. Moreover, it alters the public's perception of the museum as an open space, even if the actual form of participation in formalizing knowledge around collections remains limited (Geismar 2012, 270).

Social tagging appeals to museums because it embodies these self-directed learning philosophies by being a dialogue 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 (Trant and Wyman 2006). For example, the Steve.museum project, founded in the United States in 2005, like many similar initiatives at the time, was interested in harnessing the power of Web 2.0 in the museum by enabling tagging or user-generated taxonomies in describing collections. The Museum of New Zealand asks people to contribute images and associated documentation online, but then displays the content in the museum on a digital wall upon which visitors may choose from the archive and create their own cultural map of New Zealand, and which they can take along on a memory stick (Geismar 2012, 269–271).

The Powerhouse Museum in Australia serves as an example of on-going activity of tagging in an OPAC (on-line public access catalogue). It aimed to achieve better resource discovery as user tags allowed the recommendation of related objects by their classification by other users of the Web site. The user keywords were found to

be most often generally descriptive, allowing users to discover objects that are difficult to discover through the Museum's formal classification systems: e.g. a search for 'model train' would usually neglect to find objects formally classified as 'model locomotive.' Furthermore, these tags can be spidered by search engines as pointers to object records, allowing for their discovery and use outside of the OPAC itself.

Specifically, the comma was used as a separator of multiple tags, tags were made immediately visible, and any user could remove tags, including those submitted by other users. Tags appeared on the site as hyperlinks and could be clicked to trigger a search for the tag. As a result, 3,928 tags were submitted between June and December 2006. Of these, 537 were deleted, edited for spelling, or removed by other users and the system administrator. 2,246 objects were tagged with 3391 tags (avg=1.51, sd=1.01, n=3391). Interestingly, none of the most tagged objects are on public display within the museum.

It was concluded that user tagging and folksonomies can be used to improve navigation and discoverability, but that they work most effectively when matched with detailed collection records and balanced with the structural benefits of formal taxonomies. When combined with these features, search tracking can provide a means to improve serendipitous discovery and enhance the ability of users to find related objects and explore deeper into a collection (Chan 2007).

In order to improve discoverability, visibility of the items is often being increased first (Moffat 2006). Embedding and linking from high-traffic sites like Flickr or Wikipedia can raise visibility as well as awareness and usage (Calhoun 2014, 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 (Calhoun 2014, 24). For example, rather than edit articles, the staff at the University of Houston were primarily interested in contributing visual images to Wikipedia to accompany already existing articles, and were successful in terms of referrals to them (Galloway and DellaCorte 2014).

Likewise, a study of the users of digitized Hague Sheet Music shows that they are often interested in specific songs and songwriters. Their discovery of assets in the 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 (Szajewski 2013).

In conclusion, Ridge (2017) notices that the key trend in cultural heritage crowdsourcing is the pace and depth of constant change. The challenges reside in improvements in machine learning and computational ability to deal with tasks that were previously performed by people. Crowdsourcing projects continue to evolve to meet these challenges and changes in the digital and social landscape, and the impact on cultural heritage crowdsourcing remains to be seen.

2.4. Research Topic

Feynman pointed out in his lecture (1963) some paradoxes of evolution – we are happy with the development of medicine, but then get worried about the number of births; or we are happy about the development of air transportation, but get worried about the horrors of it. In the field of information and communication technologies (ICT) today, we develop powerful tools like the internet and all the platforms and applications built on it, and then we are worried about misuse, noise, or info-war like that accompanying the Russian-Georgian conflict in 2008. But let us focus on the positive power or, as Feynman puts it, on the value. The internet empowers people and this power can be used for the common good.

The literature refers to the knowledge gap in how society can contribute to enhancing collaboration and partnerships among and across individuals and institutions. More evidence is offered on crowdsourcing projects, on-going work in a single institution, or collaboration within one type of memory institution, but the evidence on the structural differences regarding user engagement in different types of memory organisations remains underrepresented.

The social phenomenon analysed in this study is user engagement with digital collections. This explorative and comparative study focuses on two types of memory institutions: a library and an archive. A very simple example of the common good in this context is a descriptive tag, which is attributed by a user to a digitised photo that has been made visible by the respective institution. People's knowledge of the collection items would lead to more holistic view of the collections as a result of collective action. Tags become structured metadata while being processed and, in turn, searchable data becomes useful for a number of fields, from scientific to artistic use and exposure.

As mentioned above, the key factor of discoverability is metadata attributed to the items and made available via appropriate infrastructures. The history of science has focused mainly on physics and the objects that other sciences study are themselves made up of physical entities (Okasha 2002). So are the digital platforms of memory institutions – resulting from mathematics and applied computer science. In the current project, it is planned to explore the digital platforms qualitatively from a social science point of view, while paying attention to their physical affordances.

According to Trant (2009), the research of social tagging and folksonomy can be divided into three broad approaches, focusing first, on the folksonomy itself and the role of tags in indexing and retrieval; secondly, on tagging and the behaviour of users; and thirdly, on the nature of social tagging systems as socio-technical frameworks. The second approach is the key to the current project.

2.4.1. Research Questions

Science is a method of finding things out (Feynman 1963). In this project, the desire is to find out the relationship between users' participation and the discoverability of digital collections. Our starting point is that user interaction is mediated. As discussed above, crowdsourcing projects are focused to achieve certain goals and provide limited tools for users to interact, so therefore give no solid overview of the variety of the potential user community and their preferences. This research project reaches out to more universal systems than specially-designed crowdsourcing platforms.

The main research question is: **How are users' participation and the discoverability of digital collections related?** The question implies to a hypothesis that user interaction is useful for discoverability, but on the other hand the question is broad enough to conclude the opposite, if needed.

The following sub-questions are posed:

- Where does the user interaction take place?
- What do the institutions enable the users to do in those mediums and why?
- How do users interact under those conditions and why?
- Who are the users?
- What kinds of relationships exist between the type of organisation and platform?

For answering these questions, the explanatory variables to be studied include the institutional setting, the exploited platforms and the enabled features for social interaction. Dependent variables include the amount and nature of contributions, and the interactions between users and organisations.

A high level of uncertainty is coded in the approach to compare a library and an archive. There are differences in the professional practices, both with long histories. Differences also concern institutional cultures internally and the ways to interact with their user communities externally. But we can let this uncertainty stay as encouraged by King et al. (1994) and also for the two following reasons.

Firstly, according to Feynman (1963), the more specific the rule or the more definite the statement, the more interesting it is to test. The current research project is not aimed at providing, and cannot provide, a best practice model based on two case studies, but the promising amount of detailed and structured information of the two studies would still allow numerous institutions to think hard about Feynman's grand question: "If I do this, what will happen?"

Secondly, ICT is developing fast and thus the current project relies on the statement of urgency, referred to by King et al (1994). Users of the memory organisations are never just the users of those organisations, but experience interaction daily within a number of situations in different areas. This directs the focus to the users instead of the long-developed professional practices of archives and libraries. The gathered data on user behaviour is to be disseminated to the organisations fast in order that it not become obsolete before it is accumulated.

Before laying out the methods to answer the research questions, the conceptual framework is introduced in the next chapter. To some extent, it is about challenging the digital anthropological approach by giving importance to the technological affordances as mediating tools, but still discussing the social context around it.

3. Conceptual Framework

This research is informed by activity theory. The current chapter introduces the orientation of the framework and a conceptualisation of it, which is used for analysis of the interviews in Chapter 6 and for drawing more general conclusions at the end. Secondly, the concept of participatory organisations is discussed and the models of participation are referred to throughout the following dissertation. The concepts of activity theory and participatory organisations are not directly related to one another, but they are used in this research project as complementing each other in describing the complex phenomenon of social tagging.

3.1. Activity Theory

Activity theory is a philosophical and cross-disciplinary framework for studying different forms of human practices as development processes, with both individual and social levels interlinked at the same time. Like many psychological theories, it uses human action as the unit of analysis, while also considering the context of the action. And because the context is included in the unit of analysis, the object of the research is collective even if the main interest is in individual actions (Kuutti 1996).

The roots of activity theory lie in Soviet cultural-historical psychology, founded by Vygotsky, Leont'ev, and Luria (Kuutti 1996). Today, activity theory is an approach that has transcended both international and disciplinary borders. It is applied in psychology, education, work research, and other fields (Kaptelinin and Nardi 2006, 6).

Nardi (1996) argues that activity theory is a powerful and clarifying descriptive tool rather than a strongly predictive theory. The object of activity theory is to understand the unity of consciousness and activity, whereas consciousness is not a set of discrete disembodied cognitive acts (decision making, classification, remembering) but is located in everyday practice: you are what you do. And what you do is firmly and inextricably embedded in the social matrix of people and artefacts. Artefacts may be physical tools or sign systems such as human language. Understanding the interpenetration of the individual, other people, and artefacts in everyday activity is

the challenge activity theory has set for itself. Bødker and Klokmose (2012) explain that there is no one-to-one relationship between human activity and artefacts. Artefacts are crystallizations of activity both as externalizations of operations carried out with earlier artefacts, and as representations of modes of acting in the given activity.

Mediation of human activity is one of the core concepts of activity theory, which may contradict digital anthropology that sees the digital as not merely technology, and is in opposition to approaches that imply that becoming digital has either rendered us less human and authentic or more mediated (Miller and Horst 2012, 29, 4).

Activity theory distinguishes between people and things, allowing for a discussion of human intentionality. The artefacts are designed and used intentionally. Thus the activity is always object-oriented. And the activity belongs always to a subject, who has agency – a need to act (Kaptelinin and Nardi 2006, 10, 32).

The level of action is hierarchically organized and the constituents of activity are not fixed but dynamic, and this can change as reality changes (Kaptelinin and Nardi 2006, 67–68). Thus activities are subject to change and development, which "is not linear or straightforward, but uneven and discontinuous," meaning that each activity has a history of its own (Kuutti 1996).

Kaptelinin and Nardi (2006, 10) refer to tenets of activity theory which are relevant to explain the needs of users and the possibilities of technologies: an emphasis on human intentionality; the asymmetry of people and things; the importance of human development; and the idea of culture and society as shaping human activity. These notions highlight the importance of taking context into consideration.

3.1.1. Tools for Analysis

Under these principles, several models provide tools for analysis. Models based on Leontiev's approach focus on individual activities.

The human-artifact model was proposed for the activity theoretical human-computer interaction domain. The model focuses on the artefacts that human beings use and the practices that these reflect. The model is intended to highlight tensions between intended action possibilities in the artefacts and the action possibilities expected by the user: the assumptions of use embedded in the artefact on the one hand, and the experiences or orientation of the user on the other. Bødker and

Klokmose (2012) refer to Engeström, who argues that such change processes are not fully predictable – when a new artefact is designed, its use cannot be predicted. The iteration is drove by the difference between the future artefact as it was imagined and conceptualized, and the actual artefact as it functions in consolidated use. Bødker and Klokmose (2012) believe that the practice of users needs to be brought into design, and kept there to continuously confront design visions and prototypes.

The analytical scheme of the human-artefact model combines analyses of action possibilities and mediators on three levels reflecting the activity hierarchy: activity, action, and operation. It enables the summarising of empirical findings about artefacts and humans, and the detection of matches and contradictions across the levels: motivational aspects (for artefacts) / motivational orientation (for humans) (the core question is *why*?); instrumental aspects / goal orientation (*what*?); operational aspects, including handling aspects / operational orientation, including learned handling (*how*?); and adoptive aspects / adaptation (*how*?) (Bødker and Klokmose 2011).

Activity theoretical idea that the human users possess a set of learned and adapted action possibilities and they are oriented towards certain goals and motives led Bødker and Klokmose (2013) to confront personas (i.e. the human side of the human-artefact model) to techsonas (i.e. the artefact side). Personas allow for an understanding of the values, fears, etc., of the users, which help in answering the question of how potential users reach out towards current or hypothetical future technologies. Techsonas similarly help in addressing the artefact side: how do the action possibilities of an artefact, whether a current or hypothetical artefact, reach out towards the potential user? Techsonas are needed as counterparts to personas and scenarios in order to help users and designers reason about technology on a more abstract and hypothetical level than prototypes. Like a persona, a techsona is considered hypothetical, yet rigorous and precise enough to be confronted with a variety of personas and scenarios. Applying the human-artifact model as a mediator between personas and techsonas emphasizes their dialectical relationship, and provides a structure for systematically exploring the tensions between the assumptions of the artefact and the capabilities and orientation of the user towards the artefact all the way from motivation to low-level operation (Bødker & Klokmose 2013).

The human-artefact model and similar tools focus on interfaces and evaluate the human interaction with them. For instance, an activity checklist has been used for the

analysis, evaluation, and design of a variety of technologies as it lays out a "contextual design space" by representing the key areas of context specified by activity theory (Kaptelinin and Nardi 2006, 98). But if we do not want to direct attention to the individual use of the artefact, but to run a more general explorative enquiry of the themes around social interaction with digital collections in which both organisations and community of users are seen as collective entities forming a network of activities, other options should be used.

The model proposed by Engeström describes activity as a collective phenomenon, where individuals carry out actions within a larger-scale collective activity system. It is created as a tool for describing units of complex mediated social practices, clearly identifying key aspects of the described reality, pointing to potential contradictions, and providing a visual representation indicating how these aspects are related to each other (Kaptelinin and Nardi 2006, 99–100).

The conceptualisation by Yrjo Engeström differentiates seven interrelated actors that form an activity system (Fig. 3.1, adapted from Engeström 1990 by Kaptelinin and Nardi 2006, 100). The *subject*, i.e. the human cannot control its behaviour from the inside, on the basis of biological urges, but from the outside, using and creating artefacts or *tools*, which mediate the subject and the object and which can be material as well as immaterial (e.g. tools for thinking). An *object* can be a material thing or less tangible (such as a plan) or totally intangible (such as a common idea), as long as it can be shared for manipulation and transformation by participants of the activity. *Community* shares the same object and its relationship with subject is mediated by *rules*, which cover both explicit and implicit norms, conventions, and social relations within a community. The relationship between object and community is mediated by the *division of labour*, which refers to the explicit and implicit organization of a community as related to the transformational process of the object into the outcome. The activities may form networks, matrixes or other systems noticing their relationships with each other (Kuutti, 1996).

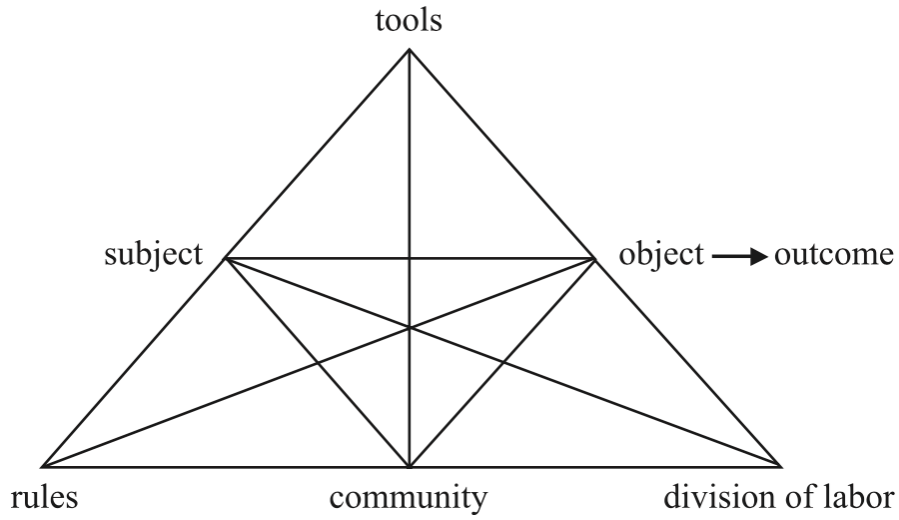

Figure 3.1. The activity system model (Kaptelinin and Nardi 2006, 100).

3.2. Participatory Organisation

The tendency of memory institutions is to move towards engaging audiences and giving them certain spaces for interaction with the organisation or with each other. The interaction can take place either on site or online, and is targeted to fulfil a specific goal for the organisation or simply switched-on functionalities because the technology enables the organization to do it. In any case, these kinds of experiments enhance the elements of a participatory institution, if not necessarily embracing the whole concept.

A participatory cultural institution is a place where visitors can create, share, and connect with each other around content. *Create* means that visitors contribute their own ideas, objects, and creative expressions to the institution and to each other. *Share* means that people discuss, take home, remix, and redistribute both what they see and what they make during their visit. *Connect* means that visitors socialize with other people – staff and visitors – who share their particular interests. *Around content* means that visitors' conversations and creations focus on the evidence, objects, and ideas most important to the institution in question (Simon 2010, ii–iii). All these components can happen also online.

Simon (2010, 26–27) differentiates 5 stages from individual to collective experiences:

1. Individual Consumes Content – visitors are granted with access to the content
2. Individual Interacts with Content – provides an opportunity for inquiry and for visitors to take action and ask questions

3. Individual Interactions are Networked in Aggregate – lets visitors see where their interests and actions fit in the wider community of visitors to the institution

4. Individual Interactions are Networked for Social Use – helps visitors to connect with particular people, like staff members and other visitors, who share their content and activity interests

5. Individuals Engage with Each Other Socially – makes the entire institution feel like a social place, full of potentially interesting, challenging, enriching encounters with other people

It is important to reach stage 5 in order to facilitate visitors helping each other. Then a community will have emerged to whom the institution can become more meaningful. It would take an audience-centric approach through personalised contact, allowing visitor profiles so that other visitors can relate, and providing tools to connect the individuals (Simon 2010, 84).

3.2.1. Participatory Models

Citizen science, where volunteers from the general public assist scientists in conducting research has been an influential model for humanities and citizen history crowdsourcing projects (Ridge 2017). The Center for Advancement of Informal Science Education (CAISE) defined three broad categories of public participation in scientific research aiming to provide multiple opportunities to increase public science literacy (Bonney et al. 2009). Simon (2010) has adapted these categories as participatory models for museums, but often refers to cultural institutions in general:

1. Contribution – visitors can provide limited and specified objects, actions or ideas to an institutionally controlled process;

2. Collaboration – visitors are invited to serve as active partners in the creation of institutional projects that are originated and ultimately controlled by the institution;

3. Co-creation – community members work together with institutional staff members from the beginning to define the project's goals and to generate the program or exhibit based on community interests.

Simon (2010, 185-187) adds a fourth category:

4. Hosted – an institution turns over a portion of its facilities and/or resources to present programs developed and implemented by public groups or casual visitors; online, programmers may use cultural object registries or scientific data as the basis for their own research or products.

These participatory models are distinct, but many institutions incorporate elements from each of them.

Participatory techniques represent an addition to the design toolkit, not a replacement for traditional strategies (Simon 2010, 349). The models are now explained.

Contributory projects are often the simplest for institutions to manage and for visitors to engage in as participants. These can be offered to visitors of all types without much setup or participant coaching, and many are self-maintaining. After their contribution (e.g. comment or post), visitors immediately see how they have contributed to the institution. There is a wide audience to use the contributed content – participants, spectating visitors, stakeholders, and researchers. Institutions may see the contributions as necessary – in which the success of the project relies on the visitors' active participation; as supplemental – in which the visitors' participation enhances an institutional project; or as educational – in which the act of contributing provides visitors with skills or experiences that are mission-relevant.

When a contributory project relies on visitors' contributions to succeed, it generates a high risk of failure, if visitors do not participate, and a level of high institutional investment with a motive to put more thought and commitment into project design to avoid that failure. Unlike projects of necessity, in which institutions often introduce constraints to ensure consistency of contributions, supplemental projects thrive when visitors are given license and encouragement to be creative or share strong reactions. Visitors who enjoy trying and learning new things are particularly drawn to educational contributory projects, which value the teaching of new skills over the contributions.

Generic requests to share a story or draw a picture are not as successful as those that ask visitors to contribute something specific under clear constraints. From a participant perspective, a good contributory project provides clear, specific opportunities for visitors to express themselves, scaffolds the contributory experience to make participation accessible regardless of prior knowledge, respects visitors' time and abilities, and clearly demonstrates how visitors' contributions will be displayed, stored, or used. The visitor-generated content is curated, removing content that staff members perceive as inappropriate or offensive, or creating a product that presents a focused set of contributions (Simon 2010, 204–225).

Collaborative projects are institutionally driven partnerships in which staff members work with community partners to develop new programs, exhibitions, or offerings. Participants may be chosen for specific knowledge or skills, association with cultural groups of interest, age, or representation of the intended audience for the output of the project. Institutions might want to engage in collaborative projects to consult with experts or community representatives to ensure the accuracy and authenticity of new exhibitions, programs, or publications; to test and develop new programs in partnership with intended users to improve the likelihood of their success; to provide educational opportunities for participants to design, create, and produce their own content or research; or to help visitors feel like partners and co-owners of the content and programs of the institution. Because collaborations often involve prolonged formal relationships between institutions and participants, institutions typically give participants more guidance than is provided in contributory projects.

Collaborative projects are broadly either consultative projects – in which institutions engage experts or community representatives to provide advice and guidance to staff members as they develop new exhibitions, programs, or publications; or they are co-development projects, in which staff members work together with participants to produce new exhibitions and programs. The roles of staff in collaborations include project directors, who manage the collaboration and keep the project on track; community managers, who work closely with participants and advocate for their needs; instructors, who provide training for participants; and client representatives, who represent institutional interests and requirements. It is particularly important to separate out instructors and client representatives from other project staff as they are authority figures, not partners.

While collaborative exhibition projects support creative skill building and story sharing, research collaborations support other skills like visual literacy, critical thinking, and analysis of diverse information sources. Integrating collaboration into visitor experiences makes participation available to anyone, anytime. On the web, Wikipedia is a good example of this kind of evolving, 'live' collaborative platform. At any time, non-contributing users can access and use the content presented while authors and editors continue to improve it. The ideal collaborative cultural experience is comparable: appealing to visitors, with a thin and permeable division between spectating and actively collaborating (Simon 2010, 231–256).

Co-creative projects originate in partnership with participants rather than being based solely on institutional goals. Cultural institutions strive to give voice and be responsive to the needs and interests of local community members; to provide a place for community engagement and dialogue; or to help participants develop skills that will support their own individual and community goals. Co-creative projects progress similarly to collaborative projects, but they confer more power on participants and result in projects that are co-owned by institutional and community partners. These projects may run into trouble when participants' goals are not aligned with institutional goals, or when staff members are not fully aware of participants' goals at the outset. The discussion should not jump to the 'how' of participation without a full investigation of the 'why.'

Co-creative projects challenge institutional perceptions of ownership and control of content. They require "radical trust" in community members' abilities to perform complex tasks, collaborate with each other, and respect institutional rules and priorities. Staff must not only trust the competencies and motivations of participants, but deeply desire their input and leadership (Simon 2010, 263–274).

In hosted projects, the institution turns over a gallery or a program to community partners. In addition to institutional partnerships, which lead to hosting, institutions can also be used or repurposed by amateur groups and casual visitors. Institutions may choose to pursue hosting models for participation to encourage the public to be comfortable using the institution for a wide range of reasons; to encourage visitors to creatively adapt and use the institution and its content; to provide a space for diverse perspectives, exhibits, and performances that staff members are unable or unwilling to present; or to attract new audiences who may not see the institution as a place for their own interests (Simon 2010, 281–298).

A precursor to evaluating participatory practice is articulating participatory goals: i.e. institutions should have a clear list of goals either per project or per participatory practice as a whole. Participatory outcomes can then be distinguished: the behaviours that the staff perceives as indicative of goals being met, e.g. an exhibition is an output, which may or may not bring along desired outcomes (Simon 2010, 304–305).

3.3. Conclusion of the Chapter

This chapter introduced two frameworks: activity theory and participatory organisations.

Engeström's approach of activity theory is relevant to this study for its collective nature, which can be applied both to community of staff and users. His conceptualisation of activity theory allows placing an activity at the centre and monitoring the context related to it. While a human-computer interaction domain would examine a specific activity in order to improve an application, for example, then the idea behind the concept also enables the formulation of an abstract action – social tagging – to study the context around it from different perspectives.

Simon's participatory models can be seen as practical formations, which are suitable to study from the activity theory's point of view. A participatory act is a unit of analysis according to this conceptualisation. The participatory models help to categorise the examples of current practice mentioned by the staff and make the research findings more easily comprehensible and applicable in practice.

Activity theory can be placed in relation to any models that provide a human activity with minimal context surrounding it. This research project makes an attempt to use the concept of participatory models for capturing certain types of activity as an input for activity theoretical study. However, both concepts are useful to frame the research inquiry, analyse the empirical evidence, and discuss the results.

The next chapter unfolds the inquiry itself by introducing the approach, research design, and methods to find answers to the research questions.

4. Methods

In this chapter, the case study approach is addressed first and the choice of the case study organisations is explained. The organisations are introduced accordingly in sections 4.1.1. and 4.1.2. Section 4.2. presents the methods for data collection and analysis. The chapter is concluded in section 4.3.

4.1. Case Studies

As outlined by the 'State of Art' (Chapter 2), the unit of comparison is the type of the organisation from the GLAM sector. From the interaction's and digital collections' point of view, archives and libraries are closer to one another than either would be with galleries or museums; therefore, it would be meaningful to study more similar institutions first. If they are similar, future research could take the next step in the direction of diversity. Thus the current project launches a comparative approach between a library and an archive.

Both cases were selected from the same country in order not to be misled by the cultural or political differences and thereby interpret them as arising from the type of the organisation. In order to fully understand the cases, the researcher must be able to understand the language of the regulatory acts, the interfaces, and the content of user contributions. The potential amount of the total information was predicted too voluminous to consider translation. Therefore, two options remained: English or Estonian speaking countries.

Estonia is considered “the champion in Europe in the online provision of public services and scores above EU average in digital skills and the use of internet by citizens” (European Commission 2017) by the Digital Economy and Society Index (DESI)[5]. The index measures digital progress through connectivity (fixed broadband, mobile broadband, broadband speed and prices), human capital (basic skills and internet use, advanced skills and development), use of internet (citizens' use of content, communication and online transactions), integration of digital technology (business digitisation and eCommerce), and digital public services (eGovernment). Estonia ranks 9th after the Scandinavian and Benelux countries, UK and Ireland. This

[5] https://ec.europa.eu/digital-single-market/en/desi

kind of profiling could raise the case for selecting Estonia as a case study organisation, unless the practice of libraries and archives does not stand out by user engagement under the set conditions, i.e. on-going integrated interaction on a significant scale. Why it does not stand out on a national scale is a matter for a different research question. In order to study the country context close to the home (University of Milano-Bicocca) and host universities (Tallinn University), an English-speaking country was preferable over other options demonstrating explicit occurrences of the phenomenon of social interaction with digital collections. Thus the United Kingdom was selected.

The organisations were aimed to be selected from a comparable level. Two national institutions were considered: the British Library, and the National Archives of the United Kingdom. As illustrated in the next two sub-sections, both organisations have variety of options for user engagement; therefore, the two organisations were selected as case studies.

4.1.1. The British Library

The British Library is the national library of the United Kingdom and one of the greatest research libraries in the world. The Library was founded in 1973 and is located in London. The Library's collections count up to 180 million items, including books, magazines, manuscripts, maps, music scores, newspapers, patents, databases, philatelic items, prints and drawings and sound recordings. There are 3.6 million library users on site and through virtual facilities, and more than 13 million unique hosts are served, excluding staff users (The British Library 2015). For reasons of history, the collections of the British Library have relevance and meaning to people and communities around the world, which is why a responsibility is seen in bringing these collections to life through digital technologies and sharing them far beyond the walls of the library (Brazier 2016).

Even if the British Library can claim to be relatively young for a national library, much of its vast historical collection was acquired long before computers were used for cataloguing. While the library's online catalogue, Explore, provides public access to nearly 57 million records, there are still thousands of items that can only be found by searching the physical card catalogues. In order to convert these into a searchable form, the Library uses crowdsourcing[6].

[6] https://www.libcrowds.com

User satisfaction surveys show that users have high expectations of what they want to experience on the Library's websites. 86% (target 88%) of visitors described the ease of finding information on the Library's website as 'excellent,', 'very good,' or 'good,' compared to 94% (target 92%) of readers who were either 'very satisfied' or 'quite satisfied' with the services and facilities they used, and 96% (target 92%) of visitors rating the enjoyment of their visit as either 'excellent' or 'good' for the exhibition visitor enjoyment rating (The British Library 2015).

The British Library website (The British Library) lists six "main digital collections" (retrieved June 2016–Jan 2017). Three collections provide a sharing tool, and four of them are complemented by a social media account. The Sounds collection enables tagging and for a limited time it enabled content contribution.

- Endangered Archives[7] is a research-oriented collection and presents sound, images, and textual works. Users' participation is not enabled on the platform, but a Facebook group[8] (a public group with over 2,000 members) and a Twitter feed[9] are run.
- Renaissance Festival Books[10] is a collection complemented with expert explanations. Tools for users' participation or related social network sites (SNS) were not noticed.
- The International Dunhuang Project[11] features manuscripts, images, etc., which are stored at the Library, but which derive from collaborations with 20 partners across the world, including museums, libraries, and research centres. Users can share the items and donate to the project. A Facebook page[12] (over 2,000 followers) and a Twitter feed[13] are run.
- Digitised Manuscripts[14] is a research-oriented collection for manuscripts and archives. No social tools or SNSs were noticed.
- Sounds[15] captures sound recordings. Social tools include sharing, tagging (requiring a log in), and reporting of problems. The UK Soundmap, the first nationwide sound map, invited people to record the sounds of their

[7] http://eap.bl.uk/index.a4d
[8] https://www.facebook.com/groups/6273301581
[9] https://twitter.com/bl_eap
[10] http://www.bl.uk/treasures/festivalbooks/homepage.html
[11] http://idp.bl.uk
[12] https://www.facebook.com/International-Dunhuang-Project-269876243911
[13] https://twitter.com/idp_uk?lang=en
[14] http://www.bl.uk/manuscripts
[15] http://sounds.bl.uk

environment, be it at home, work, or play. Over 2,000 recordings were uploaded by 350 contributors during the period of July 2010 to July 2011 (The British Library. Sound map). A Twitter feed is maintained[16].

- EThOS[17]is for doctoral theses. Users can share findings. The Twitter feed[18] provides direct links to items and re-tweets news related to PhD research.

The Library's website lists 4 main web catalogues (retrieved 20.06.2016). Two of them enable the attribution of tags and notes.

- Explore, the British Library catalogue[19], searches 57 million records for books, journals, newspapers, printed maps, scores, electronic resources, sound archive items, etc. Authorised users can log in and add notes and tags.
- The Sound and Moving Image Catalogue[20] has 3,5 million recordings in genres from pop, jazz, classical, and world music to wildlife sounds, oral history, drama, literature, language, and dialect. No social tools or SNS was noted.
- The Archives and Manuscripts catalogue [21] includes manuscripts and unpublished documents, personal papers, correspondence and diaries, family and estate papers, India Office records and private papers, and India Office prints. Drawings, paintings, and photographs were marked as 'coming soon' at the time of retrieval. Authorised users can add notes and tags while logged in. SNS was not evident.
- The British National Bibliography[22] includes books and new journal titles published or distributed in the United Kingdom and Ireland since 1950. No social tools nor SNS were evident.

Interaction with users is encouraged by a number of blogs on different topics. 14 out of 19 listed blogs on the Library's website (retrieved in June 2016) refer to collections, another 6 are either not active (e.g. were related to a specific staff member who has since left their position), post about work procedures or other matters that do not relate directly to the Library's collections, and one blog is about the UK Web Archive. All 14 collection-related blogs enable sharing (contribution to visibility); 12

[16] https://twitter.com/BLSoundHeritage
[17] http://ethos.bl.uk/Home.do
[18] https://twitter.com/EThOSBL
[19] http://explore.bl.uk/primo_library/libweb
[20] http://cadensa.bl.uk
[21] http://searcharchives.bl.uk/primo_library/libweb
[22] http://bnb.bl.uk

of 14 offer Twitter feed (contribution to visibility); 2 blogs are running 2 different Twitter feeds; and 2 blogs also offering a Facebook page. Some major events, like a Shakespeare exhibition, give grounds to several blog posts in different blogs from their respective angles. Posts by subject librarians, other staff members, and sometimes by guest authors are thorough and contribute to knowledge sharing. Posts often link to external pages, including sites like Wikipedia, YouTube, and others. The focus on the collections is important for the current research topic, i.e. the interaction with digital resources. The collections-related blogs (retrieved 17.06.2016) were the following:

- The American Collections blog[23] reports on the news on the work done, events, collections (links to the viewer of digital resources), enables sharing and posting. A Twitter feed 'News on American Studies'[24] and a Facebook page[25] links to blog posts on news, events, and items.
- The Asian and African Studies blog[26] introduces the work of curators, recent acquisitions, digitisation projects, and collaborative projects outside the Library. It enables sharing and two related Twitter feeds are run: @blasia_africa[27] reports on the news, links to the blog and/or directly to items; and @bl_visualarts[28] retweets the blogs and shares thematic news.
- The Digital Scholarship blog[29] introduces upcoming events, Open Data initiatives, projects, collaborations, and experiments enabling innovative research based on the British Library digital collections. Users can share the posts and follow two Twitter feeds: @BL_DigiSchol[30] and @BL_Labs[31].
- The English and Drama blog[32]reports on news related to the collections. Users can share the posts and follow a Twitter feed[33], which also links to the blog.
- The European Studies blog[34] provides additional information from the internet regarding the collections. Posts can be shared and users can follow a Twitter feed[35] on news, events, and links to the blog posts.

[23] http://britishlibrary.typepad.co.uk/americas/index.html
[24] https://twitter.com/_Americas
[25] https://www.facebook.com/Team-Americas-at-the-British-Library-309366835752329
[26] http://britishlibrary.typepad.co.uk/asian-and-african
[27] https://twitter.com/blasia_africa?lang=en
[28] https://twitter.com/bl_visualarts
[29] http://blogs.bl.uk/digital-scholarship/
[30] https://twitter.com/BL_DigiSchol
[31] https://twitter.com/BL_Labs
[32] http://britishlibrary.typepad.co.uk/english-and-drama
[33] https://twitter.com/BLEnglish_Drama

- The Innovation and Enterprise blog[36] presents the news on topics, interviews, and help regarding databases and publications. Posts can be shared and a Twitter feed[37] can be followed.
- The Maps and Views blog[38] publishes collection-related news, and links to the BL Online gallery, Flickr and the internet. Posts can be shared, but no SNS is reported as linked to the blog.
- The Medieval Manuscripts blog[39] refers to the collections, links to digital resources, publishes related news on digitisation, events etc. Posts can be shared and a Twitter feed[40] can be followed.
- The Music blog[41] publishes news related to the music collection, events, exhibitions with links to the internet, and some links to digital resources. Posts can be shared. No SNS is directly referred to (but on this topic, see 'Sounds' under main collections).
- The Science blog[42] offers thematic news and reflections on events, as well as news related to services (databases, collections). Posts can be shared and a Twitter feed[43] with links to the blog, services, and other BL channels, and a Facebook page[44] with events and other timely content can be followed.
- The Social Science blog[45] provides thematic news, events, reflections, and referrals to the collections. Posts can be shared. Twitter feeds are run for Sport and Society[46] and on the Social Welfare Portal[47].
- The Sound and Vision blog[48] offers news, links to subject portals, the repository and external platforms. Posts can be shared and a Twitter feed[49] followed for links to blog and timely content.

---

[34] http://britishlibrary.typepad.co.uk/european
[35] https://twitter.com/BL_European
[36] http://britishlibrary.typepad.co.uk/business
[37] https://twitter.com/BIPC
[38] http://britishlibrary.typepad.co.uk/magnificentmaps
[39] http://britishlibrary.typepad.co.uk/digitisedmanuscripts
[40] https://twitter.com/BLMedieval
[41] http://britishlibrary.typepad.co.uk/music
[42] http://britishlibrary.typepad.co.uk/science/index.html
[43] https://twitter.com/ScienceBL
[44] https://www.facebook.com/Science-The-British-Library-352385414037
[45] http://britishlibrary.typepad.co.uk/socialscience/index.html
[46] https://twitter.com/BLsportsociety
[47] https://twitter.com/blsocialwelfare
[48] http://britishlibrary.typepad.co.uk/sound-and-vision/index.html
[49] https://twitter.com/soundarchive

- The Newsroom blog[50] discusses methods to use old newspapers, events, people, and minor links to collections. Posts can be shared and a Twitter feed[51] can be followed for examples on news representations in history and today.
- The Untold Lives blog[52]makes insights into history and links to collections. Posts can be shared and Twitter feed[53] followed for links to blogs and retweets of similar content.

Two platforms (retrieved in June 2016) are specially designed for crowdsourcing:

- LibCrowds offers the task for collaborators to look at digitised Library Catalogue Cards and to find existing records in WorldCat. If no similar records were found, the card can be transcribed. 624 volunteers (from 6 continents, 41 countries and 227 cities) have participated in 11 projects, made 30,216 contributions (the majority coming from the UK at 12,632) and completed 9,359 tasks. The 10 most active users have made altogether 11,349 contributions. The reason why some specific countries like India or Indonesia are among the top 4 countries by visitors after the United Kingdom and the United States is that there have been specific tasks related to these locations: e.g. Indonesian Card Catalogue: Drawer Three[54]. Twitter feed promotes the tasks and reports on progress[55].
- Georeferencer[56] crowdsources since 2012 location data to make a selection of its vast collection of maps fully searchable and viewable using online geotechnologies. 8,000 maps were placed by users before the sixth release of maps from public domain books in 2015, of which 39% of over 50,000 maps are georeferenced.

Both LibCrowds and Georeferencer can be characterised by launching a set of tasks at a time for contributions. Georeferencer is related to the Library's Flickr collection[57], which is closely examined in Chapters 5 and 6 as an activity of BL Labs. The BL Labs project is part of the Digital Scholarship department of the British

[50] http://britishlibrary.typepad.co.uk/thenewsroom
[51] https://twitter.com/BL_newsroom
[52] http://britishlibrary.typepad.co.uk/untoldlives
[53] https://twitter.com/UntoldLives
[54] https://www.libcrowds.com/project/indonesiancardcatalogue_d3
[55] https://twitter.com/LibCrowds
[56] http://www.bl.uk/georeferencer/
[57] https://www.flickr.com/people/britishlibrary

Library. It works with researchers, artists, and software developers to actively engage users with the Library's digital content and data (The British Library, 2015).

4.1.2. The National Archives of the UK

The National Archives was founded between 2003 and 2006 by joining four historic government bodies: the Public Record Office (the national archive of England, Wales and the United Kingdom government), the Royal Commission on Historical Manuscripts (which performs the same functions in relation to private records), Her Majesty's Stationery Office (holder of the Crown copyright and official printer of all Acts of Parliament), and the newly created Office of Public Sector Information (promoting the re-use of information produced and collected by public sector organisations) (The National Archives. Our history).

The key target groups are government, the public, the wider archives sector, the academic community, and others engaged in scholarly research (The National Archives 2017). In 2010-11, over 120 million records were delivered to over 20 million online users, and for every document delivered in the reading rooms at Kew, 200 were delivered online (The National Archives. Volunteering...).

The Archives runs the main catalogue Discovery[58], which enables users to tag records upon free online registration, suggest corrections for errors, and share the records. The catalogue is accompanied by 12 thematic research guides[59] (including 'Online collections' of the most popular items) with an option to browse subject and keywords alphabetically. The Archives also captures, preserves, and makes available the UK Government Web Archive[60].

The Archives manages six Twitter accounts (The National Archives. Social media use):

- @UkNatArchives[61] tweets about blogs, news, podcasts, publications and file releases, and the latest videos and publications uploaded to YouTube and Flickr.
- @KIMexperts[62] tweets about information and records management, and cyber-security

[58] http://discovery.nationalarchives.gov.uk
[59] http://www.nationalarchives.gov.uk/help-with-your-research/research-guides-keywords/
[60] http://www.nationalarchives.gov.uk/webarchive/
[61] https://twitter.com/UkNatArchives
[62] https://twitter.com/kimexperts

- @TNAmediaofficer[63] promotes the Archives' appearances in media and makes occasional links to collections.
- @ExploreArchives[64] is the feed for the joint campaign Explore Your Archive[65] with the Archives and Records Association (UK & Ireland).
- @UkWarCabinet[66] tweeted about the unfolding events of the Second World War though the original Cabinet Papers from 1945, but is no longer active.
- @UnitWarDiaries[67] tweets accounts of daily events on the front line from the First World War as told though unit war diaries.

The two Facebook accounts of the Archives are:

- The National Archives[68], which posts updates about new content on the website.
- A community page, The National Archives Education Service[69], aimed at children, teachers, and parents interested in using The National Archives' education resources.

In addition, the following channels relating to the collections are maintained:

- Instagram account[70] featuring images and videos of records and on-site exhibitions.
- YouTube channel[71] for short videos highlighting interesting stories from the collection, new file releases and archival footage.
- Flickr account[72] with a request to help identify unknown subject matter, and Flickr group[73] to upload and share images taken by the users of the documents.
- Pinterest board[74] to find and share images from or relating to the Archives.
- Archives Media Player[75] to find, play and download audio and video podcasts. Users can add comments and star ratings to individual podcasts.

[63] https://twitter.com/TNAmediaofficer
[64] https://twitter.com/explorearchives
[65] http://www.exploreyourarchive.org
[66] https://twitter.com/ukwarcabinet
[67] https://twitter.com/UnitWarDiaries
[68] http://www.facebook.com/TheNationalArchives
[69] http://www.facebook.com/TheNationalArchivesEducationService
[70] https://www.instagram.com/nationalarchivesuk
[71] http://www.youtube.com/nationalarchives08
[72] flickr.com/photos/nationalarchives
[73] flickr.com/groups/nationalarchives
[74] https://pinterest.com/uknatarchives/
[75] http://media.nationalarchives.gov.uk/

- The National Archives' blog[76] features contributors from across the Archives, posting about their work and the wider archives sector.
- Periscope[77] @UkNatArchives for live broadcasts.
- In History pin[78], users are invited to pin their history to the world and explore images from the collection geographically.
- Glamwiki[79] welcomes contributions and improvements of the articles related to the Archives and its collections on Wikipedia.

The Archives has a long tradition of working with volunteers for cataloguing, conservation, and various project-based activities. Online volunteering takes place in Flickr regarding the description of the photographic collections. For about four years the Archives ran a wiki platform, Your Archives, as a community for record users to share their knowledge of British history, the collections, as well as other archival sources. Your Archives was built using MediaWiki, the same technology pioneered by Wikipedia. Contributors wrote almost 21,000 articles, which were collectively viewed over 48 million times. However, the rate of active participation was very low – on average, 0.5% of users contributed to the wiki. The aim of developing a knowledge-sharing community was not met, rather than creating another learning resource. Development of Your Archives was therefore suspended in order to pursue the integration of user collaboration functionality into the Discovery catalogue (The National Archives. Volunteering...; Grannum 2011).

4.2. Data Collection and Analysis

In order to design a comparable inquiry, the following sample was sieved from the overviews of the two organisations:

- the main catalogue Explore of the British Library (BL)
- the main catalogue Discovery of the National Archives (TNA)
- the catalogue Archives and Manuscripts of BL
- the Flickr collection of BL
- the Flickr collection of TNA.

[76] http://blog.nationalarchives.gov.uk/
[77] https://www.periscope.tv/
[78] https://www.historypin.org/en/person/31500/
[79] https://en.wikipedia.org/wiki/Wikipedia:GLAM/The_National_Archives

As illustrated on Figure 4.1 this selection enables not only comparison between the two organisations, but also between the type of collections (archival-library) and type of the platforms (catalogue-social network site).

		Type of collection	
		Archival	Library
Platform	Catalogue	Archives and Manuscripts (BL)	Explore (BL)
		Discovery (TNA)	EOD Search*
	Flickr	TNA	BL

Figure 4.1. Dimensions of comparison.

*EOD Search[80] is a consortial search engine of the eBooks on Demand (EOD) library network and was included in the analysis to give comparable data to the BL catalogues, because time of tag attribution was not recorded for other platforms.

A dataset with authentic user interaction data was needed for each case and was collected as follows: the Library, the Archives and the EOD network provided datasets for the catalogues on request; the Library had previously made the dataset for Flickr available on the web through figshare.com (O'Steen 2016); and similar data was retrieved through Flickr API for the Flickr page of the Archives.

Next, the document analysis of the interfaces, related help pages, and documents was carried out to reveal the conditions under which the users contributed tags. Then statistical analysis of the retrieved data was conducted by using R (R Core team 2016).

Observation is the ultimate judge of the truth of the idea (Feynman 1963), but it was not possible to observe a) the rich context around the institutions, and b) the context of individuals related to social tagging. Therefore the user data and document analysis were complemented by interviews.

For the interviews with staff, two employees from each organisation were selected for a separate in-depth interview. The selected sample of interviewees included people with experience and knowledge about user engagement regarding online collections. Two people from the same institution but with different

[80] search.books2ebooks.eu

experiences were asked for the interview, because the results were believed to be complementary to one another.

The interviews were recorded and transcribed and thematic analysis was launched according to the actors in Engeström's framework of the activity system (see Chapter 3). NVivo Starter was used for thematic analysis. Before conducting the interviews with the staff, a pilot interview was carried out with a teacher of English language in order to test the application of Engeström's framework for the interview structure and time limit of an hour to discuss all actors of various activities. It turned out that one hour was sufficient time to discuss all actors of only one activity with some side examples.

The first choice to address users was to snowball the online questionnaire to anyone who had an experience with either organisation. The pilot included reaching out to 33 people who were believed to have relevant contacts. Within a month, 13 uncompleted and one completed form with general one-sentence answers were returned. Therefore the approach was considered not suitable for the qualitative inquiry and it was replaced by semi-structured interviews at the respective organisations (see the themes in Annex 2).

The sample of users was a random selection of people on site at the organisations. Non-taggers were considered equally relevant because they could express their perception of the usefulness of tagging and their possible activity without being affected by the institutional affordances and rules that were already existent but not acknowledged by the interviewees.

The Archives were not in contact with their taggers, but the Library staff knew well some of their most active contributors and provided information to contact them. As a result, one pre-arranged in-depth interview was additionally carried out with a tagger of the BL Flickr collection. The themes were similar to other users, but more detailed examples were discussed.

Similarly to the staff interviews, all user interviews were thematically analysed according to the Engeström's framework of activity system.

4.3. Conclusion of the Chapter

Feynman (1974) claimed that, if you put your theory out, you must also put down all the facts that disagree with it. Theory formation was not the aim of the research, but the generalisation done by clustering the interview findings informed by activity

systems according to Engeström is expected to illustrate trends together with the diversity of relevant nuances, supported by a comprehensive appendix (Annex 1) of the collection of quotes from the staff interviews.

The combination of methods aims to present solid evidence on the subject matter, but is acknowledged to be non-representative. In line with King et al. (1994) and Feynman (1974), the approach taken in this research is also expected to be applicable to other kinds of research and fit phenomenon other than social tagging, which gave the idea for the research in the first place.

The next chapter launches the empirical study with the document and user data analysis.

5. The Practice of Tagging

This chapter is first of the two chapters to present empirical analysis. The chapter aims to describe users' tagging behaviour in catalogues and in Flickr, as well as the technological affordances available for them.

In order to address the statement of the State of the Art, different types of organisations are included in the analysis and all cases represent social tagging as a natural, on-going linear activity – unlike the crowdsourcing projects. Due to some limits on the available data, this chapter adds an additional case to the case-study organisations – the consortial catalogue of the pan-European eBooks on Demand (EOD) Library Network. Thus the analysis includes six platforms of two institutions and one consortium:

- the main catalogue Discovery of the National Archives (TNA)
- the main catalogue Explore of the British Library (BL)
- the catalogue Archives and Manuscripts of BL
- the Flickr page of TNA
- the Flickr page of BL
- the consortial catalogue EOD Search of the EOD Library Network.

The research questions for this chapter are as follows: what is facilitated for the users in those platforms; what characterises the user behaviour under those conditions; and what is affected by the type of the organisation or the type of the platform? Content analysis of the tags contributes to answering the main research question about the relationship between users' participation and discoverability of the digital collections. The principal ability of the tags to contribute to discoverability is assessed, but it is not intended to evaluate the use of social tags by wider user community, as this would require access to different data on information retrievals.

Section 5.1, Institutional Affordances presents document analysis of the platforms and related help pages in order to describe the conditions under which the user behaviour takes place. Section 5.2. offers an overview of the statistical query on the

users' tagging behaviour. In Section 5.3, the findings are discussed and implications are presented for organisational practice.

5.1. Institutional Affordances

First, the six platforms were described referring to the interfaces and help articles alongside them on the websites or linked pages. Some information was received directly from the institutions with delivery of data or by special enquiry.

The document analysis looked at 14 parameters: type of the platform (catalogue, social network site), the collection available for tagging (records, textual or non-textual items), online access to the items (full, restricted, partial or no access), collection size (number of items), pre-existing metadata, existence of application programming interface (API), releasing collections by small sets for tagging, time of launching social tagging, authorization of taggers (procedures of registration and sign in), publishing of social tags (immediate, verified), representation of tags to view or browse, procedure for deletion of tags, instructions to tag, and syntax (separators of tags).

5.1.1. Catalogues

The BL main catalogue Explore[81] searches around 70 million items (records for books, journals, newspapers, maps, articles, Sound Archive items, Web Archive links, etc.), being the biggest dataset in the comparison of the six platforms. The TNA main catalogue Discovery[82] holds over 32 million descriptions of records held by TNA (available for tagging) and more than 2,500 archives across the UK. The BL Archives and Manuscripts[83] catalogue includes unpublished documents, prints, drawings etc., and the number of records in the catalogue is unknown. The EOD Search[84] is a multi-lingual consortial catalogue, which runs on the open source platform VuFind and searches over 7 million records of public domain literature from 35 libraries in Europe. The records link to institutional repositories for free full-text or display a button to request digitization for a fee (Mets et al. 2014). Other catalogues in this

[81] http://explore.bl.uk
[82] http://discovery.nationalarchives.gov.uk
[83] http://searcharchives.bl.uk
[84] https://search.books2ebooks.eu

comparison provide mostly restricted access to view items. All are traditional catalogues with pre-existing metadata. No APIs are available for users.

Social tags were enabled first in Explore in November 2008, followed by EOD Search in the beginning of 2011, Archives and Manuscripts in January 2012, and Discovery in October 2012. Yet TNA may also be called a pioneer in this comparison due to launching a wiki site Your Archives[85] in April 2007. A button was placed on the Document Details page of the catalogue taking users to Your Archives to see if there was any additional information; otherwise, it created a special page inviting the user to add content (Grannum 2011). By 2012 the functionality was developed for Discovery, social tags were imported and the wiki was closed.

In the BL catalogues tagging requires signing in, which is only available to registered readers and registered document supply customers. Registration can be completed in person at the Library (The British Library. Get a Reader Pass) and registered document supply customers are the frequent users of the service, purchasing over 100 documents a year (The British Library. Document Supply Services). In Discovery, anyone can register online as providing a Reader's ticket number is optional. The EOD Search also enables anyone to register online.

Instructions about tagging are given briefly from each record's page in the BL catalogues. More detailed information is available from the opening page behind two clicks under 'Help articles.' The same information can be found behind the tab 'Tags,' which is visible throughout navigation. A comma is required to separate multiple tags. In Discovery, each record's page offers a link to sign in to add a tag, but detailed information about tagging is only available from the opening page. Another link after that page gives short tips about useful and appropriate tags, including instructions: "Simply enter a tag and click 'submit'. You can add as many tags as you like" (The National Archives. How to tag records). In the EOD Search there is a note on a field in the record: "No Tags. Be the first to tag this record!" The only instruction in EOD Search for tagging appears after clicking the 'Add Tag' button: "Spaces will separate tags. Use quotes for multi-word tags."

In all cases, social tags are published immediately without verification, mostly next to the record and in EOD Search in a field within the record. Tags can be deleted by the users who attributed them and by the institutions. In the BL catalogues, all tags

[85] http://yourarchives.nationalarchives.gov.uk

can be browsed by 'most recent' and 'most given.' Logged-in users can select to view only their own tags. In addition to tags, BL enables users to add notes, which are not indexed or searchable but are moderated (The British Library 2014a, 2014b). TNA enables users to flag inaccurate tags, which are then checked by the staff. A spam and profanity filter is also in use. All tags can be browsed alphabetically, by 'most given' and 'most recent.'

5.1.2. Flickr

Both the BL and TNA use Flickr to show their selected collections of images. The BL Flickr account[86] was established in August 2007 for corporate promotion. In December 2013, the BL Labs project added over 1 million undescribed images cropped from 65,000 volumes of digitized works from the 17th to 19th centuries (O'Steen 2016). The experiment was meant for anyone to use, remix, and repurpose and to spread new ways to navigate and display the content, and to stimulate the research concerning the materials (O'Steen 2013). First offered to Wikimedia but rejected because of the lack of metadata, Flickr was chosen next because of the tagging option, API existence, and attributing a unique URL to every image. The BL imported the metadata of the books, where the images came from to Flickr, but there was no metadata about the images. Additionally, geotags were imported for maps from the BL crowdsourcing platform Georeferencer. TNA joined Flickr[87] in October 2008 and started to present their thematic image collections since the beginning of 2011 "to give a flavour of their massive holdings" (Annex 1).

Anyone can sign up as a Flickr user and tag the images. Tags are displayed alongside the images as is the link 'Add tags.' The tags added by Flickr robots are visible together with community tags, but distinguished by their white background. Next to 'Tags' under '?' is a short description about tags and a link to some more information, including an instruction to "Separate single word tags with spaces and add phrases in quotes." Users can remove both tags they create and ones Flickr has added for them (Flickr a). The Flickr API enables anyone to write a program to present public Flickr data (photos, video, tags, profiles, or groups) in different ways and make their applications available to other users (Flickr b).

[86] http://discovery.nationalarchives.gov.uk/tags/index/howtotag
[87] http://www.flickr.com/photos/nationalarchives

5.2. User Tagging Behaviour

The acquired parameters for the user data were as follows: tags, user IDs (anonymous for catalogues), item IDs, and time of tag attribution (if recorded). The datasets with social tagging information in the catalogues of the BL, TNA, and the EOD were composed by the respective institutions on request and delivered as separate CSV files. The dataset for the BL Flickr account was composed earlier by the institution, and delivered as a TSV file. The data for TNA Flickr account was extracted by using the Flickr API. The data were imported to R (R Core Team 2016) for analysis.

All in all, 28 parameters were calculated, including total and unique tags and tagged items per person, total and unique tags and contributing users per item, users and items per tag, returns and tagging activity per person across catalogues (for BL), number of tags and people per books for BL, correlations between parameters, and themes of tags. Calendar Converter[88] was used for calculating periods of returns.

5.2.1. Overview

According to availability of the data, the period of observation varies from 28 to 99 months:

- 28 months (Dec 2013–Mar 2016) for the BL Flickr page
- 52 months (Oct 2012–Jan 2017) for Discovery
- 60,5 months (Jan 2012–Feb 2017) for Archives and Manuscripts
- 74 months (Jan 2011–Feb 2017) for EOD Search
- 75 months (Jan 2011– March 2017) for TNA Flickr page
- 99 months (Nov 2008–Feb 2017) for Explore.

Total numbers are presented for social tags on Figure 5.1. and average numbers of social tags per month are presented on Figure 5.2.

[88] https://www.timeanddate.com/date/duration.html

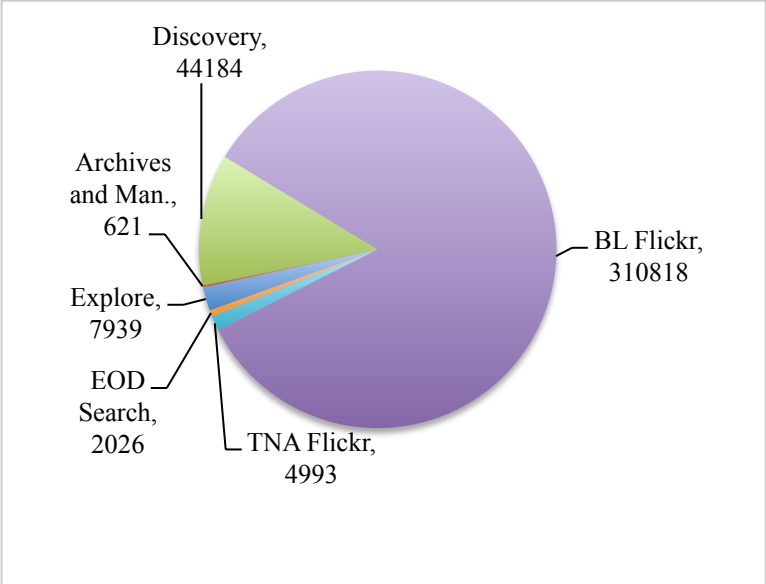

Figure 5.1. Total number of social tags.

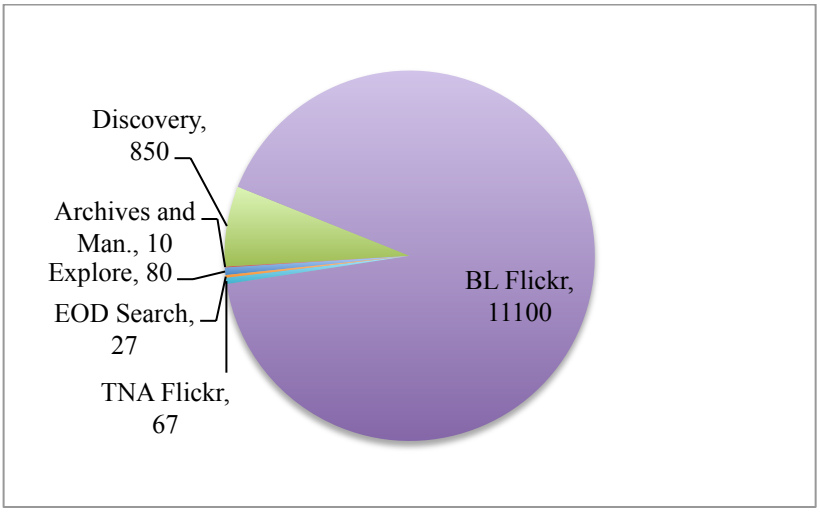

Figure 5.2. Number of social tags in average per month.

The figures exclude tags, which were incorporated from other platforms and ingested by a single user account: 15% of total tags (n=51,990) in Discovery (i.e. mostly tags attributed by users to Your Archives, then imported to Discovery by a single institutional user account), 43% of total tags (n=540,446) in the BL Flickr page (i.e. mostly tags attributed by users in Georeferencer, then imported to Flickr by a single account); 96% of total tags (n=141,603) in TNA Flickr page (i.e. collection names etc. attributed as tags by TNA). 100% of total tags is believed to be social tags in Explore, Archives and Manuscripts and EOD Search. The figures reveal that there is no significant distinction in the proportions between total and average numbers of social tags; therefore, the total number of engaged users is presented on Figure 5.3.

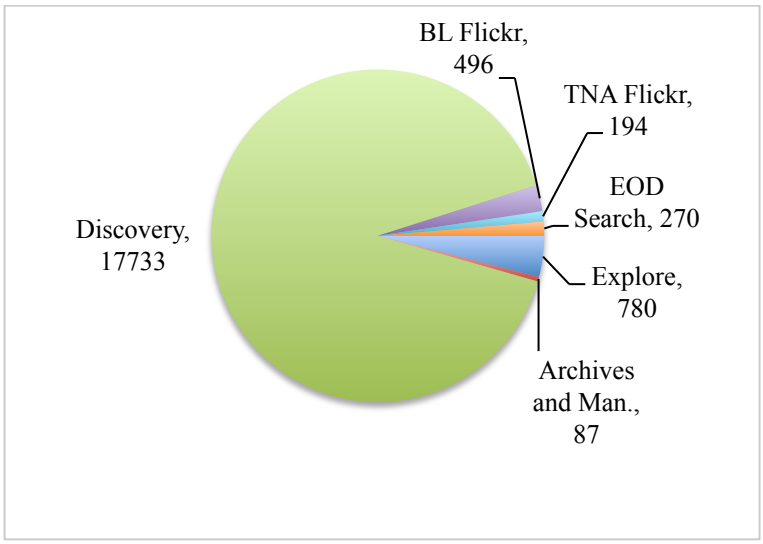

Figure 5.3. Total number of engaged taggers.

The figures expose the drastic mismatch between the number of users and the number of social tags. The content analysis and discussion of the chapter will provide an explanation.

Due to the vast amount of tags in the BL Flickr page, a query was in run Gephi to see the overview of tag division by people and the patterns of tag attribution (Figure 5.4). Nodes represent taggers. The size of a node quantitatively points to the activity rate of a user – the bigger the node the more active the user. The distance of the node illustrates qualitatively the different behaviour. All tags were included for 145 most active user accounts. The nodes on Figure 5.4. point out that 8 to 10 people form the group of top taggers, whereas the most distant active user is doing something really different from others – that would be the account set up by a volunteer to tag maps, to incorporate the georeferencing data, and to mark workflow and some technical parameters in collaboration with the Library.

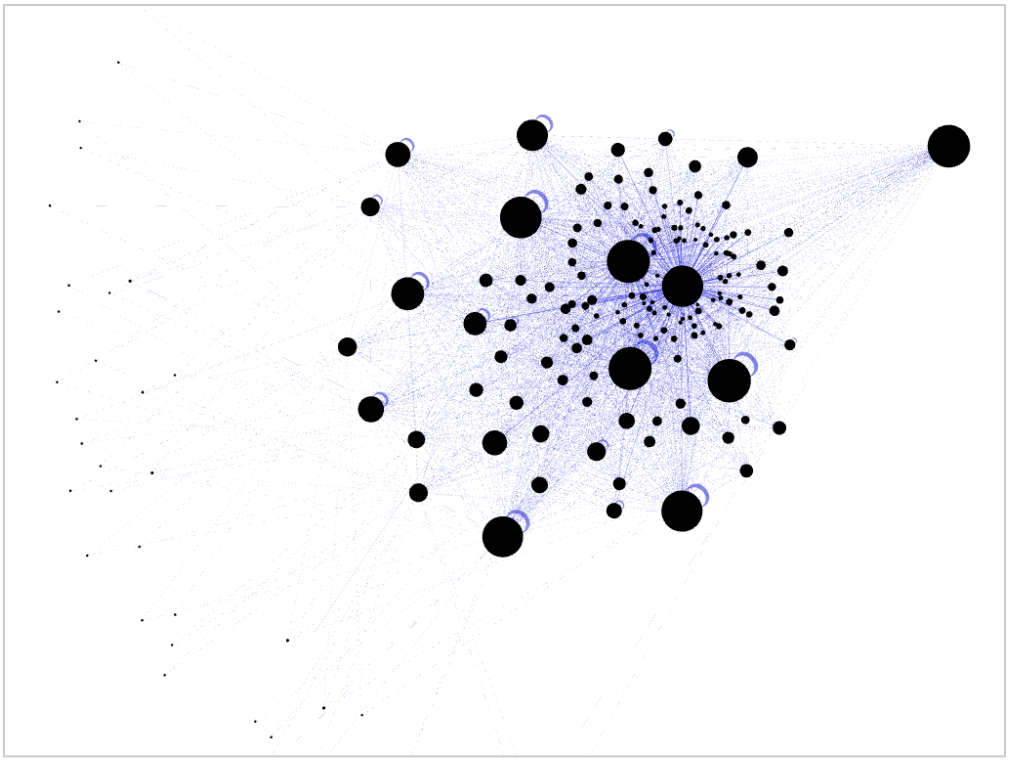

Figure 5.4. Visualisation of taggers in the British Library's Flickr page by tag attribution.

It can also be claimed for other platforms that on average top taggers are up to 8 people, either taking into account the attribution of total tags or unique tags or the number of tagged items (Fig. 5.5).

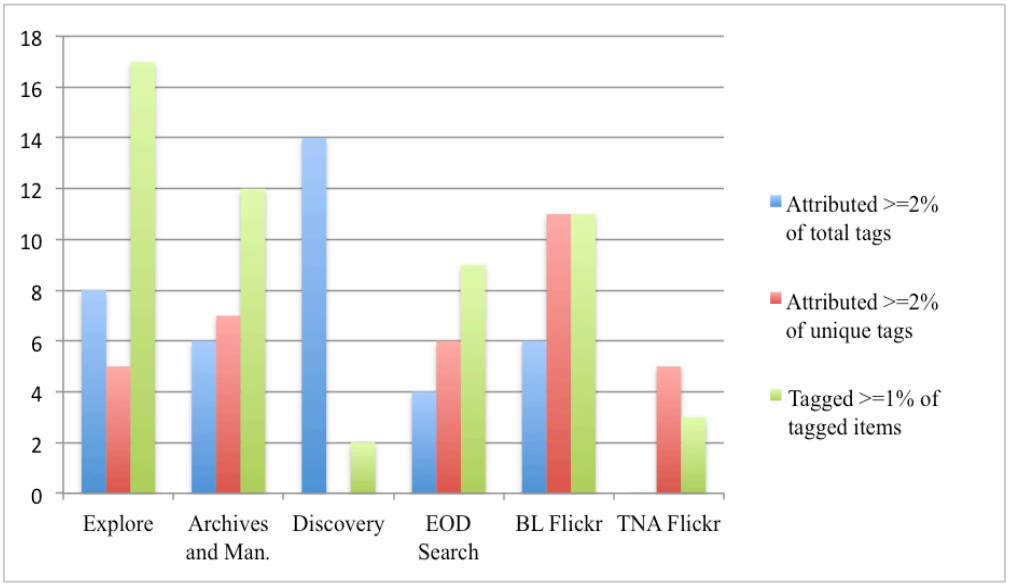

Figure 5.5. Top taggers by total and unique tags and tagged items (absolute numbers).

In total, the majority of users attribute up to 10 tags (Fig. 5.6). If we exclude the tags by institutional accounts, the median value is 6 tags per person for the BL Flickr page, 3 for TNA Flickr, 2 tags per person for Explore and EOD Search, and 1 for others.

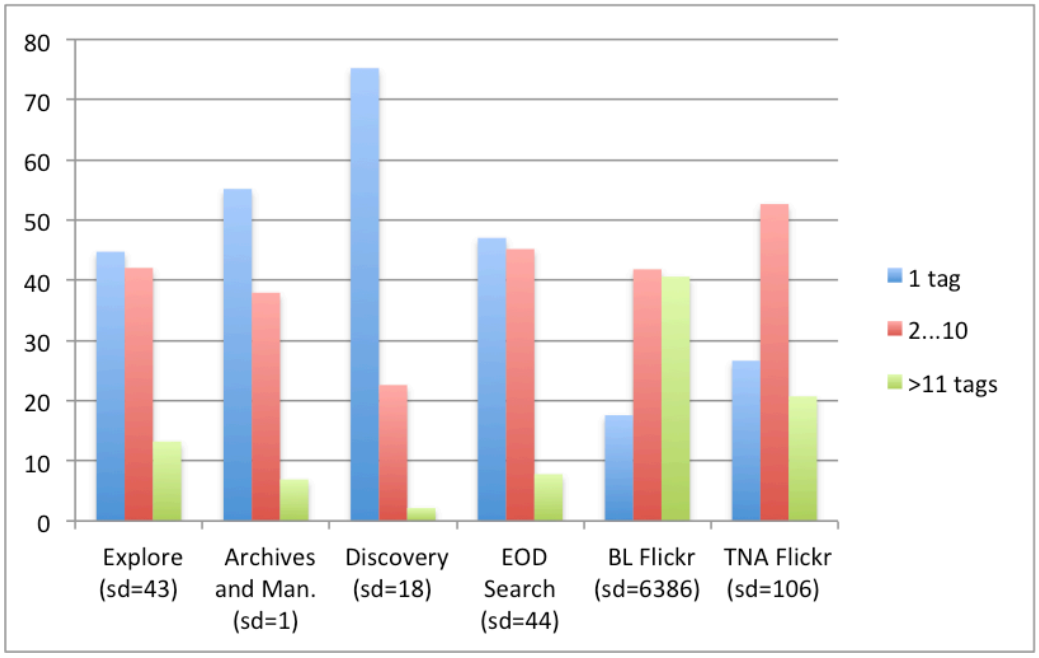

Figure 5.6. Number of tags per person (percentage of total taggers).

Time of tag attribution was analysed only for Explore, Archives and Manuscripts, and EOD Search, because in other cases the time parameter of tag attribution was not available. In Explore there were 956.5 total tags per year on average (sd = 529), in EOD Search 305 tags (sd = 315), and in Archives and Manuscripts 123 tags on average per year (sd = 152). The number of engaged users per year is illustrated on Figure 5.7.

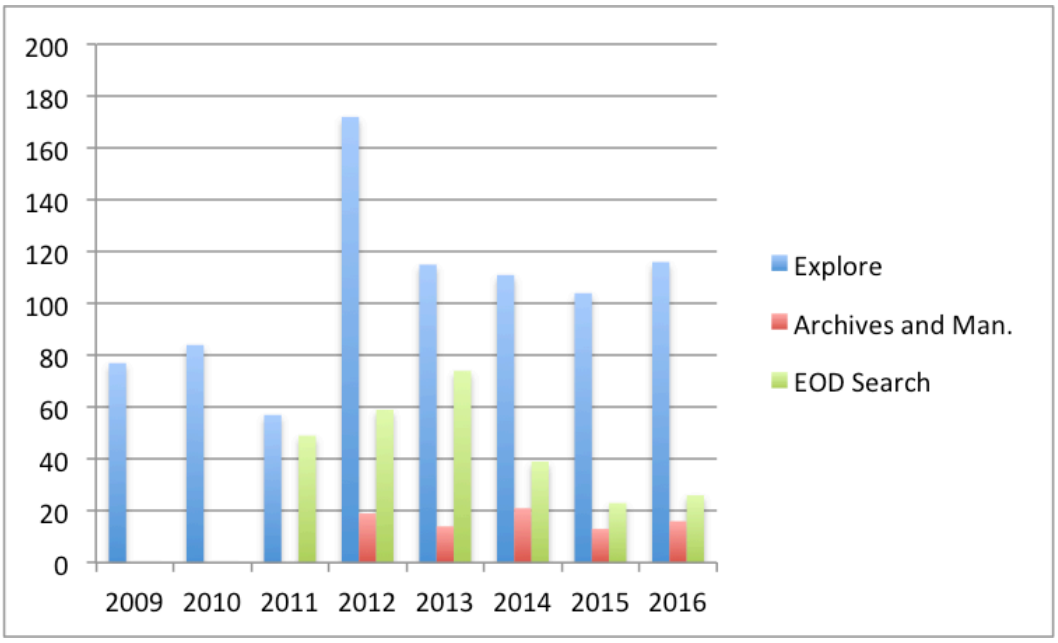

Figure 5.7. Distribution of engaged taggers by years.

In all three cases, users tagged mostly on one day: on average per person on 1.15 dates 1.58 (maximum in EOD Search: 112 dates per user, sd = 6,79). The users who returned to tag on a different date are 165 people (21% of total taggers) in Explore; 13 people (4.81% of total taggers) in EOD Search, and only 11 people (1.4% of taggers) in Archives and Manuscripts. Figure 5.8. illustrates the time interval for returns on average per person.

The period from the first to last day of tag attribution varies as follows: 260 days on average per user (sd = 373, max. 3.7 years) in Explore; 252 days (sd = 489, max. 4.5 years) in EOD Search; 24 days (sd = 23, max. 2.3 months). The correlation between the number of days from first to last day and days with tagging activity is stronger in Archives and Man. (r = 0.68) and Explore (r = 0.54), and weak in EOD Search (r = 0.23).

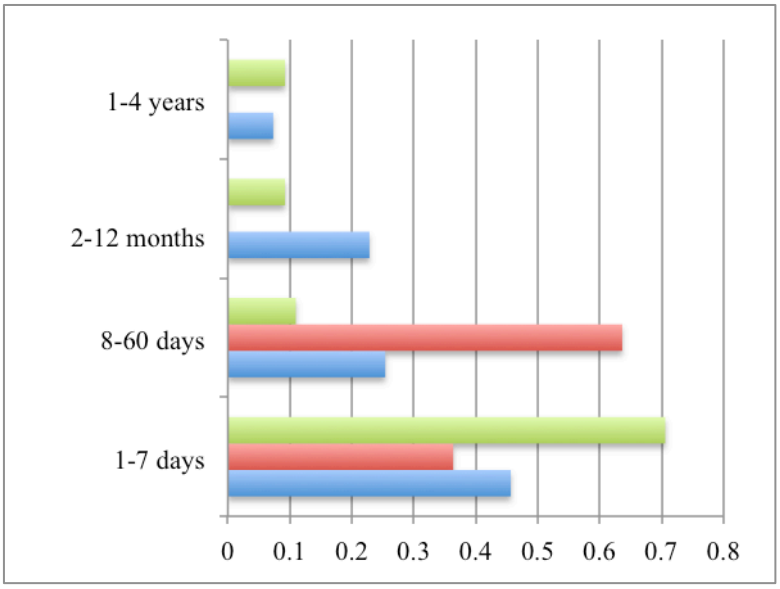

Figure 5.8. Time interval for returns to tag.

Correlation between the number of returns and attributed tags is modest (Explore r = 0.46, EOD Search r = 0.36, Archives and Manuscripts r = 0.25). For instance, the 18 top taggers (by total tags) in Explore divide into 3 people who returned on >10 dates, 12 people on 2 to 9 dates, and 3 top taggers who gave all their tags on one day. In Archives and Manuscripts, one person added tags on 4 dates within 2 months. All other 10 of 11 users tagged on 2 dates and returned within the same or subsequent month, including the top contributor (attributed 55% of tags) on two subsequent dates.

Data for Explore and Archives and Manuscripts, as well as data for Flickr pages, allows looking at tagging across catalogues. 16 people have attributed tags both to Explore and Archives and Manuscripts, i.e. 2% of taggers in Explore and 18% in Archives and Manuscripts. 2 people appear as top taggers in both catalogues. 14 people tagged images in both the BL and TNA Flickr pages, which is 2.8% of taggers in the BL Flickr page and 7.2% of TNA Flickr taggers.

The BL Flickr page is first by the sum of unique tags, but both Flickr pages have least unique tags out of total tags (TNA 5% and the BL Flickr page 8%, Explore 26%, Archives and Manuscripts 28%, EOD Search 46%, Discovery 60%).

The correlation analysis in Figure 5.9 illustrates whether a) the people, who added more tags in total also added more unique tags; b) those who added more unique tags tagged also more items in total; and c) those who added more tags in total, tagged also more different items in total. The latter is true for all cases. But the first two questions

have rather distinct answers: attribution of unique tags per person is valid for archival content.

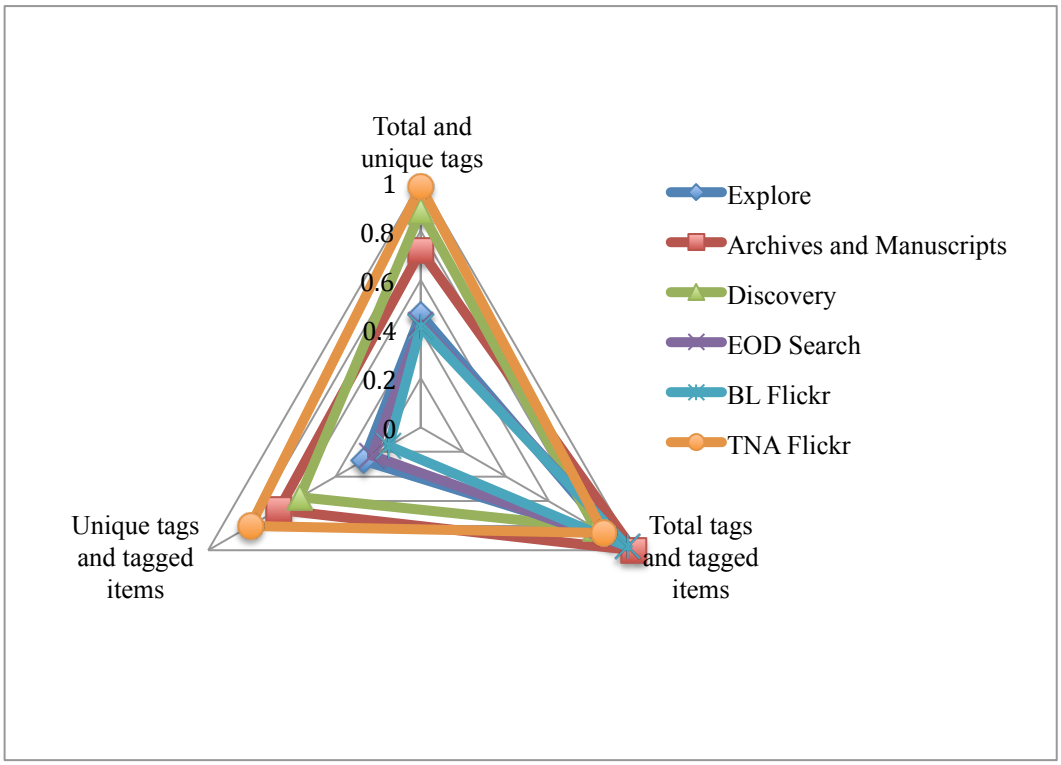

Figure 5.9. Correlations between users' attribution of total and unique tags and number of tagged items.

5.2.2. Tags

Looking at the content of tags, first, the most frequent tags were found out and second, the tags attributed by the most people. Insight into the Library's main catalogue illustrates that the most frequent tags in total and tags attributed by most people differ significantly (Table 5.10). The latter includes also tags for personal use which are meaningless to other people, e.g. 'to read', 'check.' Likewise 'dissertation' or 'thesis' may refer to the need to check the item for the user's research project, rather than the item itself being a thesis.

	Tags attributed most in total	**Tags attributed by most people**
Explore, British Library	>=1% of total tags	>=1% of total users
	silent cinema	dissertation
	business directory	>=0,51% total users <=0,77%
	war of 1812	fantasy fiction
	telecommunications industry guide	history
	This	read
	anarchist newspapers & periodicals (english language)	to read
	propaganda 2013 exhibition	biography

	Construction	check
	renewable energy industry guide	music
		1
		business, directory
		darwin
		literature
		psychology
		thesis

Table 5.10. Top tags in Explore.

The catalogue Archives and Manuscripts has only one tag, which is attributed by at least two people (Table 5.11). Furthermore, the name 'john' does not refer to the same person, because other tags attributed by the taggers to the same record did not match.

Archives and Manuscripts, British Library	>=1% of total tags	All tags attributed by at least 2 people
	sialkot	john
	gujranwala - 1	
	sargodha	
	sialkot - book	
	jhang	
	lyallpur	
	gujrat	
	faislabad	
	sahiwal-1	
	sialkot - 1	
	trade unions	
	west punjab	
	auditor notes	
	jenny	
	miscpoems	
	sialkot - map	
	war of 1812	
	asaf	
	brits on buddhist before 1922	
	manila	
	sahiwal	

Table 5.11. Top tags in Archives and Manuscripts.

Most frequent tags in Discovery refer to the import of categorical tags from the previously run wiki Your Archives, e.g. ‘a level - korean war,’ while tags attributed

by most people refer to the main interest of the users of the Archives – family history (Table 5.12).

Discovery, National Archives	>=0,46% total tags <=-0,93%	>=1% of total users
	a level - korean war	grandad
	a level - chartism	>=0,24% total users <=0,92%
	movcon	grandfather
	cycle	dad
	haiti list	father
		ww1
		great uncle

Table 5.12. Top tags in Discovery.

The findings for EOD Search refer to the multilingual use of the catalogue as well as not following syntactic rules while trying to write a phrase as a tag (Table 5.13).

EOD Search	>=1% of total tags	>=1% of total users
	incunable	1
	1	2
	CS	of
	>=0,4% of total tags	12
	officium	Geschichte
	http://search.books2ebook	a
	2	des
	Band	in
	Ainsworth	the
	Hubertusburg	von
	alphabets	
	of	

Table 5.13. Top tags in EOD Search.

Top tags in the Flickr page of the Library refer to georeferencing and to mark-up for workflow, e.g. the tag 'rotate' refers to the need to rotate the image in Flickr and 'synopticindex' refers to linking to Wikipedia (Table 5.14). If we exclude geotags by the BL in Flickr, the list of top tags remains similar. And if we manipulate that dataset further by losing the computational parts of tag strings, more geographical locations appear as the most attributed tags in total.

	Tags attributed most in total	Tags attributed by most people
Flickr, British Library	>=1% of total tags	>=4,23% of total users
	map	map
	georefphase2	portrait
	togeoref	church
	rotate90	bird
	hasgeoref	london
	wp:bookspage=geography	castle
	geo:continent=europe	river
	coatofarms	dog
	portrait	boat
	colorful	horse
	colourful	ship
	people	woman
	lettert	cathedral
	>=0,43% total tags <=0,87	bridge
	wp:bookspage=synopticindexusa	letter
	rotate270	man
	geo:country=uk	egypt
	geo:country=unitedkingdom	split
	music	geology
	fauna	child
	lettera	children
	letteri	fish
	wp:bookspage=synopticindexukandireland	scotland
	georefphase1	world
	wp:bookspage=synopticindexfrance	africa
	geo:state=england	birds
	wp:bookspage=anthropologyandethnology	fashion
	lefthalf	flowers
	righthalf	france
	letterw	japan
	geo:continent=northamerica	landscape
	initial	lion
	wp:bookspage=other	*etc.*

Table 5.14. Top tags in Flickr page of the Library.

Similarly to the Library, the most frequent tags in case of the Archives' Flickr page refer to computational tags, which in this case were added by the Archives themselves. That is why an extra column is added in the middle with the most added tags by users in total in Table 5.15.

Flickr, TNA	Tags attributed most in total, incl. TNA as a tagger	Tags attributed most in total, excl. TNA as a tagger	Tags attributed by most people
	>=1% of total tags	>0,02% of total tags	>3% users
	thenationalarchivesuk	kuching	woman
	tna:departmentreference=co	kenya	horse
	tna:seriesreference=co1069	anglofrenchboundarycommission	railway
	tna:divisionreference=cod32	nigeria	road
	africathroughalens	guinea	school
	asiathroughalens	sierraleone	street
	asia	astana	boat
	tna:subseriesreference=co1069ss1	vintagekenya	bridge
	tna:subseriesreference=co1069ss3	wwii	children
	tna:subseriesreference=co1069ss4	aviation	horses
	australasia	penang	victoria
	australasiathroughalens	river	beach
	oceania	hackman	car
	malaysia	kualalumpur	cathedral
	australia		church
	tna:subseriesreference=co1069ss2		falls
	caribbeanthroughalens		girl
	etc.		*etc.*

Table 5.15. Top tags in Flickr page of the Archives.

Next, thematic clusters of the tags were formed in order to find out the shares of the topics and assess the potential impact of the tags to the discoverability of the items. The topics of the clusters were informed by the previously presented lists of most frequent tags and tags attributed by most people; therefore, the clusters cannot be considered as a precise representation of themes, but as an indicative minimum division of tags. For this analysis, we focus on four of the most representative datasets: the main catalogues of the Library and the Archives and their Flickr pages.

Themes in Explore. Four groups were composed according to top tags or for comparison with other clusters. Figure 5.16 illustrates the diversity of tags, which were not captured by the four groups.

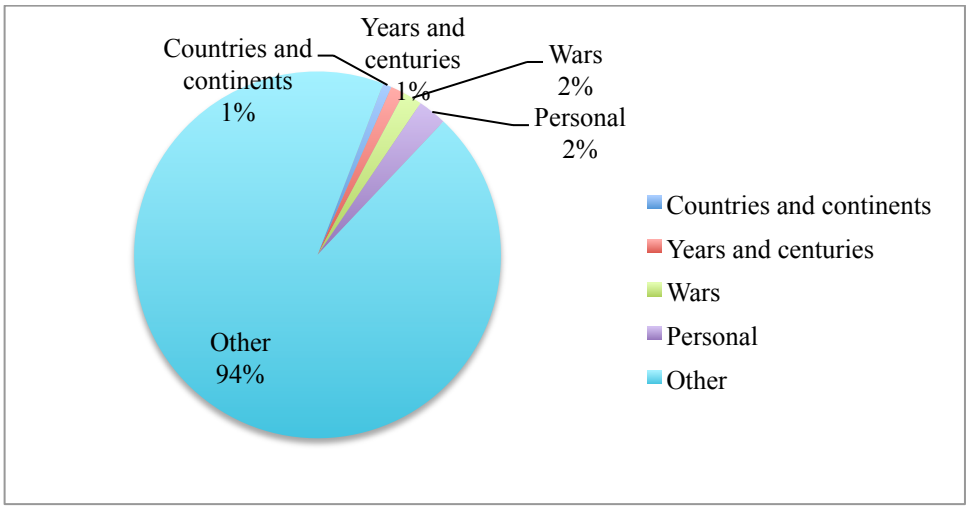

Figure 5.16. Themes in Explore.

1) Countries and continents. Countries and continents were added as tags in a few cases, 61 times in total, i.e. 0.8% of total tags in Explore. Tags belonging to this group and added at least 3 times refer to some distinct locations to UK:

europe 10
italy 9
iraq 7
afghanistan 4
australia 3
china 3
japan 3

2) Years and centuries. To compose this group all numbers between 1400 and 2000 were defined as years and all tags including the word 'century' were added. As a result, 95 tags, i.e. 1.2% of tags in Explore appeared in this cluster, and those attributed at least 4 times (followed by tags attributed 2 times) were as follows:

19th century photography 50
1812 20
18-19th century 4
dance history: 16th century 4

3) Wars. The following tags were included in this group amongst the tags, which were attributed by at least 2 people: war of 1812, world war 2, world war 1, wwi, wwii, ww1, ww2. As a result 135 tags, i.e. 1.7% of tags in Explore appeared in this group with the following total representation:

war of 1812 128
world war 2 2
ww2 2

war 1
ww1 1
wwii 1

In addition, the wars tagged in the BL Flickr page were looked up in this sample, but there were no matching occurrences.

4) Personal notes. The following tags were picked from the top tags for this cluster: to read, read, to book, check, this, read this, reading list, to order, to be ordered, and relevant. As a result this cluster is composed of tags attributed 202 times, i.e. 2.5% of total tags in Explore, with the following distribution:

this 108
to book 45
read 14
to read 12
check 8
to be ordered 6
to order 5
read this 2
relevant 2

Themes in Discovery. The following 5 groups were composed for Discovery (Fig. 5.17):

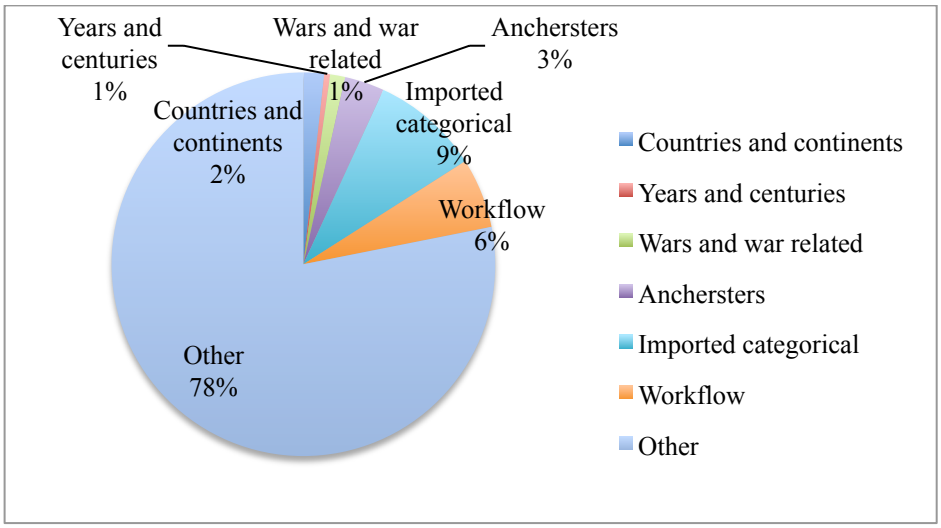

Figure 5.17. Themes in Discovery.

1) Countries and continents. Tags related to the list of countries and continents are attributed 907 times, i.e. 1.7% of total tags in Discovery. The top ten occurrences were as follows:

kenya 139
cyprus 99
turkey 89
panama 44
singapore 43
mozambique 35
poland 32
barbados 29
jersey 20
palestine 20

2) Years and centuries. Tags occurring in this group under the same conditions as for Explore are attributed 256 times, i.e. 0.5% of total tags in Discovery. The top 10 occurrences are as follows:

1956 22
1958 18
20th century 15
1936 13
1895 12
1955 11
1957 10
1948 7
19th century prison ships 7
1946 6

3) Wars and war related tag. First, this group was composed by tags attributed by at least 2 people in Discovery and similar forms for marking World War: war, african-caribbean first world war, american war of 1812, wwi, wwii, ww1, and ww2. These tags were attributed 104 times, i.e. 0.2% of total tags, in Discovery. Secondly, the most frequent words related to wars were added to this group, because they stood out in this case compared to other platforms. The added tags were as follows: regiment, squadron, brigade, battery, coy, hussars. The merged list results in 657 total tags, i.e. 1,3% of total tags, with the most frequently occurring as follows:

regiment 308
squadron 121
ww1 58

4) Anchersters. There were 1736 occurrences of the following tags: granddad, grandad, grandfather, granpa, granny, grandmother, brother, sister, great uncle and uncle, great aunt and aunt, i.e. 3,3% of total tags.

5) Imported categorical tags. This group consisted of meaningful tags incorporated by the Archives and mostly based on the tags attributed to previously run wiki Your Archives. The common feature of the tags was the beginning 'a level-, referring to the high level of the description. Those tags were attributed 4,722 times, i.e. 9.1% of total tags. The most frequent 10 tags in this group are as follows:

a level - korean war 485
a level - chartism 341
a level - records on the spanish civil war, 1936-1939 228
a level - special operations executive 224
a level - french revolution 162
a level - ireland 1916 - 1922 127
a level - suffragettes 117
a level - general strike 1926 97
a level - ufos 69
a level - unidentified flying objects 69

6) TNA workflow. This group is not meaningful for end-users and consists of tags beginning mostly with 'b) comprehensive, though incomplete, data sheets for serials...' and 'a) microfilm copies of...'. Altogether these tags form 3,069, i.e. 5.9% of total tags.

Themes in Library's Flickr page. For the Flickr page of the Library, five clusters were formed, which constitute a significant part of total tags (Fig. 5.18)

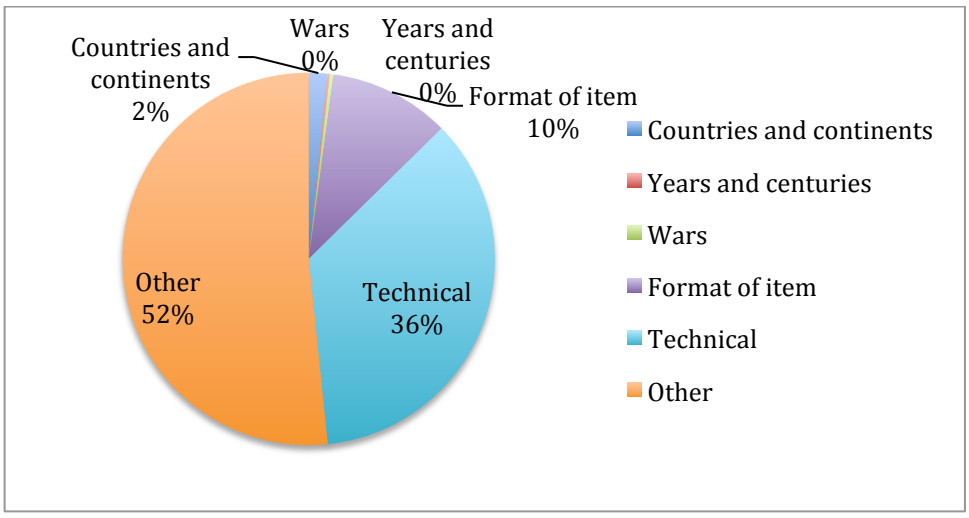

Figure 5.18. Themes in the Flickr page of the British Library.

1) Countries and continents. Pure names of countries or continents were marked as tags 3,627 times. If we include also those country and continent names in the

analysis, which were part of a computational tag strings like 'geo:...' and 'synopticincex,' the result is 9,156 tags, i.e. 1.7% of total tags.

The 10 most frequently occurring tags in this group are as follows:

france 1647
germany 647
italy 559
spain 369
india 324
canada 304
ireland 281
egypt 262
australia 244
switzerland 241

2) Years and centuries. As a result, 723 tags were defined as years and in total with centuries 1,165 tags referred to time periods, i.e. 0.22% of total tags.

The 10 most occurring tags in this group are as follows:

19thcentury 220
1715 152
18thcentury 43
16thcentury 33
1880 25
15thcentury 23
17thcentury 20
13thcentury 19
12thcentury 16
14thcentury 15

3) Wars. This group was composed based on mentions of wars within the 50 most frequent tags. The wars mentioned as tags within this selection were attributed at least 4 times and a maximum of 937 times. In total, 1,898 tags were attached to this group, i.e. 0.35% of total tags. The 10 most occurring tags in this group are as follows:

uscivilwar 937
frfrancoprussianwar 394
frfrenchrevolutionandnapoleonicwars 114
americancivilwar 94
usspanishamericanwar 67
war 39
americanrevolutionarywar 27
civilwar 25
napoleonicwars 24

frenchrevolutionarywars 21

4) Format. This group refers to the tags that mention the format of the item. For this, the 200 most frequent tags were analysed and similar words were added. The tag 'portrait' was excluded from the group, because it referred both to the portraits of people as well as the orientation of any image. As a result 59,772 tags referred to the format of the image, i.e. 11% of total tags. The group included 12 tags in total with the following frequencies of occurrence:

map 53,806
decoration 1,984
typography 939
diagram 823
engraving 652
foliage 589
illustration 427
photograph 312
plan 171
photo 31
drawing 19
painting 19

5) Technical tags. As so many technical terms appeared as the most frequent tags, a separate group was composed based on the relevant findings among the top 200 tags. Any tags beginning with 'geo:osmscale=' and 'rotate' were also treated as such. The findings relevant for the group of technical tags were given 53,806 maximum and 218 times minimum. As a result 203,686 tags appeared in this group, i.e. 38% of all tags. The 10 most occurring tags in this group are as follows:

georefphase2 50,211
togeoref 43,961
rotate 29,641
hasgeoref 14,566
geo:osmscale= 14,284
georefphase1 2,990
lefthalf 2,441
righthalf 2,441
wp:bookspage=other 2,328
nogeoref 1960

Themes in the Archives' Flickr page. For the Archives' Flickr page the following 6 groups were composed (Fig. 5.19).

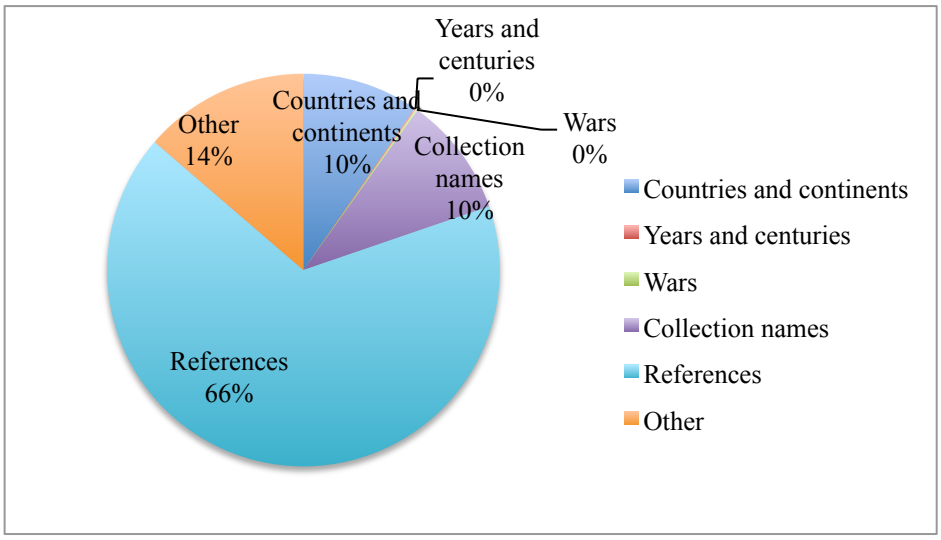

Figure 5.19. Themes in the National Archives' Flickr page.

1) Countries and continents. 13,587 tags referred to countries or continents, i.e. 9.6% of total tags. The most frequent ten tags in this group were as follows:

asia 4,144
oceania 2,461
malaysia 2,185
australia 1,847
canada 460
china 456
barbados 300
jamaica 224
fiji 188
belize 132

2) Years and centuries. 175 tags appear under similar conditions as previously set, i.e. 0.12% of total tags. Tags in this groups, which were attributed at least 10 times (followed by 5 times) are as follows:

1963 13
1930 11
1929 10

3) Wars. This group was composed by tags related to wars, which were mentioned at least 2 times (80 times maximum). These tags were attributed 249 times, i.e. 0.18% of total tags. Tags attributed more than 10 times were as follows:

boerwar 80

secondworldwar 64
wwii 42
firstworldwar 26
ww2 13

4) Collection names. This group was composed by collection names attributed by the Archives, which included the phrase '...throughalens'. 13,944 tags were attributed as such, i.e. 9.8% of total tags, divided as follows:

africathroughalens 4737
asiathroughalens 4152
australasiathroughalens 2703
caribbeanthroughalens 1556
mediterraneanthroughalens 737
europethroughalens 59

5) References. This group consists of the tags attributed by the Archives with the beginning 'tna:...' (attributed 78,879 times in total), e.g. 'tna:SubseriesReference=CO1069ss4,' which refers to the collection references and 'thenationalarchives' (attributed 15,386 times), totalling 94,265 tags, i.e. 66.6% of total tags in their Flickr page.

5.2.3. Items

On average per month, the most items were tagged on the BL Flickr page (6,290 items per month, 176,133 in total), fewer in Discovery (794 per month, 38,106 in total), in TNA Flickr page (209 per month, 15,679 in total), in Explore (56 per month, 5,548 in total), in EOD Search (18 per month, 1,360 in total) and in Archives and Manuscripts (9 per month, 528 in total).

The tagged items gained mostly one tag per item in the catalogues (mean, median, mode < 1,5; sd < 2,2), more in Flickr (BL Flickr: mean = 3, sd = 68,5; TNA Flickr: mean = 7,2; median = 8; mode = 11, sd = 4,38). In all cases, the items were mostly tagged by one person (mean < 1,42; median = 1; mode = 1; sd < 1,69).

In Archives and Manuscripts, no records were tagged by at least 2 different people. In Explore, the maximum number of taggers per record was 4. The record was for the book “Charles Dickens” and all four people throughout 3 years had exclusively added tags 'charles dickens', 'charles dickens 1', 'charles dickens test'. On the BL Flickr page, on 18 occasions a maximum of 5 people tagged the same image (13 maps, and 5 other different types of images on Figure 5.20).

Figure 5.20. Images in addition to maps tagged by most people in the British Library's Flickr page. (The British Library. Flickr)

In TNA Flickr page, maximum 8 people tagged the same image with tags like 'Victorian', 'portrait', 'child', 'boy', 'rabbitthief' etc. (Figure 5.21). The photograph was posted together with the story of the 10-year-old boy who was sentenced to hard labour for stealing two rabbits. It brought along many 'faves' and user comments about feeling sympathy for the child or arguing, “That is what discipline is!”

Figure 5.21. “Prisoner 4100”. (The National Archives. Flickr).

Discovery is distinguished by having a maximum of 104 people per record, a high-level record “Poor Rate (Ledger)” 80 people per record “In Jos. Berington's hand, notes on French oath to the Constitution,” etc.

The correlation between tagged items and attribution of total tags is very strong in all cases (r >= 0,82). The correlation between attribution of total and unique tags is strong in the case of archival content (TNA Flickr r = 0,98, Discovery r = 0,88, Archives and Manuscripts r = 0,72), but remains moderate in other cases (r <= 0,46). Similarly, the correlation between tagged items and attribution of unique tags is strong or medium for archival content (0,8 >= r >= 0,57), low for others (r <= 0,27).

The dataset for the Archives' Flickr page also included the number of views of each item. This enabled a correlation analysis to find out whether the most tagged items were also the ones most viewed. The correlation index remained modest (r=0,32), as illustrated on Figure 5.22.

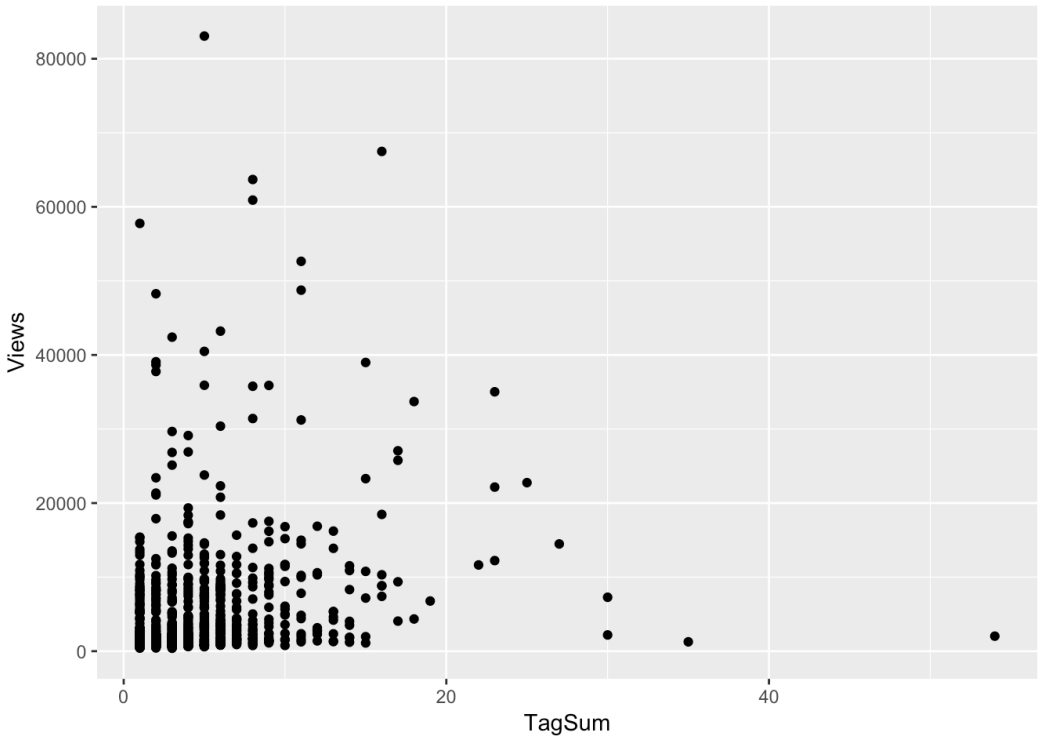

Figure 5.22. Number of tags and views of the images in the National Archives' Flickr page.

Activity per book in the BL Flickr collection. The images in the BL Flickr collection came from books, but are displayed as individual items in Flickr. Thus individuals interact with independent items and their relation to the books is not significant. However, it was noticed that tags attributed to images in the BL Flickr page refer to 18,084 books. It means that 36% of all books which were processed with Mechanical Curator got at least one tag on their images, and 64% of the books with at least one image were tagged. Most tagged books are encyclopaedia-like items, also including many maps.

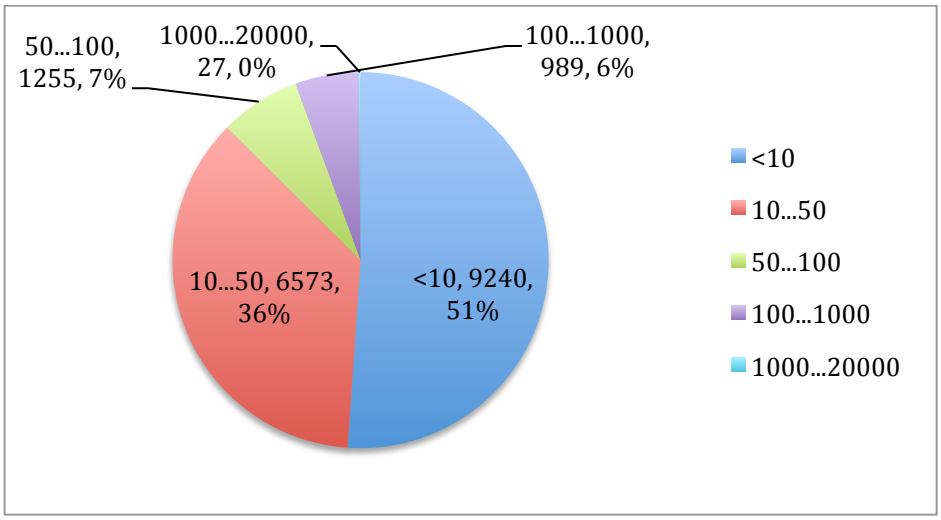

Figure 5.23. Total number of tags per book in the British Library's Flickr page.

5.3. Discussion

The comparison includes two main catalogues of an archive and a library, their Flickr pages, a smaller archival catalogue of the Library and a consortial catalogue. Derived from the availability of the data and the research questions, some datasets were analysed in more detail.

The difference in the size of collections available for tagging does not seem to have a direct impact on the tagging activity. The difference in the type of collections may have increased the tags for the BL Flickr collection of cropped images as it had no pre-existing metadata and users had the freedom to add anything, even if the image discovery may have been more serendipitous compared to described collections.

The two institutions have different user authorization procedures for the catalogues. The low number of taggers in the BL catalogues may be caused by the registration procedure, which requires personal presence or being a frequent customer of document delivery service, compared to Discovery where anyone can register online. However, Flickr offers free online registration, but engaged less users than the catalogues. The devoted subject communities may not be as used to using Flickr as catalogues.

Even though the datasets differ by number of participants and tags, they result in similar numbers of top taggers by attribution of total tags (up to 14 people), unique tags (up to 11 people), and tagged items (2 to 17 people). Analysis of the top images by the number of tags and number of users showed that there were just a few most frequently tagged images. Also taking into account that 64% of the books with images

got tagged, we may conclude that the tagging behaviour is rather random. Given that people tag different images from a variety of sources means that crowdsourcing projects, which release only a particular set of items available for tagging at a time, prevent increasing discoverability of a range of items.

It is common in catalogues that most users add only one tag. Flickr users tend to add more tags per person, and there are more tags per item in Flickr than in catalogues – both trends possibly affected by the use of Flickr API.

It is evident in all cases that people who add more tags also tag more items. Also given that we find only a few taggers per item we may conclude that users tend to describe items briefly and choose different items from each other. The phenomenal 104 users per record in Discovery is an exception that may be explained by the biggest number of taggers in total – the higher the number of users the higher the probability of tagging the same object – or it may refer to users' mistakes in attributing tags to a collection rather than a specific item level. It could have been explained by the collection size – the smaller the size of collection the higher the probability of people tagging the same item – except Discovery exceeds the number of images in Flickr about 30 times.

The common feature for all three platforms with archival content is the attribution of unique tags compared to libraries. It may refer to the perception of users to make distinctions between unique and personal archival content, compared to the published materials which form the collections of the libraries.

Comparison between most attributed tags and tags attributed by most people reveals an interesting dichotomy. In four cases, the most attributed tags refer more often to applying computational techniques for tag attribution, mass import of tags from other sources or to the form of the tagged object, whereas tags attributed by most people tend to be more telling in terms of content: e.g. Discovery suggests that the core interest of most taggers is genealogy. And even if tags like 'grandad' are a noise for others, attribution of such tags tells us to update the instructions with the most common examples. Additionally tags like 'to read' or 'alan's summer project' refer to marking up for individual need, suggesting the development of functionality for private tags. Interestingly, these examples occur in catalogues and not in Flickr, which has more liberal or messy image than the controlled and verified ones of the catalogues. It may also rise from the nature of the collections presented on Flickr – selected images instead of records of materials.

The misuse of syntax for multiple tags or for multi-word tags may turn useful tags into noise: e.g. commas must be used in the BL catalogues, but if users follow the record and add ‘Last Name, First Name’ without quotation marks, it results in having the tag ‘john’ instead of the full name. Similarly not using quotation marks for phrases in EOD Search resulted in having prepositions as tags. Flickr has the same rules, but these mistakes are not common there. Crowdsourcing projects usually avoid this kind of mistakes by having separate text boxes for different descriptive data.

The data of three catalogues, where the time factor was available, point to the rather surprising finding that not only most users, but also most top taggers by total tags, make their contributions within a short time-frame – less than 10 days on average and in some cases only once. The timescale does not imply that tagging has become more popular over time, rather that it is unstable in terms of total tags but more stable in terms of participating users.

Relevance to discoverability. Due to the vast amount of tags, the thematic analysis of tags was done computationally in order to illustrate the general trend of the prevailing themes. It is acknowledged that it is not an accurate presentation of all the terms with similar meaning. Nevertheless, the proportions of the clusters are considered significant and relevant to determine their relation to improving discoverability.

Countries and continents are marked more as tags in Flickr than in catalogues for different reasons. First, the metadata in catalogues usually captures the location as a keyword, but the Library's Flickr collection initially had no metadata – thus there is more reason to add a country name as a tag, especially when a group of active contributors formed around maps and geotagging. This also has a significant impact on the mark up of the format of items. Secondly, on the Archives' Flickr page it became evident that the majority of tags referring to countries were attributed by the Archives' staff – possibly incorporating already existing metadata via Flickr API for discoverability on that particular platform.

The Library's Flickr page may also stand out in terms of the attribution of years and centuries, because of the missing metadata which in other cases is marked up by the staff. Both Flickr pages have used computational tags to mark up items for workflows possibly related to ingestion, analytics, or descriptions. The institutional systems probably enable the use of other features than tagging for these purposes.

It can be argued whether the technical and personal tags improve the discoverability or should be considered as noise. From the point of view of visibility, point of view it can be agreed that those groups of tags *are* noise: when someone takes a look at the list of such tags next to a record, they does not provide any meaningful additional information. To take a position from discoverability's point of view, we should return to the definition of discoverability, which says that it is improving an item's ability to be found by appropriate users. Thus the staff members or volunteers who have added technical tags have improved the discoverability of the items in order to be able to extract these items and incorporate them into other 'appropriate infrastructures' – another condition of the definition of discoverability. Likewise the people who have tagged items with personal keywords like 'granddad' or 'to read' have improved the discoverability of the items *for themselves* so that they can return to these records. Thus the content analysis suggests that tags are attributed for discoverability, but more sophisticated ways to expose the tags according to the special groups of 'appropriate users' could be developed.

Considerations for Institutional Practice. Eventually, 8 considerations based on the document and user data analysis are extracted from the presented analysis for institutional practice:

Enable and instruct

(1) Enabling tagging for everyone upon online registration seems to have a positive impact on the number of contributors compared to ID verified users only. Institutional practices refer to volunteers coming from different countries (Ridge 2014), as do the users of the catalogues who are potential taggers. About 90% of the visits to the EOD search engine (Mets et al. 2014) and 30% of visits to Your Archives were from Google searches (Grannum 2011). The description of the record may become more varied and more reliable if tagged by more people.

(2) When the taggers come, they might not come through the first door, i.e. the opening page of the catalogue, which was in one instance the only place to find instructions on tagging. EOD Search has experienced only 3% of the visits landing on the opening page, and the majority of visits came directly to single records (Mets et al. 2014). Help pages should be cross-linked.

(3) Available and clear instructions proved to be vital, especially for the use of separators. In some cases, the comma was used intuitively instead of the space, quotation marks were not used for phrases, or users did not understand the instruction

to insert one tag at a time. If user behaviour suggests that the requirements for separators must be adjusted, it will likely not affect many users, who were used to different separators and will likely not return to tag in the future as suggested by the data.

Advance and reflect

(4) Some tags, even within top tags, are initiated through individual need. It is not clear if users intentionally break the rule of 'making a useful contribution' or if they have overlooked the notion that all tags are made publicly available. Still, users seem to need an alternative option to add some tags visible only to themselves or groups defined by them, because these tags are not meaningful to other people.

(5) As the digital skills of users improve (use of APIs, running software libraries, image recognition tools, etc.), providing an API justifies the effort and significantly increases the amount of tags. It requires full availability of items, which in turn might reduce the risk of tagging based on assumptions when the item was not fully seen. It may also lead us to an intriguing option to enable social (computational) tagging not in catalogues, but in repositories.

(6) Displaying tags which have been given by most people might be more telling than showing the most attributed tags, no matter how many people or techniques. It is not the case when the amount of contributors is too low.

Monitor and maintain

(7) Recording the time of tag attribution is important for monitoring the returns of the contributors. That data were available for 3 platforms of 6, all suggesting that not only majority, but even top taggers occur in short time frames.

(8) If the goal is to keep the top taggers, their contributions should be detected quickly and, if deemed valuable, the dialogue should be started and maintained quickly before they leave.

5.3.1. Conclusion of the Chapter

The variety of tags, and especially the variety of tagged items, which becomes evident from this study, illustrates the importance of social tagging for the whole collection, not a specific set selected by the staff as is common for crowdsourcing projects. It would contribute to the serendipity of noticing unexpected but relevant findings in the search results and eventually increasing the discoverability of those items.

Overall user activity confirmed the previous studies on the small proportion of active users. The current study also showed that driving tagging activity onto a social network site does not guarantee that the activity goes viral. The power of Flickr in this case lies in its API, which was used for mass tagging, but not in increasingly engaged audiences, which could have been expected due to the social nature of the platform. But it also did not bring along noise or spam as might be expected from social network sites.

The takeaways for organizations suggest reviewing sign up procedures, making instructions clear and available, considering individual need for tagging, developing tools for computational tagging by users, defining "top tags" not only by their sum but also by the number of people attributing them, monitoring the activity in time, and cherishing the valuable contributors.

In order to measure the impact of social tags, additional comparative analysis should be done with the analytics of the usage of search terms. That would require access to specific user data.

While this chapter presented the empirical evidence about the current practice of organisations and users, the next chapter provides the context for it.

6. The Context of Users' Participation

The relevance of social tagging to discoverability was assessed in the previous chapter on the basis of interfaces and user data. This chapter aims to provide context for the results of the document and data analysis, and to assess the relevance of related activities to discoverability based on perceptions of staff and users. Therefore, the current chapter contributes similarly to the previous one in answering the main research question: what is the relationship between users' participation and discoverability of the digital collections?

The data regarding the context was collected by interviews with staff and users. Four in-depth, semi-structured interviews were carried out and recorded during 19–20 April 2017 with representatives of the case study organisations at the respective venues:

1. Mahendra Mahey, Project Manager, British Library Labs
2. Mia Ridge, Digital Curator, the British Library
3. Guy Grannum, Interim Head of Systems Development Department, the National Archives;
4. Jo Pugh, Digital Development Manager in Archive Sector Development, the National Archives.

Prior to these core interviews, seven unstructured interviews were also held via Skype with Mahendra Mahey between 24 May and 22 November 2016 in order to learn about the greater institutional context of social engagement and potential topics within it (the first three interviews were not recorded, but notes were taken).

24 semi-structured interviews were carried out and recorded during 14–16 August 2017 with users of the British Library (13 interviews, including 1 pre-arranged in-depth interview) and the National Archives (11 interviews) at the respective venues.

All recorded interviews were transcribed and coded, and NVivo Starter and Excel were used for analysis and figures.

Next, the results of the interview analysis are presented for each organization (Section 6.1) and for users (Section 6.2). The sections are divided into subsections by the seven actors in the activity system according to the Engeström's framework of the

activity system (see Chapter 3). The abstract activity that is central to the analysis is users' participation with the collection items.

Comparative diagrams of the themes mentioned by the interviewees from both organisations are provided in the beginning of each subsection for a quick overview. After each diagram, the themes are briefly explained and mostly presented in the sequence of most mentioned. Sub-topics are marked in bold in the text within theme explanations and are not included in the figures. Readers interested in learning about the practices of the organisations in more detail can follow the quotes in Annex 1 as referred to at the beginning of each subsection of Section 6.1. Hyperlinks for further information are added in footnotes in the Annex. Quotations which included personal or sensitive information were either modified by leaving out names or similar details or were left out entirely, but are referred to as themes in the current chapter. If a named person has already been mentioned publicly elsewhere in relation to the referred activity, the name also appears in the Annex. The information provided by users was more general and similar to each other; thus their quotes are occasionally mentioned within Section 6.2.

Each actor's relevance to discoverability is assessed by the criteria directly occurring from the definition of discoverability: "the description or measure of an item's level of successful integration into appropriate infrastructure maximizing its likelihood of being found by appropriate users" (Somerville and Conrad 2014, 3):

1) Does the actor address an item's level?
2) Does the actor contribute to integration into appropriate infrastructure?
3) Does the actor contribute to item's ability of being found (given it is made available)?
4) Does the actor target appropriate users?

The actors directly contributing to discoverability are marked in grey in the diagrams at the beginning of the sections. (See Chapter 2 for more information about discoverability and related concepts).

The issues mentioned by staff and users are compared and discussed in Section 6.3.

6.1. Institutional Activity Systems

The current section aims to address how staff members perceive the institutional context around the activity of social engagement, and how relevant the actors are for discoverability of digital collections.

The explorative approach presents the variety of topics that were mentioned by staff members in relation to enabling users' participation. But the scope of the themes should by no means be considered exclusive from the institutions' points of view. The aim is not to give a systematic and comprehensive overview of the organisations' activities, which can be found in the formal reports and yearbooks of the organisations, but to explore relevant actors related to facilitating social interaction.

The interviews with Mahendra Mahey were mainly focusing around the Flickr collections. The amount of information and his notion that "there's actually a little bit of a ecosystem opening in this collection" led to the creation of the network of related activities, which can be found in Annex 3.

The interview with Mia Ridge mainly discussed the crowdsourcing project around the collection of playbills which was about to be launched. In addition to the National Archives' main catalogue Discovery, Guy Grannum gave an insight into the previously run Wiki site Your Archives (see also: Grannum 2011). Jo Pugh focused mainly on the Flickr collections of the National Archives.

Figure 6.1. Word cloud of four staff interviews.

Figure 6.1. represents stemmed words (e.g. 'think' and 'thinking' counted as one and are represented by a single word). If constructed separately by organisation, the word clouds did not differ much. The interviewees referred to ‘people,’ signifying both staff members and users. In the context of this research project, it is significant that the word ‘people’ is central.

6.1.1. Subject

Here, we include both the staff of the library and archive and the collections of these institutions as the Subject, and ask who is behind the activities of engaging people with digital collections. What characterises the subject? The respective quotes by the interviewees are listed mainly from 1-52 and 129-147 in Annex 1.

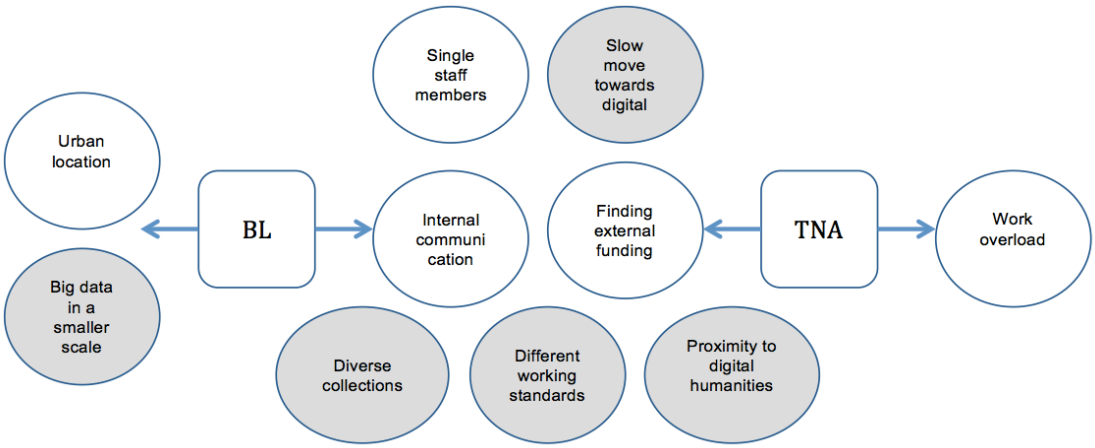

Figure 6.2. Comparison of themes for Subject.

- Single staff members

All interviewees seem to perceive the activity leaders mostly being single staff members, driven by their own interests or backgrounds. **Collaboration** takes place with other activity leaders or teams from other areas of the organisation. This individualistic approach has sometimes resulted in gaps in knowledge or work if the person left the organisation. In contrast, it has also been an enabler for new staff members to initiate activity, which had been put on hold due to preferences by former colleagues.

Discoverability: No direct contribution.

- Slow move towards digital

Three interviewees pointed out that a cultural shift is needed, policies are emerging, and changes in the understandings of staff are happening. The **importance of experimentation** and the need for a **higher tolerance of risk** in this new context of digital were highlighted. There might be fear of exposing collections to re-mix due to uncertainty about usage, or to reuse due to the uncertainty of maintaining the link to provenance; thus **heavy scrutiny can be applied first** within the organisation while launching something new. Increasing **awareness of the staff** was also mentioned in regards to already-existing digital collections, the ways of using them, and the possibilities of applying other tools. The two interviewees from the Library illustrated the **multipurpose use of the platforms** by the organisation with the examples of Explore and Flickr.

Discoverability: Experimentation and risk toleration address 'the item's level' and, together with multipurpose use of the platforms, also contribute to 'integration into appropriate infrastructure.'

- Different working standards

This issue, mentioned also by three interviewees, is technical in nature but relates to the broader planning of activities, disseminating the ultimate potentially **multipurpose goals**. **Digitisation** output formats, resolutions etc. can all be determinants of whether the materials can be reused in the future for a different purpose than originally intended or not.

Discoverability: Digitisation output addresses 'items' and the ability of being 'integrated into appropriate infrastructure.'

- Diverse collections

Two participants from both organisations highlighted the variety of collection items by format, geographical coverage, etc. Both referred to instances when digitisation of a particular collection or an experiment with it **brought about the exploitation of a new tool** or platform. In the case of the Archives it was also mentioned that the characteristics of collections might limit cross-institutional collaboration: e.g. if a theme is underrepresented in the collections of the institution or they hold only black and white photographs about the theme, but it wouldn't have the same potential on a selected platform. Likewise, the content that is potentially more viral is selected for social media.

Discoverability: 'Items' are under discussion and item processing has resulted in 'integration into appropriate infrastructure.'

- Internal communication

It was mentioned once from both institutions that either the size of the organisation makes it complicated to know what is happening, or it is difficult to know who is a particular activity leader. The issue is dealt with by involving other staff in the process of preparing new crowdsourcing projects (e.g. usability testing).

Discoverability: No direct contribution.

- Proximity to digital humanities

The BL Labs' activities focus very much around digital humanities, which also brings professors from the field to the Advisory Board of the Lab. A team from the Archives is also dedicated to learn the needs of the digital humanities community.

Discoverability: The community of digital humanities can be seen as 'appropriate users.'

- Finding external funding

The need to find external funding for developing systems or for the digitisation of more items and staff related to that was mentioned by two people from both organisations.

Discoverability: No direct contribution.

- Big data on a smaller scale

BL Labs is much more focused on the data around the collection items and the concept of big data was mentioned in this context. The largest dataset was half a terabyte, but it is not considered equal to other areas.

Discoverability: In the new context of digital, a dataset can be seen as an 'item.'

- Urban location

The issue of urban setting was discussed with the representatives of the Library. The importance of being surrounded by other significant information industries was mentioned by one interviewee, whereas the other did not see the relevance and, on the contrary, the location may influence a focus on collections about other, less known regions.

Discoverability: No direct contribution.

- Work overload

An interviewee from the Archives referred to being occupied with future projects, which leaves little or no space to take a step back and analyse what needs to be further done with the functionalities already live (e.g. social tagging).

Discoverability: No direct contribution.

6.1.2. Objective

What are the institutional goals related to enabling online user collaboration?

See quotes 53-147 in Annex 1.

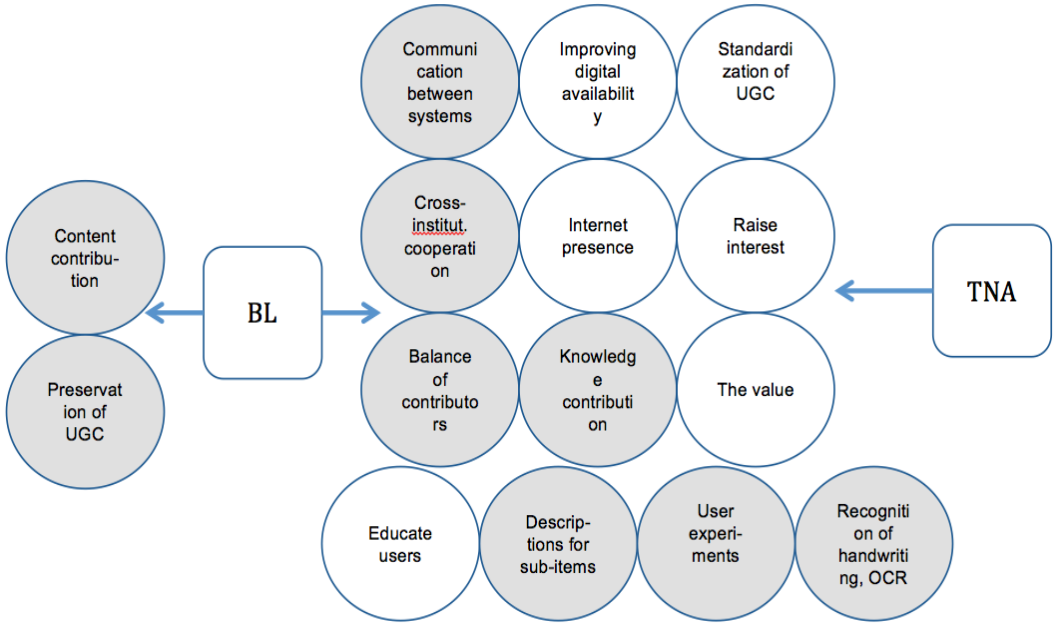

Figure 6.3. Comparison of themes for Objective.

- Communication between systems

In terms of the direction of travel, it can be considered significant that all respondents mentioned the importance of systems' ability to communicate and being **compatible for data import/export**. One interviewee raised the question whether the outputs are equally suitable for humans and machines to process – something that should ideally be kept constantly in mind and which is believed to be different for digitized and born-digital materials. The example of usability was likewise raised by BL Labs, where computational snipping and publishing of an item makes a landscape image available as portrait, for instance, if it was published that way, but was not suitable for a quick glance by a user.

All respondents welcome the **use of different systems in parallel by users**, which are ideally integrated to the relevant extent. An integrated approach with research guides or other web resources, and the creation of semantic links facilitating recommendation of items ideally contribute to discoverability, if put into practice, but it is challenging for the organisations at the moment for various reasons. The links

back to the platforms of the organisations also ensure the visibility of the items' provenance.

Discoverability: Items' ability to be found by integration into appropriate systems is ideally addressed.

- Improving digital availability

Digitisation is a precursor for discoverability. All respondents, from different viewpoints, mentioned the objective of making items digitally available. Collections can be digitised on the basis of the interest of the funder or for preservation purposes (e.g. for old materials with vanishing text), but it may not necessarily match with the interest of users. In the case of crowdsourcing project like Playbills, the visibility will be increased for **a set of items at a time,** instead of a whole collection, in order to complete a logic unit of items together (e.g. a volume into which the playbills are bounded by the library). It was also the case for photographic collection revealed on the Archives' Flickr pageThe institutional catalogues and projects of BL Labs are allowing serendipity and making the **whole collections available** for contributions.

Discoverability: No direct contribution.

- Internet presence

Visibility is increased by publishing items on the **platforms with existing potential users** who are already present. BL Labs also welcomes **users to set up own environments** where copies of the items are presented. But it goes against the objective of creating integrated systems. Alternatively, it was mentioned that users should be empowered with tools that enable them to share and promote the content.

Discoverability: No direct contribution.

- Cross-institutional cooperation at the item level

Both representatives from the archives and from BL Labs mentioned the need for cross-institutional cooperation at the item level. The Archives practice the **ingestion of records from smaller institutions** who do not have catalogues. But this brings along another issue of avoiding double entries in search results. It is predicted to take time working cross-institutionally to **make public domain items visible and available**.

Discoverability: Item level and integration into appropriate infrastructure are addressed.

- Balance of contributors

Both representatives from the Library and one from the Archives saw the need to contribute to a better balance of voices of people with different backgrounds, **regions, gender, race**. This issue relates to the **digital divide**, as people with knowledge about the items may not have technological capability to contribute.

Discoverability: Addresses potentially 'appropriate users.'

- Knowledge contribution

This issue relates to the balance of contributors from the perspective of the digital divide. Apart from that, the issue varies from a more general mention of dialogue and "saying something about something" – to detailed knowledge that people share about items.

Discoverability: Item, ability of being found and appropriate users are addressed.

- Educate users

From both institutions, it was mentioned that users should get some extra information about the items, which is useful and interesting, and which enhances the existing knowledge.

Discoverability: No direct contribution.

- Raise interest

Two aspects were raised by both organisations: first, the presentation of materials should enable users to get a hook; secondly, the contributions by users should help others to get a hook. The Archives have also predicted the interest of users and composed different sets of items from various aspects.

Discoverability: No direct contribution (the qualities of the item do not change).

- Descriptions for sub-items

The differences in the structure of library and archival catalogues define two separate questions for each organisations. The Library struggles to find a way to make available and searchable **descriptions of items within an item**: e.g. the

description of an image or map from a book or a playbill belonging to a volume which has been bound together by the Library. The archival catalogue is hierarchical already, but the question is to what **level of detail** should the descriptions go, as there are numerous potentially interesting aspects to the items (e.g. naming a person in a will versus mapping possessions and social networks).

Discoverability: Item in detail, appropriate infrastructure and ability of being found are addressed.

- The value

Interestingly, the issue of value was raised by three interviewees from two aspects: what is the value of opening up collections and trade-off with the "crown jewels"; and what is the value of user contributions? These questions remain unanswered. Articulating the output of contributions to the contributors themselves was also seen necessary; e.g. making a sourced mark-up of Playbills available through the main catalogue of the Library and notifying volunteers of their contributions becoming visible articulates the usefulness of their contributions and might motivate them to do more.

Discoverability: No direct contribution.

- Recognition of handwriting, OCR

Both organisations are involved with initiatives to improve computational ability to recognize handwriting. BL Labs sees OCR (optical character recognition) as a stumbling point for printed materials, because the texts recognised with older technology might produce results like 'd?g' if a letter "o" was obscured due to the folding of a paper. Therefore, OCR correction is needed.

Discoverability: Item's ability of being found increases.

- User experiments

Both organisations mention the importance of giving users an opportunity to experiment with collections. The Archives staff, for example, switched on the social tagging feature without much direction of or limits on users so that they could go back after a while and see the patterns, which might **result in concrete objectives with which the organisation** can proceed. A threat that has turned into reality is that staff become occupied with other things and the step back for analysing the user behavior

is not taken. It is also planned to create a generic comment field for Playbills and see how it is used and which, if any, tags are attributed.

Ideally, in social media the user community might become self-regulating and, for instance, decide the appropriateness of tags themselves. The Labs see their role in **partnering with users** as getting and solving new ideas or research problems, facilitating discovery in a new way, repurposing the content, enabling multipurpose outcomes, and getting publicity due to engaged users who are looking for fun and entertainment.

Discoverability: Yes, if an item's ability of being found is increased. User experiments possibly integrate the items into additional infrastructures.

- Standardisation of UGC

The user-generated content (UGC) comes in a standardised way if a crowdsourced applicationlike Playbills gives contributors limited freedom of action. At the same time, users may choose between **different levels of tasks** according to the time or knowledge required. The Discovery catalogue faces the need to start standardising the UGC by first enforcing category selection similar to Wikipedia and creating an option for private tags.

Discoverability: No direct contribution.

- Content contribution; preservation of user data

One interviewee also mentioned a project for content contribution (quote 79), and the need to preserve UGC. During the data analysis for the previous chapter, it was also witnessed that some users had deleted their user accounts and that the UGC attached to them then also disappears.

Discoverability: As content contribution is required together with the metadata, it addresses the item's level and its ability of being found, and likewise the preservation of previously attributed metadata.

6.1.3. Tools

What mediates the subject and its objective?

This topic concerns also two previous issues and therefore the topic is split into three parts: Tools; Subject-Tools, Tools-Objective. See quotes 148-288 in Annex 1.

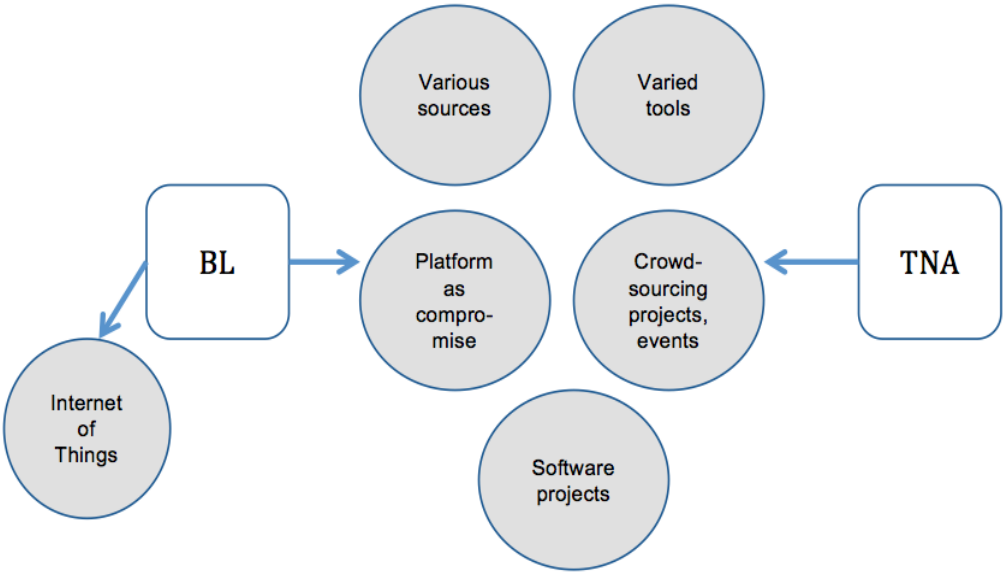

Figure 6.4. Comparison of themes for Tools.

- Various sources of input

Taking an example from georeferencing, this **special kind of tagging** needs special software and therefore takes place in another interface, but the outcomes are also intended to be visible from the main catalogue. But not all collections need it: for instance, if the whole volume of playbills is around the theatre in the same region, it is not meaningful to ask people to spend their time locating each item. The same platform may not be suitable for both items and sub-items at the same time: "Catalogue is one thing, object is another." In one instance, the need to use more UGC was mentioned, with the obstacle being that the two systems of tags and user data were not communicating with each other.

A problem with various sources is seen in producing all search results, including **overlapping results** without an option to see the latest by content or unique entries only (e.g. webarchive results). A 10-year-old catalogue needed to be closed and a new one built in order to facilitate developments. The issue is also related to digitisation outputs mentioned under different working standards of the Subject: images files of the text cannot become input for future systems.

Discoverability: Item's ability of being found is increased by different kinds of tagging; integration into appropriate infrastructure may contribute to discoverability, but may also put users off, when search results deliver too much noise.

- Varied tools

The variety of existing and self-developed tools were mentioned by the library, including the mechanism for snipping out images and exhibiting them via Tumblr, use of software libraries to describe the images, machine learning technologies, and finding similar images and relating them. The IIIF international interoperability framework is in use by the Library for easy image sharing and download. Artificial intelligence was used by a group of users, who tagged images and launched their own interface so that other users could suggest improvements and contribute. Wikisource was mentioned by the Archives, but wiki as such is considered to be quite niche and Your Archives wiki was closed even though it delivered UGC. The Archives is missing a feature to endorse tags: a number of similar tags attach to the item and possibly give it a better ranking in the search results.

Discoverability: Image processing with software libraries as part of the 'appropriate infrastructures' results in new descriptions.

- Use of platform as compromise

Flickr is not seen as the best solution. Insufficient searches, broken stats, the risk of changing its policies in the case of a change of ownership, a place for photos rather than for illustrations, and missing the option for anonymous contributions were mentioned. But it was used because of a lack of options within the institution, it was available immediately (it would take too much time to develop their own system), and it was easy to start using it; it had a URL for every image and API for user experiments; it directed a lot of traffic, and was also promoted in the beginning by the Flickr team; the collections are live and easy to find anytime and anywhere; and similar institutions were using it, which made collaborative online events possible and brought a group of people from different institutions into a joint mailing-list to change ideas and best practices.

In one instance, Wikimedia Commons refused the Library's contentbecause these were illustrations from books but without any metadata for those illustrations.

Discoverability: Flickr API as 'appropriate infrastructure' enables mass-tagging of items.

- Crowdsourcing projects, events

The Library has actively organised events to use and describe the collections, e.g. thematic wiki edit-a-thons using mediawiki to create entries in Wikipedia, engaging teachers as mediators to teach their students geotagging, and a BL Labs Awards competition to get feedback on outcomes. Making the tagging event known beforehand brought a significant contribution even before the event took place. Once deployed, software for one project is reused for another for cost efficiency and staff could reuse their code. Playbills was designed in a way that the risk for spam is low and therefore users can also contribute anonymously.

Discoverability: Item's improved 'ability of being found' is a direct outcome of these events.

- Software projects

The Zooniverse platform is used for collaboration to develop tools by both organisations. One of the burning issues is augmenting item-level descriptions. Projects for enhancing printed and handwritten text recognition are also taking place. The quality of OCR is sometimes an obstacle to building something new upon it and enabling communication between systems (see next).

Discoverability: OCR improves the ability of being found for textual items.

- Internet of Things

Presenting the BL collection in Flickr together with the link to the viewer made Flickr robots go through the text on the image page and provide tags of its own.

Discoverability: Image's ability of being found increases by making the text around it searchable.

Subject-Tools

How do the characteristics of the Subject influence the exploitation of Tools or vice versa?

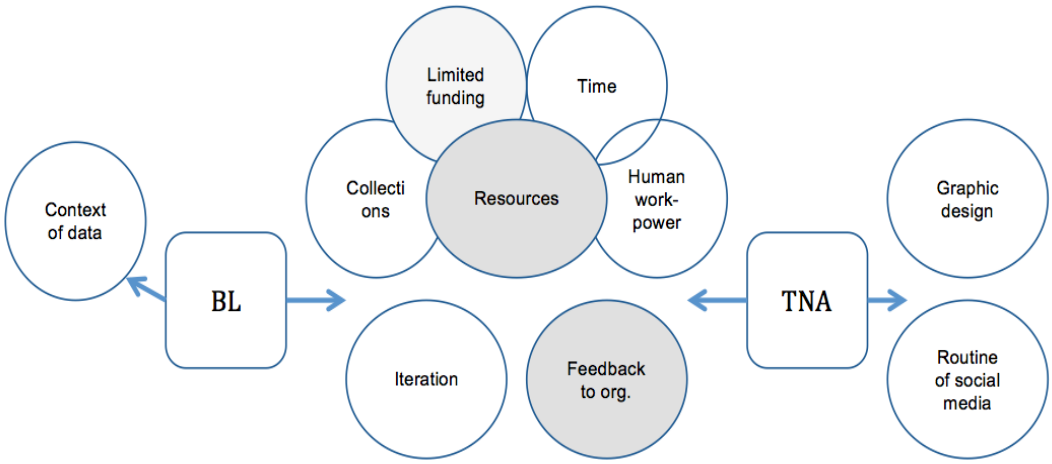

Figure 6.5. Comparison of themes on the scale of Subject-Tools.

- Resources

All respondents referred to different resources related to Subject that affect the choice or use of tools. **Limitations on funding** excluded possible describathons or other events related to the content of newly-launched crowdsourcing platforms, and drive institutions to figure out low-cost opportunities to motivate and acknowledge significant contributors or users who have reused or remixed the content in a new way. On the other hand, external funding is believed to enable longer planning and more scrutiny. The Archives staff was concerned about the resource extensiveness of the previously run wiki "more from technical rather than moderation point of view," and the cost-benefit ratio was questioned due to niche audiences involved, even if UGC was perceived as useful and valuable.

The Library staff also raised the following issues related to resources:

- **Time**: selection criteria for a platform may be related to limited time, especially under a project framework. That is why the concept of perfect data does not fit into the context of experimenting and getting some quick feedback. Likewise, quick experimental projects require an opportunity-to-fail mindset.
- **Human workpower** is needed: humans are still better than machines in many areas, and systems need to be constantly maintained by humans. If a system is not maintained, it is unlikely that staff will find time to exchange outdated links. The necessary availability of a software engineer gives an opportunity to start preparing a project.

- **Collections**: some collections can be excluded from the opportunity to become available or visible due to their nature (e.g. digitisation may take place according to the theme required by the funder of the project or by physical qualities like only single sheet materials due to the respective availability of a scanner etc). Additionally, as mentioned previously, the variety of collections sets the case for needing different platforms.

Discoverability: Lack of resources to organise describathon events, where people add descriptions to the items, directly affects the item's ability of being found.

- Feedback to organisation

It was noted from both organisations that experimentation with platforms and tools serves foremost the aim to give the organisation feedback about relevant functionalities together with preliminary results on proof of concept. After experimenting in a lightweight platform, it is considered meaningful to start developing on their own: e.g. testing workflow-related ingestion of UGC and how to make it available. A lightweight way may still be time-consuming: BL Labs projects demand a lot of prep work.

Discoverability: Opportunity of creating 'appropriate infrastructure' contributes to discoverability.

- Iteration

The need to design iterative processes was mentioned once by both institutions. Before launching the Playbills project, the initiators already had a post-launch to-do list ready. In the other case, it was claimed that there is no time for evaluation of progress or finding ways to use outcomes as input; and later developments need a checklist so some available features are not forgotten to be "switched on" for different formats.

Discoverability: No direct contribution.

- Context of data

On one occasion, it was highlighted that a tool must provide an option to publish limitations and context together with the data.

Discoverability: No direct contribution.

- Graphic design

On one occasion, the issue of graphic design was raised as it may limit people to specific roles: for instance, when the logo used as a profile picture is presented as dominating, then staff tend to speak more formally and users may get a perception of formal communication with the organisation instead of the free-form talk expected in social environment.

Discoverability: No direct contribution.

- Routine of social media

It was also mentioned once that engaging in social media has become routine, and that reaching out is believed to be not as effective as it used to be initially.

Discoverability: No direct contribution.

Tools - Objective

How do the tools employed affect the objectives or vice versa?

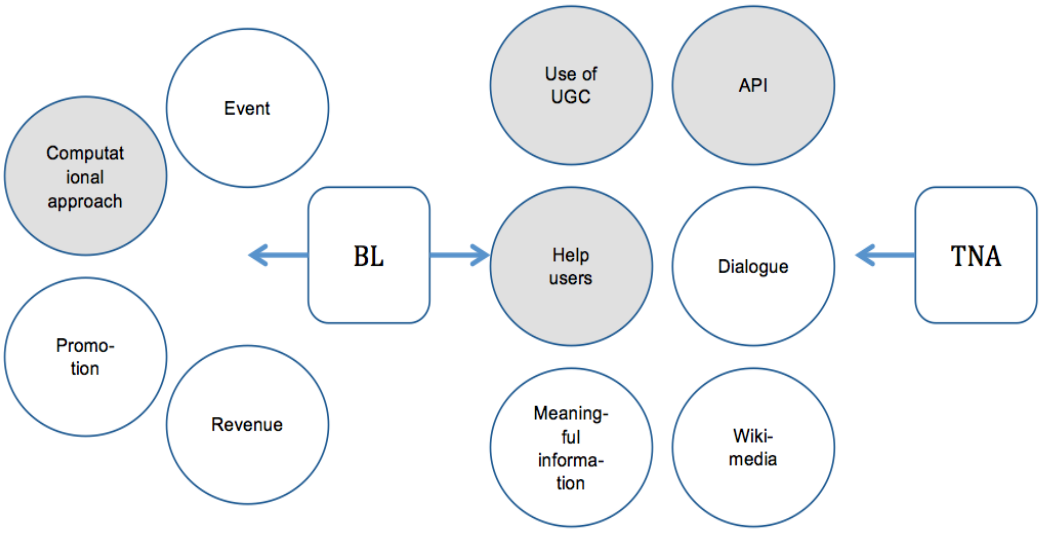

Figure 6.6. Comparison of themes on the scale of Tools-Objective.

- Dialogue, user feedback

The Archives staff finds it important to facilitate dialogue with users, mentioning that the catalogue does not provide an option to have it, but dialogue was enabled in the wiki Your Archives. Dialogue is not something to be moved directly to the catalogue, but links from the catalogue to the discussion can be provided. The option

for users to flag inappropriate tags in Discovery is not much in use. BL Labs added the need for highlighting the importance of user feedback regarding what has been done by the users with the collections and the data about them.

Discoverability: No direct contribution.

- Helping users

Another concern in common for both organisations was helping users understand where they have landed, contextualise the setting, and giving options as to what users can do there and how. Some collections need more detailed help pages so that users can be sure, if they contribute, that they are doing it properly and in a useful way. The graphic design should support an understanding of where the UGC appears and confirm the knowledge that is will be made public, if it is to be. Surprisingly, it is not the case for Discovery: tags are not searchable by others, but only visible from the record page. Besides, the need for flexibility around the presentation of UGC was mentioned for serving the interests of different communities.

Discoverability: If help articles avoid noise and contribute to the attribution of relevant tags, then the item's 'ability of being found' increases. Flexibility in presenting UGC for serving communities with different information needs contributes to targeting 'appropriate users.'

- Use of UGC

If tags are not searchable by other users, discoverability is not improved.

When the previously-run wiki Your Archives was closed, the tags were moved to Discovery but, together with tags attributed to the catalogue, they were not made searchable. UGC in Your Archives also included notes, comments, and quality annotations appearing with the contributor's name, but there was no mechanism to move these over to Discovery. The web page of the wiki was archived, so the content is available but it does not contribute to discoverability unless someone in the future re-launches another platform which incorporates the UGC from the archived wiki page.

Discoverability: Ideally yes, but in reality in some cases there is no contribution.

- API

From both organisations the need noted for API was foremost in facilitating experiments, not only by users but also for the organisations to run their own experiments or easily collect data about UGC. Any option to extract data is essential for organising hackathons or similar events, where collection owners need to expose their data. API is important for increasing discoverability, but not for the main goals of the Archives to improve visibility and facilitate dialogue. The Archives has just launched a new API, which enhances more collections than previously and is therefore expected to be used more.

Discoverability: API experiments have resulted in increasing discoverability of items.

- Wikimedia

Wiki Commons was perceived important output by both organisations for visibility. The obstacle for the Library was missing metadata, which was needed and acquired first by crowdsourcing. The downside for the Archives was that a dialogue option was not enabled. Serendipity seemed acceptable over discoverability in a high traffic platform.

Discoverability: No direct contribution.

- Computational approach

The computational approach was discussed in relation to findability by Mahendra Mahey, who believes that we are still a long way before computers can be trusted enough to be able to fully improve discovery. He points out that 100% good data is the ground truth needed for a computational approach, and if the organisation does not have the relevant data then it is preferable to borrow good data from other institutions rather than waste time improving your data first (like BL did for a user project with British Museum). In addition, the systems, which are built and tested for 5 years or so, probably become redundant and were never launched publicly.

Discoverability: This example of cross-institutional collaboration empowers computational tagging.

- Meaningful, relevant information

The issue around a computational approach relates to providing meaningful and relevant information to the users, which was mentioned by both institutions. Computationally it is easy to apply and is applied for the BL Flickr content to create a tag, e.g. ‘sysnr00332277,’ by which all images from the same book can be brought together. But is this kind of output comprehensible by humans? On the contrary, some features are both very easily understandable by humans and curated by computers, but their relevance is questionable, e.g. browsing tags alphabetically or by 'most recent,' which also uses a lot of memory. Additionally mentioned, computer-curated content ideally summarises the themes in a text for a human, not overwhelming the user with information. Likewise, ranking the results is misleading if full text is included because it is always prioritised.

Discoverability: No direct contribution.

- Event

In an event by BL Labs, the users collectively decided on a common goal to detect special kind of items (maps among other illustrations from the books) and tag them accordingly. So, the objective was set by the community.

Discoverability: ‘Ability of being found’ was improved for certain types of items.

- Revenue

In one instance, the tool was ideally seen as part of a system to earn additional funds, which can be directed back into digitisation and then made available for user contributions.

Discoverability: No direct contribution.

- Promotion

BL runs several blogs, including the Digital Scholarship Blog, which reports the BL Lab’s activities and new tools developed by and/or for the users. Guest authors are sometimes invited.

Discoverability: No direct contribution.

6.1.4. Community

Who is the community, and who contributes to achieving the goal or benefits from the goal?

How do the organisations see the community? What characterises the community? See quotes 289-426 in Annex 1.

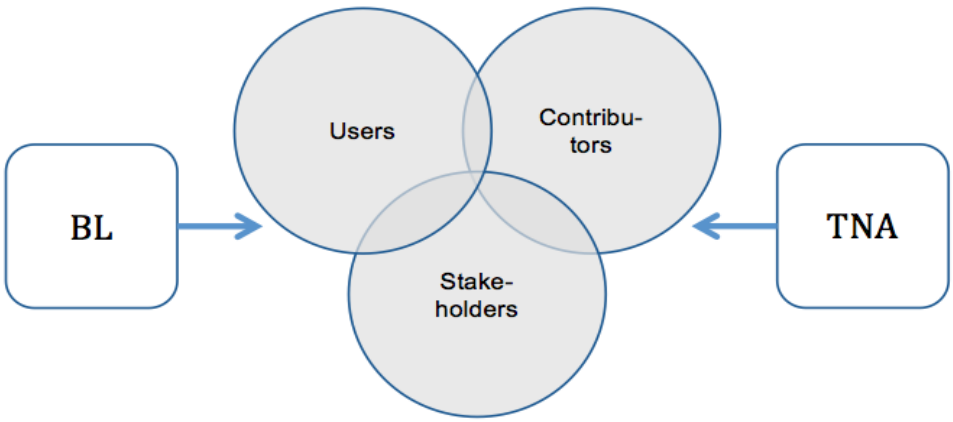

Figure 6.7. Comparison of themes for Community.

- Contributors

Contributors make up a small proportion of the total users: this is "niche sourcing rather than crowdsourcing," as pointed out by one interviewee. The number of taggers in Discovery is many times higher than in any other platform compared in Chapter 5, but the Archives do not have an explanation for this as nothing specific to that has been done. Some collaborators do a quality job out of their own passion, while others might be asked or be paid to do it (e.g. editors, archivists etc.). Contributors are not necessarily users.

From the practice of BL Labs, some contributors are retired or part-time employed people and not necessarily from the fields of humanities or social sciences, some are members of staff, and some contributors become mentors to others. Some collaborators are individuals, who create their own datasets (for hobby, entertainment, or research) and might need assistance to develop or maintain the datasets. This kind of activity sometimes brings along tagging, e.g. 'hat on the ground,' which means, according to old times, that something bad was going to happen and this situation was looked for on the images. The Labs staff steps into dialogue with top contributors,

gets to know their stories, values their contribution, and makes an effort to acknowledge them.

Discoverability: tag attribution takes place in an 'appropriate infrastructure' and addresses the item's level of being found.

- Users

Interviewees mentioned the public, "anyone interested in our collections," academics, historians, family historians, local historians, and people with personal connection to the location of a collection (this can be in relation to the historical colonies of the UK). An intriguing aspect was brought out by the Archives' representative: academics have the need for community tags, but they are not always willing to contribute because they want to publish on their finding first and do not go back to tag the catalogue retrospectively. BL, being a research library, has a lot of examples from artists. They also name independent scholars, entrepreneurs, other memory organisations, commercial organisations, charities, digital scholars, and creative maker communities.

Potential audiences are also targeted thematically for wiki events (e.g. food, fashion as an idea, etc.); for Playbills, this would be people working on theatre performance. The local community in physical proximity to the BL benefits from the outcomes of technical projects (i.e. the celebration of finding things in messy data) through cultural events.

Discoverability: 'Appropriate users' who need to find the items are addressed.

- Stakeholders

Everyone mentioned the exchange of experiences within GLAM sector, whereas the Labs' representative talked about the whole sector and others mainly about their own type of organisations – other libraries or archives respectively. An Archives representative mentioned the GLAM sector jointly as organisations which were active on Flickr sharing their experiences and ideas in regards to that.

Government organisations were discussed as institutions which transfer their records to the Archivesand, on one occasion, by BL Labs for organising a competition to start-ups for which the BL was expected to set the challenge. Those named as contributing directly to discoverability were Wikipedians in residence and geography

teachers, for instance, who were participating at events about geotagging and also teaching their students to do it later on. The media also contributes to promotion.

The Playbills project engages local libraries, local historical societies, and possibly some theatres to reach the target audiences. In addition, some universities holding collections of playbills of their own are addressed in order to reach users already engaged with their collections.

Discoverability: some stakeholders initiate activities directly contributing to the 'item's ability of being found'.

Community - Subject

What characterises the relation between the community and the subject?

See quotes 331-376 in Annex 1.

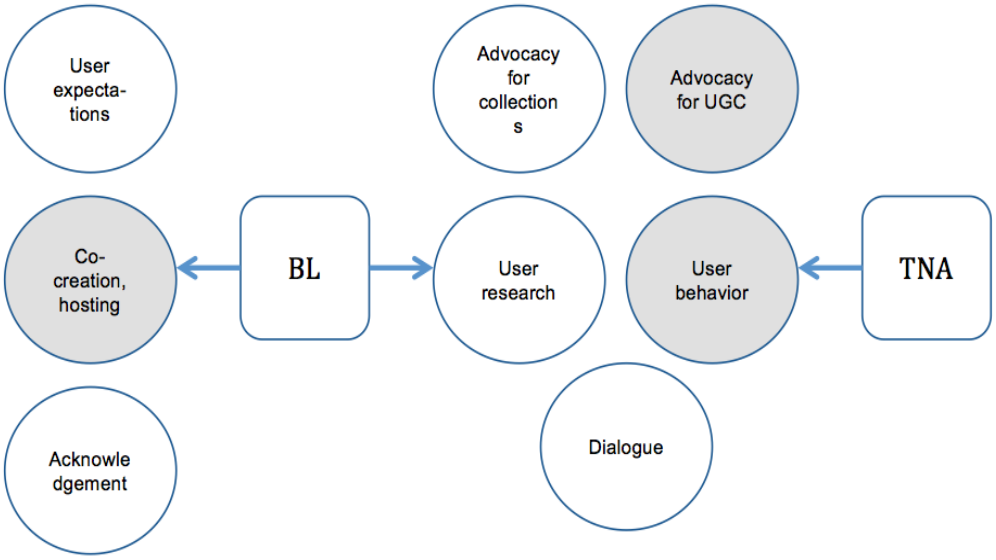

Figure 6.8. Comparison of themes on the scale of Community-Subject.

- Advocacy for collections' content

The most discussed theme by interviewees was dissemination of the content of the collections. The topics included advocacy for collections in order to **raise new ideas** about how can they be used; bringing **stakeholders closer to target audiences**; **different dissemination activities** needed for different user groups; and people with potentially specific knowledge about collections needing to be very **proactively targeted**. Advocacy is also needed to engage audiences with collections in a high

traffic environment. On the other hand, it is not seen as a straightforward issue because the content of collections is so varied and can be disseminated in numerous ways relevant to different target audiences.

Discoverability: No direct contribution.

- Advocacy for UGC

Examples of contributions may give others a push to contribute the same or in a new way. An interviewee spoke of a more holistic view of what participants can do, related to articulating the value of the expertise of participants. A mechanism is needed to know the outcome examples. It is the ethics of crowdsourcing that UGC will be preserved and put to use, not left in systems that might not be maintained and only work for a while. Equally important is to articulate this point. Knowing more about the specific needs of special user groups enables the dissemination of relevant and often specific UGC to according users.

Discoverability: This theme addresses all the criteria of discoverability.

- User research

It was claimed that the chance that someone's contribution is serendipitously valuable to someone else is small. User research is needed about current communities and systems, as well as among external communities of other platforms. User needs must be understood community by community. Predictions are also needed. Other interviewee concluded that it is equally important to understand the reasons behind user behaviour in order to make more of that.

Discoverability: No direct contribution.

- User behaviour

All interviewees accepted that users do surprise. In a positive scenario, community involvement and responses encourage staff from other departments to join the initiative or replicate similar experiments with other collection items. In addition, it may become an argument about which direction the organisation's own systems must be used or which features to consider. On the other hand, users may have a different perception of the system and may define, for instance, the appropriateness or accurateness of tags otherwise than the organisations had envisaged.

Discoverability: ‘appropriate infrastructures’ and the ‘item’s ability of being found’ are addressed.

- Dialogue

The dialogue with contributors should last longer than asking for UGC and receiving it. First, it should already have started *before* launching a crowdsourcing platform to introduce the concept and what kind of relationship the organisation wants to have with contributors. Secondly, users should also get feedback when the organisation has achieved some greater benchmark due to contributions and direct users. Dialogue held by BL Labs with top contributors has led to further collaborations with volunteers in finding and describing images.

Discoverability: No direct contribution.

- Acknowledgement

BL Labs highlighted the importance of acknowledgement, but also respects users who wish to stay anonymous. The Playbills example is to publish a leader board, but further acknowledgement by the organisation is not seen as essential. UGC in the former wiki of the Archives was published together with the name of the contributor.

Discoverability: No direct contribution.

- Co-creation and hosting

BL Labs practices co-creative projects to learn the needs of users, experience reoccurring problems, and get outcomes. It is time consuming, taking about 6 months. Some initiatives may become hosted when a member of community leads others to complete a specific task related to describing collections.

Discoverability: ‘Item’s ability of being found’ is addressed.

- User expectation

The mismatch between users' expectations and the organisation’s capabilities, which was discussed by one interviewee, lies in digitisation. Most items are not digitised and/or made digitally available. On other occasions, access to large image files is restricted for commercial purposes.

Discoverability: No direct contribution.

Community - Tools

How does the community relate to the tools?

See quotes 377-426 in Annex 1.

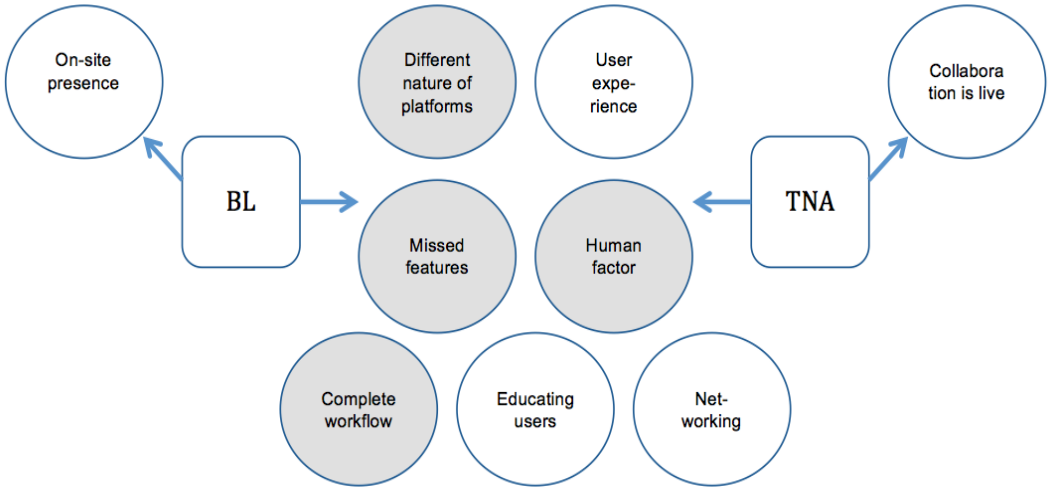

Figure 6.9. Comparison of themes on the scale of Community-Tools.

- Different natures of platforms

The interviewed representatives do not see a problem in having records, items, and discussions separately. Furthermore, it is perceived that discussion takes place elsewhere than in the catalogue, because the catalogue is not for experience but for discovery -- not for interaction, but for data. Conversation about images can take place on different platforms. People need a tool that brings community and content together, facilitates discussion rather than simply a plain search, as claimed by an interviewee. Sometimes, even signing up for an account in Flickr can be a barrier for users. Yet the effect on outcomes by anonymous users is not known. It is known from the data analysis and is confirmed by an interviewee that users on Flickr generally tend to tag in more descriptive ways than in catalogues.

Discoverability: 'Infrastructures' and 'item's ability of being found' are addressed.

- Missed features

Generally, the set-up of former tools may not meet the needs of new user communities. Specifically at the moment alternative ways are needed to access the materials, because there are frequent enquiries about easy access. Together with

releasing datasets, there should be an opportunity for users to describe how they cleaned the datasets and to upload their own outputs. Tools should enable users to give feedback not only about the correction of data, but also to endorse contributions by others. The interviewee from the Archives is also interested in getting an overview of those items that are more in the interest of users for the time being. Additionally, as became evident from the data analysis, users might attach tags to the wrong level of item. This is believed to be because a) record levels of the item are not descriptive enough, b) the record level is not created, or c) users have been transferred by an external link to the higher level.

Discoverability: Improved tags raise the 'ability of the item to be found.'

- User experience

It is believed that users expect from library and archival materials the same experience as from other well-known platforms: e.g. if people get involved with live interactive collections equipped with tools which enable them to intervene or experiment, then a bad user experience might put them off from a link to the library system which has a different design. At the same time, introducing new features benefits from new well-known opportunities, like users instantly knew what tagging was after the feature was introduced on Facebook.

Discoverability: No direct contribution.

- Human factor

A tool alone is not enough to find things in messy data, but diverse techniques are needed. Probably, human efforts are needed in parallel with computational work for a long time (mainly because the collections are not digitised or accessible, or have poor OCR quality) to benefit from or improve discoverability.

Discoverability: 'Item's ability of being found' is addressed.

- Educating

An interviewee pointed out, when everything becomes computationally readable-recognized, then new ways need to be invented to engage users. Volunteers often get a hook into a theme or material via a simple crowdsourcing task, like transcription of someone's handwritten text. This highlights the importance of the process, where that kind of involvement gets audiences historically curious, gives new knowledge on the

topic, and teaches them new skills. The process might sometimes be more important than the outcome, as was supported by another interviewee.

Discoverability: No direct contribution.

- Complete workflow

BL Labs has an example to extract collections, put them on a third party's platform (Flickr), crowdsource tags, extract images with specific tags ('map'), put them on a specialised crowdsourcing platform (Georeferencer), deploy tools (Google Fusion table) to increase visibility elsewhere (Wikipedia), and at the moment the final link of a complete workflow is being tested – how to put it back on to the library's system (Explore). Thus users are integrated into the process in many steps and the final step "makes most of the bits and pieces, which are got back." A community's active involvement and use of alternative tools might also change or postpone the decision to develop an institutional platform. The Archives has previously directed users from records to the wiki Your Archives to describe items and when the wiki was closed, the tags were incorporated to the catalogue. On the contrary, both organisations mention merely having 'switched on' the tagging feature in their catalogues.

Discoverability: 'Item's ability of being found' is improved in the process several times.

- On-site presence

The issue is related to describathons concerning literature in physical form. Security measures must be followed in order to work with non-digitised materials. An on-site presence is also required for some digital materials that are not made available online, but often that personal presence is not possible. Then the cooperation can take place by building trust and finding alternative ways.

Discoverability: No direct contribution.

- Collaboration is live

It was mentioned by the Archives that collaboration emerges on live platforms, where no push for that is needed. Therefore, in order to facilitate discussion, the content should be brought to people on platforms with already existing communities, instead of making the effort to build their own digital communities. Closed spaces

with certain type of materials in them is not believed to be "a recipe for success." Users might prefer to find instead a smaller, familiar circle of people with whom to collaborate.

Discoverability: No direct contribution.

- Networking

The experimental approach by BL Labs might have brought along the mention of networking by them. On one occasion, the Library could not support a team of users developing techniques for improving discoverability because the ground data needed was missing from the Library. But the Library had the knowledge where to request it from and thus was able to help. In another example, Wikimedia was considered difficult to use as the open source software by the Library and it is not easy to consult the authors of the code. Thus the Library can again lean on its network and ask for good practice examples from other libraries. Similarly, the Archives took part in a mailing list of memory organisations participating in Flickr.

Discoverability: No direct contribution.

6.1.5. Rules

What regulates the relationship between subject and community?

See quotes 427-446 in Annex 1.

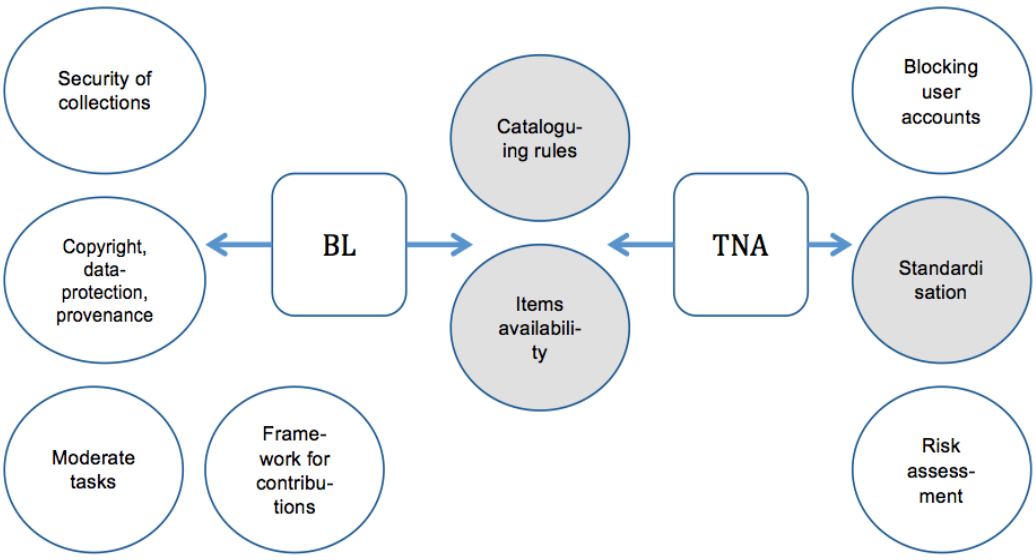

Figure 6.10. Comparison of themes for Rules.

- Cataloguing rules

This issue was mentioned by both organisations. Community contributions change the former working procedures of the library and library catalogues have child records about items inside, but there is still resistance to that happening. The push is to separate professional and community tags because there is a lack of trust in community tags among professional as well as user communities. Equally, the acceptance of data that is not perfectly clean will be a step forward.

Discoverability: If community tags are considered not trustworthy and not published, or if they are not made visible or searchable, the 'appropriate infrastructure' and 'item's ability of being found' is affected.

- Items' availability

Both organisations have examples of making the whole collection available for crowdsourcing or making a set of a collection available at one time. The purpose of the latter is mainly going through the collection and describing its items in a systematic way, getting an entire set of items described at a time.

Discoverability: In either way, the 'item's quality of being found' can be improved.

- Security of collections

Working with physical collections individually or at an event requires personal presence and following security rules.

Discoverability: No direct contribution.

- Copyright, data protection and provenance

Copyright and data protection rules have to be followed both by organisations and community, and provenance can be made clear and supported by the links back to organisation's systems.

Discoverability: No direct contribution.

- Moderate tasks

One interviewee referred to an ethic of care for the participants. Therefore, organisations should not demand too much from the volunteers or at least provide a choice of option in how deep to go with the contribution.

Discoverability: ‘Item’s ability of being found’ will be increased in either case.

- Framework for contributions

In addition to giving participants an option to choose the level of involvement, it was pointed out that the organisation should also provide a clear framework within which to be creative. This would avoid possible disappointment from the users with non-realisable ideas.

Discoverability: No direct contribution.

- Blocking user accounts

The accounts of inappropriate taggers are to be blocked, but this has not been an issue.

Discoverability: No direct contribution.

- Standardisation

The Archives, based on their experience with Your Archives, mentioned the standardisation of crowd-sourced content. It is believed that content management is still needed, especially in the form of assigning tags into a category. In case of social media, it is believed that users follow the community norms that exist there.

Discoverability: ‘Item’s ability of being found’ will be increased if tags can be also browsed by categories.

- Risk assessment

The fear of not knowing the context and purposes to which the users would put the digital images and fear of publishing controversial images might hold staff back from increasing the visibility of the content. It can be solved by assigning the task of monitoring the reactions to controversial content and putting a take-down policy of images into effect in case of demand.

Discoverability: No direct contribution.

6.1.6. Division of Labour

How is the work divided among the community in order to achieve the objective?

See quotes 447-468 in Annex 1.

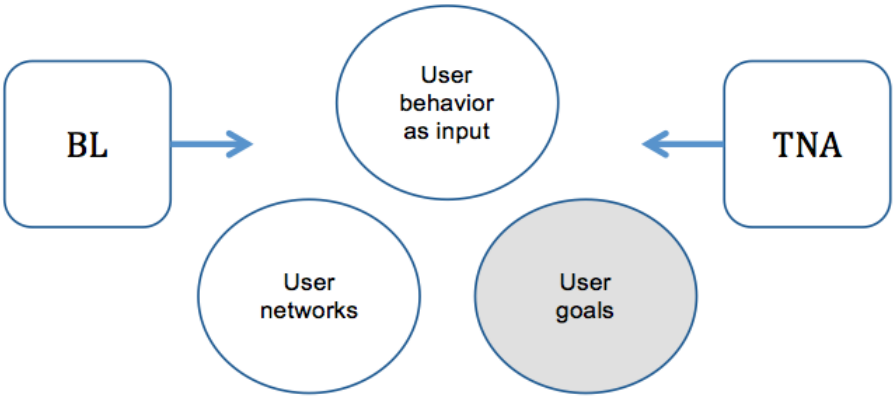

Figure 6.11. Comparison of themes for Division of Labour.

- User behaviour as input

The Library captures and prioritises its objectives by following the user behaviour. Co-creation gives the staff the knowledge needed about the specific needs of the users, especially those called "digital scholars." A raw idea by a user might become a task for the staff. Missing feedback from users stops staff from adjusting the practice of the Archives as well. Horizon scanning by staff is needed.

Discoverability: No direct contribution.

- User networks

In case of BL Labs, a volunteer has become a mentor to others, teachers might involve their students, etc. The Archives mentioned an example of a local newspaper picking up the Flickr collection topic due to an engaged volunteer. User contribution might also get others hooked to explore the collections and then add their own say.

Discoverability: No direct contribution.

- User goals

The data provided by the Library has to be **flexible** to enable users to define their own goals around what to do with that data. It is not clear if the same users would return to collaborate when new sets of items are made available. More descriptive annotations and transcriptions marked up appropriately to allow people to use them is seen as the future by the interviewee from the Archives. But defining value first and then retrospectively improving metadata is challenging in both size and time.

Discoverability: 'Item's ability of being found' as well as being used for 'appropriate infrastructure' (including those created by users) are addressed.

6.1.7. Outcomes

What outcomes emerge while achieving the objective?

See quotes 469-498 in Annex 1.

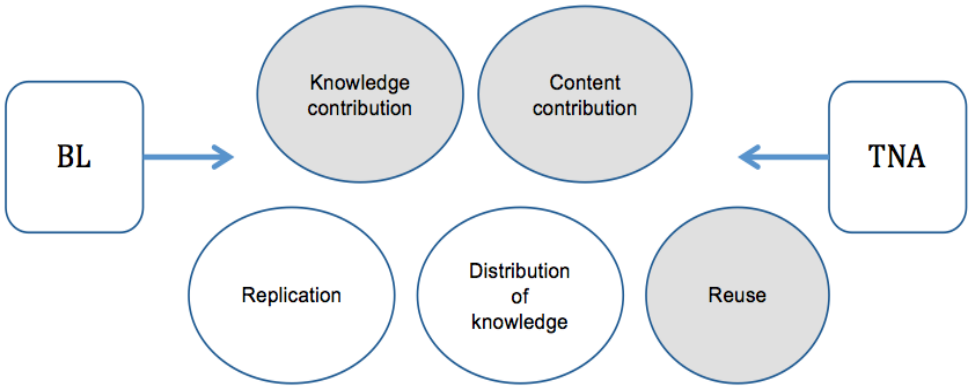

Figure 6.12. Comparison of themes for Outcomes.

- Knowledge contribution

Tags, especially if they can be put back into library systems, are considered important for discoverability. The Archives' representative points out that categorical tags are useful. Attaching a topic to a record is useful, otherwise the tags are too unique and should be standardised.

Discoverability: 'Item's ability of being found' is increased.

- Content contribution

When the public has been invited to contribute new materials, they have to be submitted together with descriptions.

Discoverability: The items and their 'ability of being found' are addressed.

- Distribution of knowledge

For instance, new articles are published on Wikipedia related to the BL collections on Wikimedia Commons.

Discoverability: No direct contribution.

- Reuse

60 examples of reuse of collections or reuse of the data related to the collections for research, commercial, artistic or teaching purposes have been presented over the last years to the BL Labs Awards. The Library has also organised events on top of these reuse outcomes. One example includes a co-created user experiment to find Victorian jokes and let comedians present them to the local public.

Discoverability: The examples include increasing 'ability of items of being found.'

- Replication

The BL Lab's story of the one million collection in Flickr was not planned as such in the beginning. But, reflecting all the elements of their activity system, it formed into a well-replicable experience and was indeed replicated.

Discoverability: No direct contribution to the collections of the initiators.

6.2. Users' Activity Systems

The current section aims to answer the questions, how do the visitors of the British Library and the National Archives perceive the context around the activity of social engagement with the collections, and how relevant are the actors for discoverability?

24 semi-structured interviews were carried out. The number of interviewed visitors is not representative. The qualitative approach explores the themes raised by the interviewees in relation to users' participation with the collection items. The questions can be found in Annex 2. The questions varied a little according to the responses of the interviewees and were not asked in the same sequence for the sake of fluent conversation. All interviewees signed a consent form for recording the conversation and using it anonymously for the current research project.

One interview out of 24 was pre-arranged with one of the top contributors of the Library, James Heald, who agreed with the request to appear non-anonymously. That in-depth interview focused on the BL Flickr collection due to his engagement with it, and lasted for two hours, while other interviews mostly took 10-25 minutes. The Archives did not know the top contributors, which is why they were not contacted.

Altogether, the randomly selected and one pre-arranged interviewees include:

- 18 men, 6 women

- 13 people up to 40 years, 11 people 40-70 years (by observation)
- 3 Asians, 1 Latino
- 3 people living outside the UK (Norway, Switzerland, the United States), 4 people mentioning coming from other parts of the UK
- 9 PhD students, 1 part-time doctoral student, 2 Master students, 5 academics, researchers or university teachers, 1 play script writer (met in BL), 2 book writers, 1 participant in a genealogy course, 1 person researching own family history (both met in TNA)
- Areas of interest of visitors of the Library include the following: environmental sustainability, the construction industry, struggles with pipeline construction, cognitive sciences, the humanities, Indonesian history, and Christianity. Visitors to the Archives were interested in Brazil-West African relations, transAtlantic relations, military culture and history, the intelligence relationship between Atlantic partners, genealogy and their own family history, architectural practices, and British policy and railway negotiations 1911-1914.

Figure 6.13 represents the word cloud for visitor interviews. Stemmed words are counted. Interview questions are excluded from the word count. The in-depth interview is excluded from this word count because of the uneven interview time. The result is remarkably similar to the analysis of word frequencies in staff interviews.

Figure 6.13. Word cloud of visitor interviews.

6.2.1. Subject

What characterises the users?

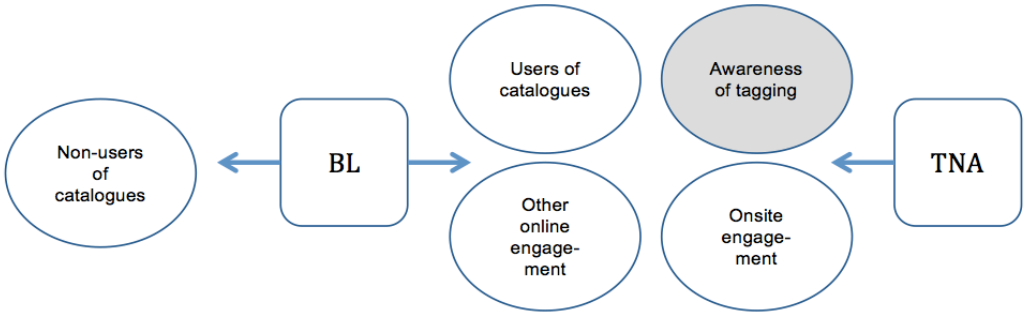

Figure 6.14. Comparison of themes for Subject by users.

- Users of catalogues

Interviewees from both venues are mostly catalogue users, on-site and remotely looking for resources for studies and research.

Discoverability: No direct contribution.

- Non-users of catalogues

A couple of people in the Library were willing to be interviewed about their interaction with collections, but had not actually used the catalogue. Their main purpose at the time being was to use the Library as a space to work or study.

Discoverability: No direct contribution.

- Awareness of tagging

Interviewees from both venues represented people who knew about the tagging functionality but had not added tags, and those who had not noticed the option to tag. One interviewee who was a user at both institutions had not noticed the tagging option in either catalogue. One person amongst the randomly interviewed visitors at the Archives had attributed tags to Discovery. The same person had also used Explore for ordering specific books, but not tagged there.

Discoverability: Taggers increase 'item's ability of being found' and users of the catalogue are 'appropriate users.'

- Other online engagement

Some interviewees follow social media accounts of the visited organisation or British Library or other libraries. One user of the Archives mentioned taking a look at the blogs of the archives she was about to visit. Mostly they do not interact on the posts of those pages. One person at the Archives mentioned an occasion when he had clarified a fact during his doctoral work and then contributed the suggestion for a correction to slavevoyages.org. A non-tagger from the Archives has contributed to Wikipedia if he notices a mistake to correct.

The tagger of the Library has contributed to Wikipedia since 2004 when he was using it and didn't always find things there of personal interest. Then he searched the Internet and brought pieces together to Wikipedia. He also filled in some bits in Italian Wikipedia to practice his pre-existing language skills. In the early days, he contributed to articles related to his occupation – physics. Previously he has been involved with GLAM by using online resources for researching his own family history.

Discoverability: No direct contribution. Visibility is mainly increased in the social network sites.

- Onsite engagement

The interviewees had mostly not participated in events organised by the organisations. A few people mentioned having visited free-of-charge exhibitions at the Library.

Discoverability: No direct contribution.

6.2.2. Objective

What has or what would make people interact with the catalogue or social media and possibly to tag items?

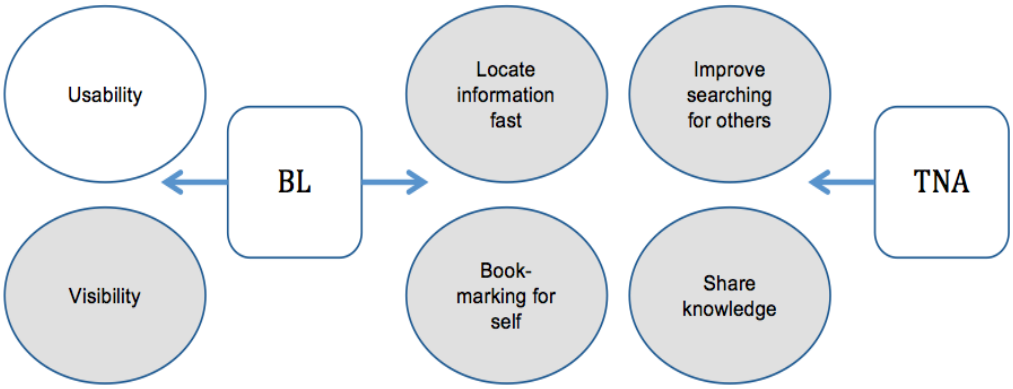

Figure 6.15. Comparison of themes for Objective by users.

- Improve searching for others

In general, tagging was perceived as contributing to making collections more searchable, and to improving ease of classifications and ease of research. The main argument from both groups of visitors was tagging to possibly improve searching for others in case a keyword was missing. Many mentioned adding **more specific keywords** than were already represented. "There are certain issues that are not included too easily in searches however hard you do the keyword search," noted an interviewee. "I find it for me a little bit hard to find subjects, so tags would be for convenience for me and others." This was supported by another person. The BL tagger was involved with **georeferencing**, i.e. matching points on an old map with another, often contemporary map. The goereferencing accuracy is dependent on the quality of the map against which the georeferencer is georeferencing.

Alternative keywords were believed to contribute in better findability or "in directing people towards resources that they didn't know existed or **didn't know how to search for.**" It was also suggested that tags could contribute to the **ranking of the**

search results. Interviewees from both institutions pointed out that tagging helps users not to miss the treasures and lets people find materials that they did not know precisely and let them know the materials at which the tagger was looking.

Discoverability: 'Item's ability of being found by appropriate users' and being integrated into appropriate infrastructure are addressed.

- Bookmarking for own interest

The second most-mentioned objective for tagging was to bookmark items for oneself. For instance, a doctoral student found, "I would use it for my current project at some point to kind of keep track what I've used and what I'm looking for, so I don't need to keep searching the same things over and over again". He believes that some tags could be useful to others. An academic believes the same: "It would be quite selfish to make sure I could find that next time. I'd like to think it's a benefit to other people, but I can't pretend that I'm there to search, but if it does help that's great". Searching the archival collections was mentioned as challenging by many interviewees; therefore, tagging was believed to be useful, e.g. "to help me with my own memory when I can't remember what was the collection." Some people believed they were **tagging for future projects** while searching for something else for the moment.

Tagging for bookmarking for their own sake was also the main motive of the tagger interviewed in the Archives, but he sees it as a benefit for future scholars as well: "If they come and need anything. When I started my PhD 4 years ago, I was looking, what people looked at these documents. Now leaving my trace for the next ones."

A university teacher believed that if she tagged, it would probably be for a research project or to prepare **for teaching, reading lists** for her students, or organisation of materials: "Being on top of everything. That would be my goal."

Discoverability: 'Item's ability of being found by appropriate users' is addressed.

- To locate information fast

The third most-mentioned topic by visitors from both organisations was to save time, including examples like, "to spend less time researching specific information," and "to make research quicker, looking through archival documents is very time-consuming."

Discoverability: ‘Item’s ability of being found’ is addressed.

- Share knowledge

Two people from both institutions referred to sharing knowledge by tagging. One interviewee said, “We are helping all the civilization to get new knowledge.”

Discoverability: ‘Item’s ability of being found' is addressed.

- Visibility

An interviewee from BL found tagging useful for resources to improve their sociability. The BL tagger mentioned being thrilled that the Labs group has made the stuff from the public domain available, but frustrated that there is so much else in the Library that is not accessible in the same way. He saw **improving discoverability and visibility for reuse** of the items as a goal on its own:

> ...and as I was probably under-occupied and needed something to do. [...] It [the BL Flickr collection] had some great things obviously, but it was almost impossible to find anything. So I thought, well, this is a good idea start writing index [in Wikipedia]. [...] I'm quite pleased of it being unlocked from the library and put into system somebody else can use, via Wikipedia to Humanities Commons for instance, create new material that has to be under a license that anybody can use for whatever purpose they want, commercial, not commercial.

The index, called the Synoptic Index, enabled people to find books about a preferred location, Germany, France etc. The important condition for the taggers was that the whole collection was visible and anyone could pick a country or region of interest to look for in the books' metadata.

Discoverability: Appropriate infrastructures and users are addressed.

- Usability

In order to contribute to improving usability, the BL tagger added tags for images to be rotated (e.g. ‘rotatec’ for rotate clockwise or ‘rotatecc’ to see the image rotated counter clockwise): “It was for convenience to view images and for georeferencers to be able to set common points between old and new maps.” He also mentioned being frustrated by the lack of overview of the progress: “we still don't know how many aren't the right way up”.

Discoverability: No direct contribution.

6.2.3. Tools

What mediates the subject and its objective?

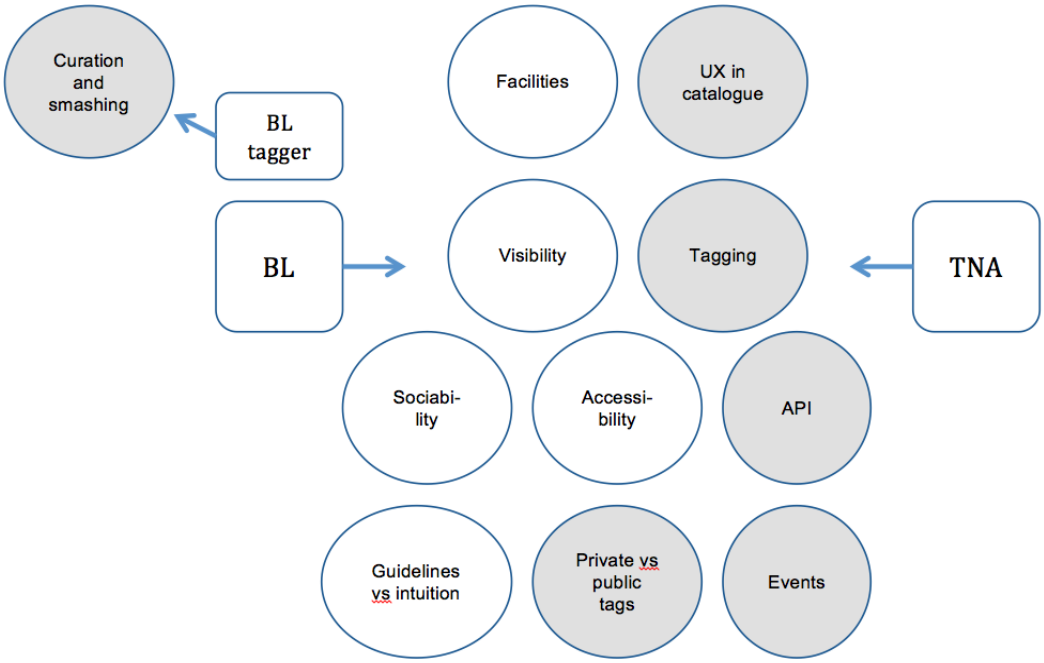

Figure 6.16. Comparison of themes for Tools by users.

- Facilities

The visitors mentioning the facilities were willing to use their own laptops as well as the facilities in the reading rooms. The tagger of the Library expressed disappointment with the **Internet speed** in the Library, which slowed down the progress of a hackathon, and many people continued from home.

Discoverability: No direct contribution.

- User experience in catalogues

Many users found that the catalogues of either institution are not easy to use. "It's not straightforward how to speak to catalogue," pointed out a Library user, who discussed the degree of free thought, instinct, and knowing how the boolean operators are treated. Another user was concerned about a variety of issues with the Library catalogues, including them being too impersonal and not easy to navigate, not understanding where to begin, not finding the right articles even if they knew they

were there. Another user found that the Library's catalogue is quite good for certain type of books, but not for everything, like old materials. Therefore she suggests **tagging be done systematically and comprehensively**. One visitor just pointed out, "I haven't bothered to working out how to use the search tool yet." The visitors of the Archives complain about finding specific items, e.g.: "It's hard to find information in Discovery if it doesn't know exactly what to look for," and "if you are going in blind, Discovery is a hard system to use well." One visitor claimed that it took her ten years to learn the catalogue, but now she is happy with it.

Discoverability: Item's ability of being found in an appropriate infrastructure is discussed.

- Missed features for visibility

Derived from the previous point, a user from the Archives suggests an interactive map for newbies including **tree of items** with clear indicative phrases under it describing what's in it. Another user says that **cross-institutional links** in the catalogues are very helpful, and that they were a reason why he had come to the Library on that day instead of working in the Archives as usual. A remote catalogue user supports the idea that items should be made **visible on multiple platforms** - she had come across items elsewhere and coming to the Library this means that she already knows what to ask for. A visitor from the Library drives attention to the fact that visibility gives a taste of what's in there, but **good subject loops** are needed because the choice of *everything* doesn't take you further: "It's the same as having it like Spotify, I suppose like having a record collection and iTunes. You have the choice of everything on your phone or computer, but like there's no like menu of options".

The BL tagger raises the issue of the **relevance of the selected platform**. Flickr was selected by the library because of many existing technological features, but he believes that Flickr would have preferred, for instance, a nice image of Venice rather than how it was mapped in a 18th century's cheap travel book. Geotags may also be annoying on the platform. And the Library's Flickr collection is far away from its best quality materials and is presented for serendipity. He also finds it important that users are able to link to the items, and that the items should **have URLs** "to make things visible, available, more friendly, accessible, machine friendly". He finds that URL is

also needed for provenance – to show clearly where the item belongs – and for personal notes to return to the item later on.

Discoverability: No direct contribution.

- Accessibility

Users value digitised materials. Many interviewees at the Archives pointed out that they were willing to **pay for the access** if it enables the institution to digitise more, and it also saves their travel costs. A visitor of the Library mentioned, "If there are things available online, you **have to be here to access** them. So that can sometimes be a hindrance." Another user of the Library, who would prefer "to read paper books, feel the paper, read the marginalia," uses many online platforms from different countries and says "digitisation makes things incredibly accessible, and I'm really used to that way." A user from the Library may have got confused with the rules about access and found himself "locked out" when trying to use the items remotely.

Discoverability: No direct contribution.

- Guidelines vs. intuition

The majority of respondents go into tagging as an intuitive process: "Like on Facebook, you don't read guidelines." Some people wanted to have a help section or short instructions at hand to take a brief look beforehand or in case of questions.

Discoverability: No direct contribution.

- Private vs. public tags

This dilemma divided into three options. First, people who prefer **tags being public** refer to the aim of helping others. The BL tagger used public tags to mark workflow in Flickr and there was no option for private tags, but he believed that couple of specific tags would not create that much noise. Some people were willing to **share their tags with a specific community**, e.g. a teacher wants to share reading lists with her students "so it wasn't just the whole world got to see that I think it's important to read." Or in case of the private sector, "it would be useful to tag stuff so it's like easily sharable with the people I'm working on with that project. People whom you know." Mostly it was believed that all people should be able to add tags to the catalogue.

Secondly, the **private tagging** option was believed to be needed for couple of reasons. Firstly, people want to add tags for own interest to find the same items again. These tags can be meaningless to others as noticed in the data analysis, e.g. 'check,' or a visitor of the Archives believed to use the function to add brief descriptions to remind himself "what's in there." Another user sees publishing the tags as a two-step process: first, attribute tags on assumption about what's in there, and later publish relevant tags after seeing the document. Secondly, private tagging was also seen as relevant by many interviewees who were doing research and being unwilling to share the findings before having a chance to publish them. One interviewee believed that he would go back and make his private tags public after publishing the article about the items. Another researcher agreed to publish tags for the items which are not available online, but to keep tags private for the items that are fully available. Another reason to use private tags was suggested for data protection, not from the user's point of viewbut because of the person included in the item. For instance, relatives of a criminal might not want the information to be easily found, but if researchers start tagging related items and persons, it might become a problem. One interviewee, who is believed to use private tags, was tagging but felt a moral tension at the same time: "although I'd feel it would go against sort of equals of the IT."

A couple of people believed they would **not tag**. One person felt it is more for younger generation, and the other person (from the younger generation) felt that he does not have time to waste on sharing anything before completing his PhD. Additionally, he mentioned using multiple platforms and therefore uses his own notebook instead to have everything written in one place. He still added, "Perhaps, if I get good at navigating the site then why not helping the others?"

Discoverability: 'Item's ability of being found by appropriate users' is addressed.

- Events

Many of the visitors to the Library had previously visited **free exhibitions** for personal, not research-related interests. Paid ones were mostly judged to be too expensive to attend. One visitor preferred exhibitions to hackathons. Others were not that exclusive and were interested in **events for increasing their digital skills**. Some mentioned that the benefits should be clearly articulated before the event, and other said those events, while being nice to have, were not their primary motive to come.

A couple of visitors expressed a preference to participate in an **event on site**. An elderly academic believed that he would like to learn new digital skills and "tagging for images sounds useful, so I would step in a public workshop on site." The other respondent thought of a probable brief engagement: "half an hour participation would do, on-site, not remotely. I'm not into computers so much." One person pointed out the need for on-site events outside the Library, in universities or closer to people.

The BL tagger appreciated an event on site to kick off the community engagement with maps. The event included **talks** about what to do with maps, and introduced an app from Japan and the Open Street Maps initiative. **A tour** in the maps department was an asset, to see the valuable maps and interesting 3D models of old maps with some explanation of how they were used. The total number of participants was about 20 people, some being an already-existing group of georeferencers. The tagger appreciated the **free atmosphere** of the event with cookies: "it was a fun day," he concluded.

Other interviewees mostly preferred online events to improve their digital skills. A visitor at the Archives said that she would participate online, if at all, because when she is present, she wants to take advantage of time to see the documents free of charge. One visitor of the Archives pointed to the need for training:

> Online videos to help would be very useful. Training videos and things like that would be very helpful for historians. Some of the stuff I use is old enough that you have to use finding aids and they are endlessly confusing to me and to the staff too, they don't know actually the way through them -- more training to everybody.

Discoverability: Increasing 'item's ability of being found in appropriate infrastructure' is partly addressed.

- Missed features for tagging

All interviewees would expect that **tags are included in the search**. A visitor of the Library would like to see tags also included in the **recommendation system**: "If I open a book record with a tag 'sustainability,' I would like it also to say here are three other books with tags of 'sustainability' in them." Another user of the Archives might want to see multiple files on screen and be able to **tag them simultaneously**, e.g. attribute a tag to all items belonging to the record.

Discoverability: 'Item's ability of being found in an appropriate infrastructure' is discussed.

- Missed features for sociability

A couple of interviewees pointed out the potential social aspect around digital collections: “archival research is very individual work, sometimes I miss that people could work together on the same topic.” Another user saw a relation to improving user experience in the catalogue of the Archives: “So things that help you using it are obviously good. Bringing groups together digitally to talk about what's in the archives in particular subject or something like that.”

Discoverability: No direct contribution.

- APIs

Most interviewees expressed interest in **learning new digital skills** and many of them were willing to experiment with APIs if they were made available and **support was provided**. “People go and identify objects for public science and so it'll be cool to see that happen here,” said an interviewee at the Library, with the idea that online guidance is needed for people using the system elsewhere. Another visitor who was willing “to expand to that level of catalogue use” was motivated by learning a new digital skill as well as **contributing through that to her subject area** of interest. Many interviewees mentioned this kind of double motivation. The direction to increase digital skills was also mentioned by universities in regards to creating digital humanities labs and similar. A university teacher was ready to try, if an API was useful for bibliography and reading lists.

Visitors, who were **hesitant** were afraid that it would not be very straightforward and simple, and did not believe they could find time for it, or were “not necessarily convinced that algorithm has the fine judgment for trained historians.”

The BL tagger finds an API good for **tagging in bulk** in Flickr. He is concerned that API use is not important for the platform owners and therefore there has not been much of response to the complaints if API had stopped working. He also points out that machine tags could be **low prioritised** as Flickr changes their representation in every 6 months or so.

Discoverability: Learning to use the ‘appropriate infrastructure’ and improving the ‘item’s ability of being found’ are addressed.

- Curation and smashing

The BL tagger sees providing data in datasets in an **acceptable file format** as an important precursor for experiments "so people can run their own scripts." For instance, the whole list of the books with the titles for Flickr images enabled him to think how to start and enabled building a list of books in Wikipedia based on location. Then he used the list for finding maps and added manually the tag 'map' to Flickr. He used Wiki editor to mark the workflow: "this book has been checked for maps, this hasn't"; therefore, the tags in Wikipedia showed the progress and kept track of what had been done.

He finds it important that Flickr is entirely **free of charge** to use for these purposes. The strengths of wiki include ease of uploading and scale, ease of use, and the creation of lots of inner links. He points out that with wiki you can make a contribution as big or small as you can, which **does not have to be perfect**, as someone else comes and adds what is missing, building upon it.

From these constraints, the tagger points out that there are tools to help with **curating collections**, e.g. creating a subset of decorative first letters in the case of the BL Flickr collection, but he perceives here a contradiction with the mission of BL Labs, who see their mission more in "producing interesting data." Additionally, he sees that the user community should **release their codes**, so that their experiments can be reproduced. He mentions two examples relevant to discoverability which are affected by the constraints of resources. First, to re-OCR the collections would enable to read the caption better for keywords, and would give a good clue that the image should be rotated so the caption would run along the bottom. Secondly, the BL georeferencing service provided by the Klokan company is financed as a project, not as core work. Thus it becomes expensive when changes are needed in the software, content is varied, and changes are needed.

Discoverability: 'Item's ability of being found and incorporated into appropriate infrastructures' is addressed.

6.2.4. Community

Who is the community, and who contributes to achieving the goal or benefits from the goal?

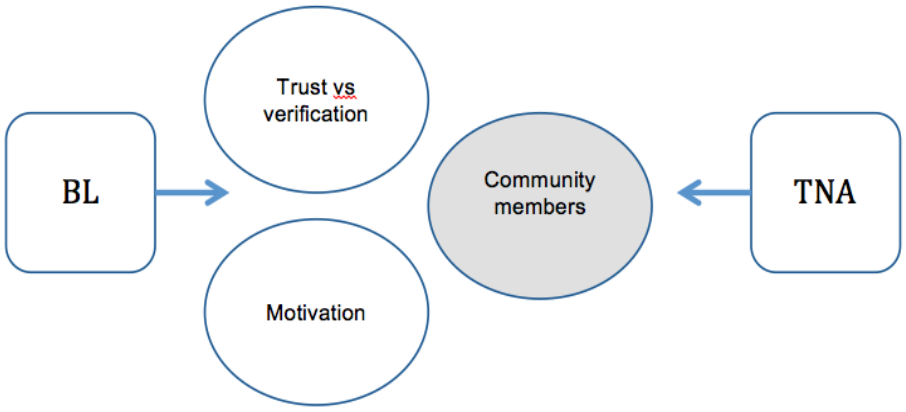

Figure 6.17. Comparison of themes for Community by users.

- Community members

The interviewees mostly see the beneficiaries as the community around the tagging activity: **all people interested in the same subjects**, users of images, colleagues, other graduate and PhD students and graduates from the same school, subject communities, and people from the same research area. A teacher suggested that it is possible that some tagging becomes a classroom task in the university, but in a restricted way, through access code or similar, and not visible to others. Some respondents said that the subject community is already known from the literature and conferences and thus it is not helpful to know the tag authors. But another opinion was that "it would be really positive to have a way of knowing who is into something similar – that would be a motivation to tag." She also saw a threat in it: "if it's not anonymous and if there's particular people who are sort of elite within it, that would put me off."

Another person commented, "I guess that finding something and then sharing it, I don't know so much about that world, but I know about programming and sort of open source coding, computer world that's a kind of general ethics of a lot of those systems. I wonder if we can learn from that to take over into this tome of knowledge, but yeah, it's really interesting." It doesn't consider crowdsourcing in memory organisations, even with the other sectors, and may refer to a need for a paradigm shift regarding the collections of these institutions. This opinion was supported by another interviewee with the notion that, if we use literature, we trust other scholars, but with tags, it would be nice to have them checked.

Interviewees mentioned collaborating with their communities by conferences, publications, or Facebook groups.

The BL tagger sees anyone interested as users of crowdsourced information. But the most **active contributing group** was, for instance, about 5 people in the case of the event for finding maps. These people were not the end-users, but included for instance a Wikipedian in residence and someone from Commons suggesting techniques to others. Wikipedians do have a habit of meeting and engaging socially, e.g. at their monthly pub meetings in London. Some of them are retired people, who have more free time. Usually a core team of volunteers forms within Wikipedia activities, which sets the direction for how things are going to be done and how other people's contributions can be used usefully. They write upwhat they are doing rather than having a live communication. The tagger expressed frustration with the categorisation in Wikipedia, because people create it without knowing in advance the eventual level of depth. But he believes that the professional librarians have other jobs to do rather than help there and, equally, "the Wiki community is stubborn-minded and does its own thing and mistakes." He concluded that the ecosystem of communities has drained away and commercial companies are providing direct answers to the needs of different user groups.

Discoverability: 'appropriate users' and contributors are addressed.

- Trust vs. verification

Most respondents would **trust community tags**: "I would trust all of them. I'm keen on the community engagement in these projects and open source software and that sort of things. I would treat them equally"; "We live in a world where lot of things are crowd sourced these days and there are community tags on websites and forums." Some interviewees thought they would trust community tags, because the resources in their interest are so specialist that others with such interests would be credible taggers.

A couple of people said that they would either trust the simple or more general tags, like country names, but would double-check more detailed descriptions. One respondent pointed out that she would trust the tags, because they help to do the first selection of materials, but eventually would read the documents anyway. People who do not think that organisations should verify the tags explain that there should be no inference or comment that "it's resource extensive, so no need to ask community to

do it, if resources are spent on it anyway"; "there is no point, if people have already looked into the items"; "you would have to allow the community to monitor itself rather than library being so authoritative." The "Library should facilitate community engagement rather than rule over" it was one opinion. One respondent suggested that the organisations should check which tags to include in the search.

A few peoplewho tend **not to trust community tags** are, for instance, "afraid that the community tags are just too big and too idiosyncratic, too many of them," or express a general concern: "I feel hesitant for community contribution without any order." One person thought that she would be not up to the task: "I'm generally more hesitant, like I'm not sure it's right, I would probably assume that it would be checked off by someone who works here." A couple of people suggested a computerised approach for verification, correcting spelling, blocking spam or doing random checks. 6 people remained unconcerned. There were also some interviewees who believed that the organisations should remove inappropriate and out-of-date tags, verifying all tags, and two people thought the organisations should also give personal feedback.

Discoverability: No direct contribution.

- Motivation

Some people miss the **social aspect** in catalogues and believe to tagging, knowing that the community of academics who work on similar things would look at it, too, is "thinking the more collaborative way, not to be in the individual work" and could "possibly grow own network." One person guessed that some would tag for having more sense of being more authoritative; she also thought that it might depend on a subject and "perhaps family historians are doing more of it."

The BL tagger believes that **interaction** is a tremendous driver:: someone noticing you, hunting together can be energising; a staff member meeting for a coffee and being interested; but also a button to 'say thanks' to and editor in Wikipedia – are the things he perceives motivational. He has also been interested in who the taggers are and how much they are contributing. Neither in Wikipedia nor in Flickr was there a leader board for that, but there are tools to find it out.

A few people from both institutions who remained **hesitant** had different reasons: "I guess I just wouldn't think that my intervention would be kind of valuable enough, that it would be done better by other people"; "I'm selfish right now, but when I tag for someone who is related to my research project, for them actually it would be

benefit, beneficial for them. They are substantially competitive"; "I haven't quite known what the others might find a useful tag."

Discoverability: No direct contribution.

6.2.5. Rules

What regulates the relationship between subject and community?

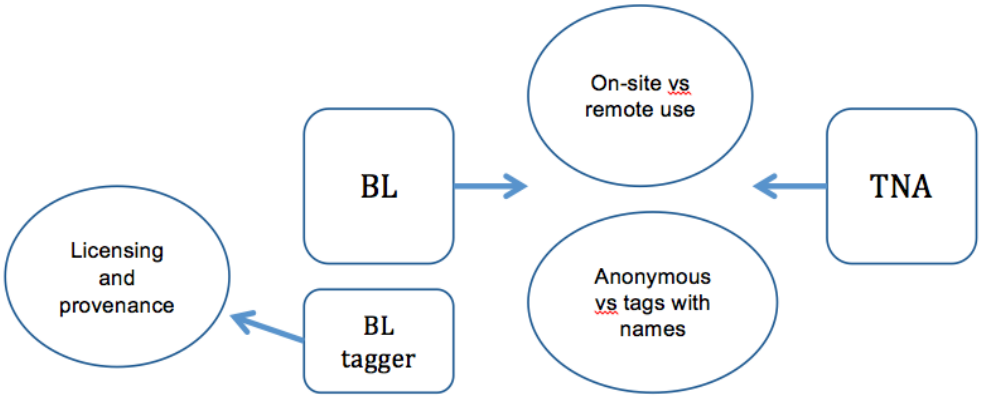

Figure 6.18. Comparison of themes for Community by users.

- Anonymous vs. tags with names

Mostly people believed that tagging is a collaborative effort and should not include names of the attributors of the tags. But there were also different opinions – that the name should be noted to add more credibility to the tags, or out of interest in knowing who works in a similar field and what materials have they looked at. For some people, it didn't matter if tags came with their names or not "as long as tags are accessible."

Discoverability: No direct contribution.

- On-site vs. remote use

Some people mentioned the access to the British Library as being too restricted. The tagger of the Library also saw the need to loosen the rules to work on site with materials: "readers have alternatives," he said.

A visitor at the Archives, who lives outside England, had trouble ordering materials in advance because between her visits her card had expired. Then she settled an agreement that staff can order the items in advance. But because of not being able

to come on site, she has arranged a yearly subscription to Ancestry to see many documents instead of paying a fee to the Archives to see one document at a time.

Discoverability: No direct contribution.

- Licensing and provenance

The tagger of the Library discussed these issues. He draws parallels with the Wiki environment, which asks anyone to provide information and where the content was coming from, and at the same time lets everyone use it. "The National Library of Scotland makes a certain amount of stuff smashable," but he sees the British Library as being too controlling about what anyone can do. He considers structure is important to follow, as are the templates on Wikimedia commons which also make provenance clear. Additionally, he finds that if different organisations would collaborate more, it would also increase the quality of the contributions by users. For instance, if a quality map from another institution would be available to compare against what people can georeference the map against, it would be much easier and more accurate than a satellite image, which might not be right on top of the object, e.g. a church.

The Library's tagger acknowledges that there are risks with large collections that people might put items into an unexpected context. For example, somebody composed a collection of topless African women by tagging the relevant items as such in Flickr. He believes it aroused hesitations in the BL Labs staff, but it was a legitimate representation of people in the 19th century.

Discoverability: No direct contribution.

6.2.6. Division of Labour

How is the work divided within the community in order to achieve the objective?

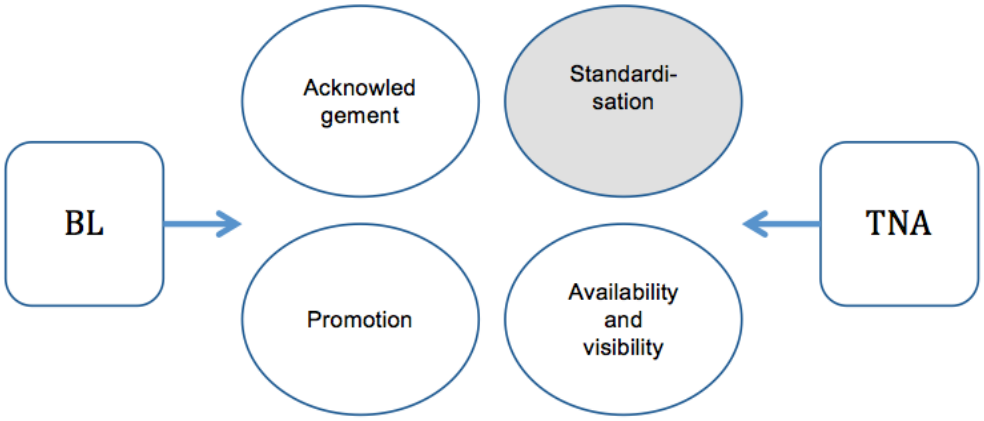

Figure 6.19. Comparison of themes for Division of Labour by users.

- Acknowledgement

Most respondents thought that **acknowledgement or awards are not necessary** in a collective culture. Some people suggested that contributors who are standing out by their massive amount of work should be acknowledged. Some people remained unconcerned.

The tagger of the Library was recognized at the BL Labs Awards competition. He thought it was nice to receive acknowledgement and a prize, but still he didn't continue to contribute and was feeling guilty about that. He believes that "just saying thanks is a huge motivation. It is appreciated. Awards can be hard on people who didn't get the award." A couple of people suggested that the **community should** have some means to acknowledge taggers or endorse tags, if it was said within the system: "something like it was suggested by so and so on such a date." Some people from the Archives felt that it is important to **acknowledge people publicly** or give **feedback in person**.

Discoverability: No direct contribution.

- Standardisation

Organisations were expected to manage tagging by standardizing the tags, because "you can have a wide array of different ideas of what should be keywords," and "guides would be helpful for knowing which tags are used currently." Suggestions for tags were considered important "so you don't just type in something that no one else uses, e.g. tags, which are most popular by the number of different

users to avoid spam by mean people, [...] so you don't end up having like 6 different tags that are all quite similar, like carbon and CO2."

Discoverability: 'Item's ability of being found' increases if the choice of tags corresponds to what 'appropriate users' would use.

- Promotion

"The knowledge of subjects are drawn up by librarians, so if they have sufficiently wide view, great; if it's too narrow, not so great. Then users can step in and help," said an older academic. In order to help, many respondents saw a role for the library in raising awareness of tagging activity as well as other features of the catalogue.

The organisations were seen as facilitators of tagging functionality, and were expected to "**inform exactly what you include** and what you can't." Many respondents mentioned that people need to be **encouraged** to tag. Sometimes people appreciated a personal approach: "probably now, spoken to you I probably try something. It would be helpful to sit down with somebody sometimes and look how to use the system." A person suggested that software engineers should be encouraged to do more mass tagging. Another person pointed towards **ethnic diversity**: the library should be "looking for sort of non-Western people, who are producing similar knowledge to this. Sort of expanding this pale-male-... I think the library should be doing something on this. Not the same people always replicating the same…"

The Library's tagger believed he would be motivated to contribute again if the activity would be **more organised and useful to other people**. He would love to give structure to chaotic data, improve discoverability and analysability, putting some order in it. His involvement with BL maps was very much related to a need and a collaboration with a map librarian, who was running out of maps for the existing community of geotaggers. So he thought of finding maps from public domain books and making links in Wikipedia to the location of the books. He feels sorry that the "georeferencers had vanished within 2 years time" after the map librarian had left. The tagger believes it is challenging to **recruit new contributors**, but it is necessary because people move from one life cycle to another, there are changes that might "take them away from editing screen," take their time away, and they no longer will be contributing, even if they wanted to. And so did he: "I got busy with other things." He thinks it is challenging to build a community around your own project, but easier

to go on to a platform where a community already exists. The tagger appreciated **direct contact** with the staff to learn about the work done in the library and, due to the contact with BL Labs, he got to know about concerns in the map department. He also mentioned an exchange of techniques with staff.

Some interviewees found that **users themselves could promote** tagging features within their communities. Another person would appreciate a feature in the catalogue to share findings easily. The tagger of the Library got to know about the BL Labs Flickr initiative through his wiki community: "Somebody posted on the daily information board in the daily high-traffic discussion boards on Wikicommons I think it was: Hey look, the British Library has done this wonderful thing, a source everyone should find images and put them on articles and so forth." Wikipedians had also started an index, a little page of some interesting books that might be worth taking images from. "And I suppose I thought I could just help out with a little bit," he said. While engaged, he wrote a post to the Wikipedia community about the start of the weekend event at Halloween.

Discoverability: No direct contribution to the item's level.

- Availability and visibility

In general interviewees wanted to see **more items digitised**, findable, and, whenever possible, also accessible. In terms of finding the items a visitor at the Archives thought, "they have done a pretty good job actually," but was concerned that something more should be done to very directly **help other people to find** the resources that they wanted. Another respondent supported the statement that "the Archives has done their best, giving us brief lines about it. In Poland, US etc. they have only titles, no idea what's in there." Some thought that the library should facilitate monitoring, what are the most looked-at items, and give higher ranking to those items. The Archives was expected to make information retrieval more understandable by "pooling more this expert knowledge that there would be more guidance for a general audience who doesn't know the codes." One person suggested that the archivists should also tag more themselves.

Instead of the approach of letting people play with the data, the BL tagger prefers to see collections curated better and data improved so that they are also usable for Wiki projects. He thinks that georeferencing was interesting but, for instance, the National Libraries of Scotland and Wales have georeferenced much more interesting

body of maps and are now making them more accessible to the national communities. "There is probably **more value**," he suggests, but understands that the digitisation of quality content takes resources: Labs' mission was interesting, but not what the library's role should be, to make **the collections as accessible as possible.**" He finds that there are many wonderful maps in the Library that they could have been digitizing and then georeferencing instead of cheap travel book maps. But that is different focus and he would not know what is right. He highlights the usefulness of monitoring the process: "If you want people to do something you have to know what they're doing, what they are looking at, what to prioritise to make available." He believes that **memory organisations** mostly **maintain control** and "don't want to let things go [...]; once the stuff has escaped to the public domain, it is difficult to monitor it." He believes that the BL Awards are a nice way to do it, but it is tricky because, in that case, they do not have much control. He proposes the Rijksmuseum in Amsterdam as a good example of letting things go and making collections available. On one hand, they got frustrated how wrong all the yellows were on representations of the painting "The Milkmaid," but it also brought more awareness of their collections, more referencing, and more people coming to see originals, going to their shop, or purchasing online.

Discoverability: No direct contribution.

6.2.7. Outcomes

What outcomes emerge while achieving the objective?

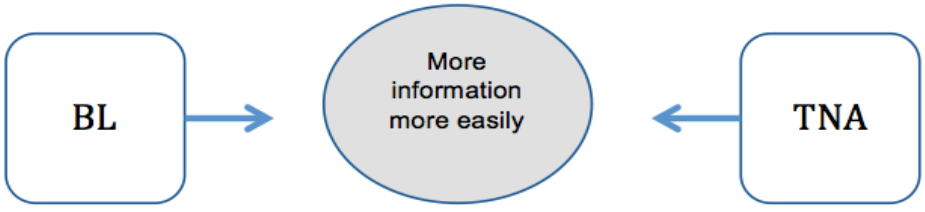

Figure 6.20. Comparison of themes for Outcomes by users.

- More information more easily

Most respondents saw that tagging enables users to find **more information more quickly** than only institutional keywords. One respondent considered that tags become important actors between paper and digital.

Some people highlighted the importance of **organised content:** "categories would be relevant." "To make the collections more accessible and more searchable, a reliable way that you can be pretty sure would lead you to not everything, because an exhaustive search would be impossible, but really a good range of stuff that you were looking for," summarized an interviewee. Some saw their research benefitting by specific information being found more easily, "list of documents or abstracts or references coming together," "more information about my research subject," "better, more detailed research with wider choice of sources," and "in the records more capturing elements, to make research more easy to other people." One person emphasised the adequate keywords: "Not too general, not too specific, but the subject community must understand, not everyone necessarily." Another person mentioned "connection to others working on similar things, expanding my own knowledge in terms of that area, knowing more about that particular area."

A completed family tree and 50,000 maps identified amongst one million images within 2 months were brought as examples of concrete outcomes. The Library's tagger thought that reuse is good publicity for the Library. A visitor of the Archives argued that it would be interesting to see what the people look up most after introducing tags in the search.

A respondent from the Archives remarked, "if everything becomes tagged, then the process of research changes." She also saw a danger that students would only look at the tagged items. The Library's tagger questioned whether the tags are used and whether they are valuable? Are the public using it for discovery? He believes that "it's too loose, what individuals have done."

Discoverability: 'Appropriate users' will find the items.

6.3. Discussion

The thematic analysis of the interviews exposed topics which answered the relation to discoverability either positively or negatively. The aim of these analyses was to define this relation, instead of evaluating the performance of the case-study organisations.

For the institutions, 84 themes, including 43 themes related to discoverability (see Figure 6.21.), were composed in total for seven actors and 4 groups, indicative of the relation between the actors. If we divide the total number of themes per actor by the number of themes related directly to discoverability for the same actor, we find that the actors related more to discoverability in institutional context are (index<2): Objective, Tools, Community, Outcomes.

The majority of themes are in common for the institutions. There are more issues mentioned by the Library, which derives from the fact that several Skype interviews with the interviewee from BL Labs were included in the analysis, which resulted in covering more topics than in the one-time in-depth interviews with the employees of the Archives.

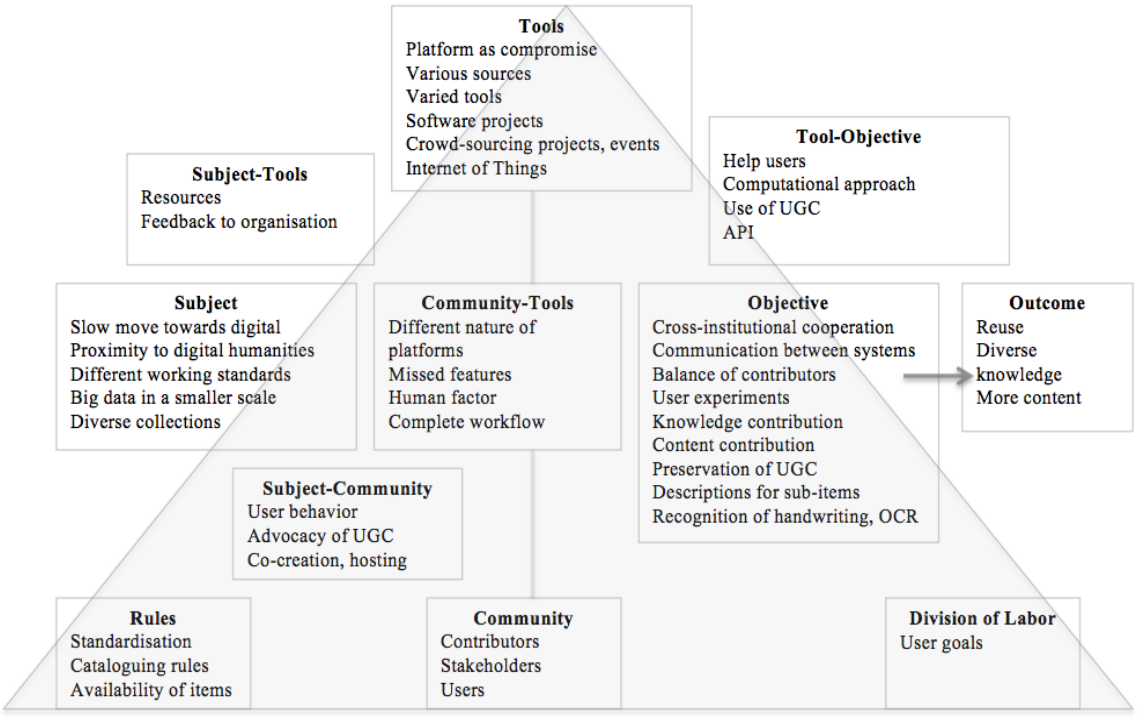

Figure 6.21. Institutional activity system for activities addressing discoverability.

For users, 33 themes, including 15 themes related to discoverability (see Figure 6.22), were composed for seven actors. More relevant actors to discoverability seem to match with the institutional context (index<2): Objective, Tools, Outcomes. The majority of the themes are also in common for the users. Again, the visitors of the Library raised more issues because an in-depth interview was carried out with one of the top contributors there.

In general, staff raised more issues than users because they had better knowledge and more experience of users' participation with the collections. At the same time, all interviewed users, except the tagger of the Library and one visitor at the Archives, had no previous involvement in tagging and therefore no varied experiences to share. They expressed their preferences briefly, if they were involved. Also, the tagger of the Archives used tagging very straightforwardly for bookmarking and had no broader view to share.

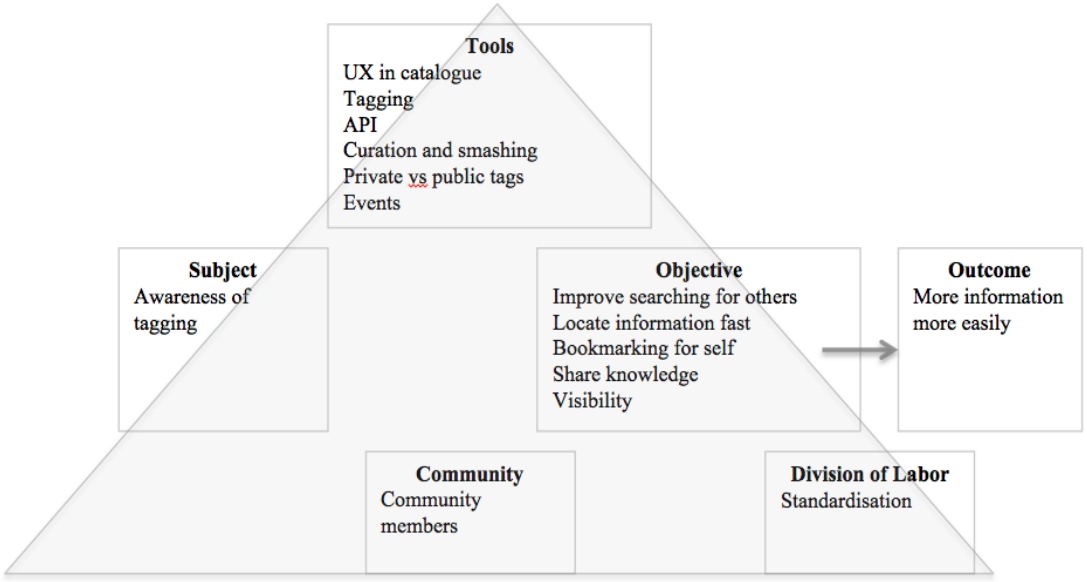

Figure 6.22. Activity system of users for activities addressing discoverability.

If we compare the activity systems of the institutions with the systems of the users, we find matches as well as mismatches. The following generalised discussion takes a note of the actors in the brackets: first, the actor of the institutional system, and second, the actor for the system of users.

Issues of common concern. Most issues were discussed from a common point of view by staff and users. Institutions felt the need for a paradigm shift towards digital, which is slowly happening in the institutions. Users perceived the collaborative culture around them, but some did not yet fully see themselves as contributing to open knowledge, using public tags. Staff, as well as some users, pointed towards raising awareness of digital among staff so they could help the users better (Subject-Tools). Different working standards within the organisations may make the digitisation outputs or datasets incoherent and thus not reusable for user experiments (Subject-

Tools). Proximity of institutional initiatives to digital humanities corresponds to users' imagination of data reuse on Humanities Commons platform and relates to training activities in the universities (Subject-Objective; Tools).

Institutions are mainly expected to make things available and accessible. The organisations are also acting upon this and digitising the items or catalogue cards. Institutions want users to contribute to different platforms, which are ideally linked to each other. Georeferencing is a good example of this. Many people noted the need to be able to search and use collections cross-institutionally. Making public domain items accessible cross-institutionally was emphasised by one interviewee. Given that even within the group of 24 visitors who were interviewed, we had overlapping users of both institutions, there is clearly a need for cross-institutional solutions (Objectives).

Both agree that users have knowledge about the items to be shared. Aiming for a greater variety of contributors and balancing the "white voices" in the descriptions was mentioned both by staff and users (Objectives) – an area often related to the digital divide, not to institutional or user preferences. The staff felt that monitoring of the engagement process and UGC is important, but it is complicated to find time for that. Users thought that it is the role of the institutions and perhaps some of the monitoring could be automatized (Objective-Community). A clear statement from both sides was that UGC should be standardised, e.g. categories introduced and most used keywords suggested during tag attribution (Rules; Outcomes-Division of Labour; Outcomes).

Co-creation, and any partnering and dialogue with users, contributes to learning more about the users and learning skills from each other (Objective-Tools). Users may not perceive whether a better user experience comes from developing the system or improving their search skills, but both serve the ultimate goal of a good user experience: e.g. a user claiming that it took her 10 years to learn the catalogue but she she is not happy with it, whereas the Archives had meanwhile launched a new catalogue system (Tools). Some staff members and a user mentioned Flickr not being the most appropriate place for non-photographic materials, like illustrations or images from books, but technological affordances overruled the content-related argument. Still, many respondents said that they would not have thought to look for the content of these organisations on Flickr. The use of Flickr API was seen as an asset and most

respondents would like to learn the institutional API option, if it was available (Tools).

Better ranking of search results, possibly using tags, was deemed necessary by both sides (Tools-Objective). Technological development is fast and so the full texts are expected to be reprocessed with OCR technology for better text recognition and search. Organisations have provided options to tag the whole catalogue and contribute to crowdsourcing projects, which focus on specific sets of collections at a time. It corresponds to diverse preferences: some users want to work with the collection items in their interest, tag them, and increase their visibility in Wikipedia, while others expect a comprehensive and systematic approach (Rules-Objective).

Mismatch. The essential mismatch between some of the mentioned Objectives of the organisations and users is that users want to do something useful, they trust that the organisations give them tasks that are really needed to complete, that the contributions will be put into use, and that anyone with knowledge can contribute, because the knowledge about items can be out there without a user being a registered reader of a particular organisation. Users do not see themselves that much as experimenting for fun, especially without instructions and without others having shared their codes for reproducing the experiments (Objective-Tools). The approach of BL Labs relates more to seeing collections as a mass of data and to Auer's (2013) vision of a digital library, which should not duplicate the physical library's functions of collecting-preserving-making accessible, but rather be a hub of linked open data which enables users to use the data in a new way.

Staff raised questions about serving the needs of very different user communities, and the user communities did expose different attitudes indeed. For instance, some see the whole objective of tagging as helping other researchers and enhancing knowledge, while others prefer to keep their tags private. Also, the same people use catalogues for research and prefer events for personal interests (Community-Objective; Tools). The Archives is struggling with the question to what level of detail to describe and from which aspects of the materials, while the users are claiming that it is very difficult to find specific items unless you know the collection to search from (Objective-Tools). Some people expressed a preference to participate in a public workshop and use physical books, while the security rules of the Library would not allow that free interaction with physical literature (Rules-Tools).

One-sided topics. Some issues raised by the users were not discussed by the staff, most relevant perhaps being encouraging users to interact with the catalogue for a specific task; making the catalogue less impersonal; facilitating faster information retrieval; and limiting the visibility of tags, which could conflict data protection.

From the institutions' point of view, cross-institutional cooperation was mentioned within the GLAM sector, and BL Labs is open to any new half-year experiments, but tighter collaboration with third parties was not mentioned. Also a Wikipedian did not see the memory organisations bringing their expertise to classification, which was mentioned as a bottleneck in Wikipedia.

Many visitors said one thing hesitatingly at the beginning of the interview, but after imagining themselves in the collaborative context of the organisation, they saw tagging as more useful and were more willing to contribute. The users may not yet expect to have an option to collaborate on institutional platforms of memory organisations, although some respondents drew parallels with social platforms like Facebook or Spotify to illustrate what could or how this should be facilitated for users. At the same time, they are not collaborators on the social network sites of the case-study organisations either, but many thanked the researcher for informing or reminding them to look up social media pages or a subject blog by the organisations and said that they would go and check it out.

Other remarks. Given that people mostly would not read guidelines unless questions arise, but they do expect examples of good practice in tagging, then the suggestions could be incorporated into the system in a smart way, popping up when someone is writing a tag. Analysis like that presented in the previous chapter gives sufficient data to work out automatic suggestions. Many users were concerned that it is too resource-extensive if staff verify the tags, but none of them were concerned that verification takes place, as was suggested in Chapter 2 by Matusiak (2006).

If we synthesise some staff and user statements, it may well be that before the shift towards digital has fully happened within the memory organisations and their user communities, machine learning has become so advanced that text is recognised to the extent that tags are not needed and commercial companies are filling in the gaps that different users need.

6.3.1. Conclusion of the Chapter

The thematic analysis revealed that half of the themes discussed around users' participation with collections were directly contributing to discoverability, from the perspectives of both staff and users. This is a significant enough proportion in order to put discoverability into the spotlight of organisational practice as well as future research for studying the relations with different actors in more detail.

There is no direct answer whether to prefer systematic crowdsourcing projects or to open up the whole collection for contributions. But, in either case, users certainly expect clear information about the expectations of the organisation and that their contributions will be put into use, mostly because they saw the objective as helping other people from their subject communities. This in turn may help to bind subject communities more tightly around the institutions. When the specialist users come to help each other, the organisations have direct access to them with endless opportunities to engage them for improving discoverability, visibility, feedback and so on.

This concludes the empirical analysis. The next chapter discusses the contributions of the whole research project.

7. Discussion

First, this chapter summarises the results of the study in a synthesised way by answering the research questions. Detailed description of the results can be found in the discussion section of previous chapters. Secondly, this chapter discusses the findings as making contributions to four aspects: to organisational practice, to methodological innovation, to activity theory, and to future research. Furthermore, the final part of the chapter will point out the limitations of this research project.

7.1. Summary of Results

Before trying to answer the main research question – **how are users' participation and the discoverability of digital collections related? –** the findings are summarised by sub-questions.

Where does user interaction take place?

Preliminary document analysis was carried out on the websites of the two case-study organisations. As regards to online collections-related participation, the analysis revealed that users can interact on many institutional platforms (catalogues, blogs), including platforms of the Library which were specially designed for crowdsourcing (LibCrowds, Georeferencer), as well as in well-known social media sites where the institutions are operating (Twitter, Facebook, Flickr etc.). It was evident that users' participation is different on the platforms which are specially designed for crowdsourcing compared to the platforms where the participation is an on-going additional feature. Sufficient evidence in the literature refers to focused management of crowdsourcing platforms, leading to more coherent user communities who are focused on achieving a specific aim. The current project wanted to learn more about the natural behaviour of users under the condition of not being directed by the specific goals of a project. Five platforms were selected in total for further analysis on the basis of being comparable and close to the phenomenon of user participation in relation to discoverability, i.e. enabling social tagging. These platforms were as follows: the main catalogues, Explore of the Library and Discovery of the Archives;

the specialised catalogue Archives and Manuscripts of the Library; and the Flickr pages of both organisations.

What do the institutions enable the users to do in those media and why?

Document analysis of the 5 platforms and related help pages revealed what was enabled for the users in those mediums and how. All three catalogues enabled social tagging. While in Discovery and in Flickr, anyone can attribute tags upon online registration, the Library had authorised only those users who had been granted a reader's pass or had registered as frequent customers of their document supply service to tag catalogue records. The platforms had different syntax rules for the attribution of multiple tags, which on some occasions turned meaningful tags into useless in case of the misuse of the syntax.

The most popular tags were displayed on the basis of how many times they were attributed, but not alternatively counting the number of users who had attributed the tags. Comparison of the top tags based on their occurrence and based on the number of people who attributed the tags reveals an intriguing dichotomy. In four cases, the most attributed tags refer more to applying computational techniques for tag attribution, mass import of tags from other sources, or to the format of the tagged item. Tags attributed by the most people tend to be more telling in content, but also, in the case of the ,catalogues revealed an individualistic use of tags, e.g. 'to read' or 'granddad.'

Tagging guidelines were more thorough for catalogues and more brief in Flickr. The comprehensive help section in Discovery was found only on the opening page, and was not visible to users who landed directly on the record page. Discovery turned out to be the only platform that did not make the contributions available to other users, as social tags were visible only to the user who attributed them. API was made available only in Flickr, and the BL Labs team promoted its usage in multiple ways.

In order to answer the 'why' part of the research question, the interview method was adopted. Thematic analysis of 4 in-depth interviews and 7 online consultations with members of staff from the respective organisations was conducted.

The organisations referred to being in the middle of a paradigm shift towards a digital mind-set. Several factors of this process concern the platforms in question:

- Fear of losing control over collections remains in the organisations, which sometimes slows down the full application of available technologies. This

includes fear of losing the link to provenance and uncertainty of the new context for collection items.

- Only a small part of the collections is digitised and even less is fully available online. Digitisation is resource-extensive and time-consuming, and the selection of materials or forms of output may be dependent upon agreements with external funders.
- Different working standards within the organisations may make the digitisation outputs or datasets incoherent and therefore not reusable for multiple purposes or for user experiments.
- The image of the organisations in the eyes of users is as a place of physical things, but in the future not only digitised, but also born-digital materials will prevail.
- Similarly to digitisation, it is time-consuming and resource-extensive to develop platforms and deploy new tools. That is why experiments are done on platforms which are not specifically designed for the respective content. For instance, Flickr has been a great experimentation and learning space for the organisations, but staff expressed concerns whether a photo-oriented platform is the most appropriate place for illustrations. Likewise, users had not considered using Flickr for that purpose.
- Sometimes in order to make innovative things happen – and when the institution has no existing procedures, the only option is to trust the participator. The question of trust also relates to internal tolerance of experiments and failures.

Probably similarly to developments in analogue world, initiative in digital is taken and led by enthusiasts among the staff. It contains the risk that after the person leaves the organisation and there is no replacement with similar enthusiasm, the investment of building a participatory community does not show a return.

Under these circumstances, the organisations enabled social tagging in order to experiment and learn more about user behaviour. A contributing factor to launching this experiment was existing technology which had the ready-to-go option of tagging.

How do users participate under those conditions and why?

Tagging data analysis, including over 744,000 observations (i.e. tags attributed on the platforms in total), illustrated how users interact under the conditions set by organisations or platform owners. The findings echoed previous studies by pointing to the small number of active contributors in all platforms.

The Library launched social tagging 5 years before the Archives' user count starts in Discovery, and the Library offers more than twice as many records for tagging than the Archives, but the Library has nearly 23 times fewer engaged users than the Archives and 5 times fewer social tags in its catalogue. Because of this, restricted user authorisation is believed to be the reason for lower participation for the Library.

Is the quantitative difference reflected in qualitative outcomes? This assumption drove the Library to set the restriction, but the data analysis of the current study did not confirm that selected group of participators provided more quality tags than random group of participators. The prevailing tagging behaviour was similar for the catalogues of the two organisations with different practices for user authorisation.

The time factor was recorded only for the Library's catalogues; thus an additional dataset was introduced into the study for tagging data analysis, i.e. EOD Search, a consortial catalogue of pan-European eBooks on Demand Library Network. Tagging in both catalogues of the Library and in EOD Search implied that tagging has not become more popular over time, but rather that it is unstable in terms of total tags but more stable in terms of participating users.

The analysis of the time factor also pointed to the rather surprising finding that not only most users, but also most top taggers make their contributions within a short time-frame – less than 10 days, and in some cases only once. This finding is even more surprising considering the fact that the time-factor was available for the datasets of the Library catalogues, where taggers are mostly registered readers and would thus be expected to relate with the collections longer than casual online users.

It is noteworthy that spam words did not occur in any of the five platforms, but it is also acknowledged that spam filters were applied.

Who are the users?

Thematic analysis of 24 semi-structured interviews with visitors of the organisations (including one in-depth interview with a top tagger of the Library) broadened the understanding of who the users were and why they did or did not

interact with the digital collections. The random sample of visitors came mainly from the academic community, including doctoral and master students.

When the institutions seem to be in the paradigm shift in the scale of analogue-digital, then the users seem to move in the scale of individual-collective:

- Researchers with individualistic goals do not tag at all, because they do not want to reveal their findings before having published in an academic domain which assures them the rightful reference. It can be considered a matter of values and self-determination characterising the individual rather than to collective culture. It may change over time according to general trends of society and scholarly communication, including open science initiatives and the field of digital humanities, which are essentially about openness and stand for ease of research. Not all researchers who were interviewed within this project fell into this categorisation. Many did not mind tagging openly for the sake of their subject community.
- Some respondents had consciously adopted the collective mind-set in regard to many areas, but not necessarily in regard to the collections of memory institutions. They sometimes even feel moral duty, but do not yet participate.
- Significantly, users shared the concern of the institutions to balance the dominant "white voices" in annotations or on Wikipedia. This notion also relates to the wider issue of a digital divide.
- Integrated approaches cross-institutionally and with third parties, and multi-purpose use of the platforms, are addressed by staff and appreciated by users, many of whom had ended up in the catalogue having been referred from elsewhere.

As regards to more detailed expectations and preferences, visitors perceived access to the collection items as an even more burning issue than having discoverable collections. The lack of discoverability was seen more as a problem in connection with the archival catalogue for low-level items.

Users seemed more to appreciate standardisation of tags than acknowledgement for tagging. Respondents agreed that there is no purpose to asking people to tag if the same amount of time is used for checking the tags. Thus a general monitoring of the process, somewhat automated procedures for standardisation, or random checks were

proposed by users. But, most importantly, users were willing to contribute if the value and usefulness were articulated, but not for experiments.

The sample was not representative for making a statistical overview of users but, more importantly for this research approach, it served to provide qualitative information about how users perceive the role of the organisations and their role as regards to participation. They expect organisations to facilitate participation, manage the crowd-sourced content, and provide assurances that it will be put into use. Users themselves fall into three broad groups:

- researchers who do not want to participate before having published in academic domain;
- researchers and other users who feel positive about other users' participation, but it is not yet their behavioural pattern even if they are socially active in other areas;
- researchers and other users who have participated or see themselves doing so in the future to help others with similar interests.

What kinds of relationships exist between the type of organisation and platform?

The differences brought along either by the type of the organisation, the type of the collection, or the type of the platform were monitored along the way and the application of activity theoretical model of activity system completed the answer.

The type of collection seems to influence the tagging outcomes. Archival materials, which are generally unique in content, get more unique tags than published materials in the collections of libraries. The fact that cropped images from public domain books of the Library had no pre-existing metadata might result in them being tagged more often, even if the image discovery may have been more serendipitous in the first place compared to described collections available for tagging.

The type of platform was assumed to play a role in the number of taggers in favour of the social network sites because of their social role to engage new audiences, but this turned out not to be the case. The advantage of Flickr was its API, which was used by just a few people but enabled mass tagging and was used for couple of user experiments: to mark workflows, and to incorporate tags from other sources, e.g. geotags from georeferencing platforms. The content on social network

sites is sometimes believed to be messy and noisy compared to the structured and verified content in catalogues but, in the current case, the individual type of tags like 'to read,', 'check,' etc. were common in catalogues, but not in Flickr.

The distinctly hierarchical nature of archival catalogues may have caused the addition of tags at the wrong level and results in, for example, Discovery having over 100 tags in some records. Mistakes against syntax rules when adding multiple tags at a time and multi-word tags were noticed in all platforms, turning useful multi-word tags into noise. Words within phrases were considered as separate tags, if they depended on a system not putting phrases between quotation marks, or if a comma was used in the middle of the phrase.

The main research question – **how are users' participation and the discoverability of digital collections related? –** implied a hypothesis that user participation is useful for discoverability. The synthesis of the user data and definition of discoverability confirmed this hypothesis.

Technical tags (e.g. 'geo:osmscale=,' 'lefthalf') and personal tags (e.g. 'borrow,' 'granddad') are noise from the respect of visibility. When someone takes a look at the list of such tags next to a record, it does not provide any meaningful additional information. But the definition of discoverability refers to improving the item's ability to be found by appropriate users. Thus staff members or volunteers, who have added technical tags, have improved the discoverability of the items for themselves in order to be able to use the items for other 'appropriate infrastructures' – another condition in the definition of discoverability. Likewise, users who have tagged items with personal keywords like 'granddad' or 'to read' have improved discoverability of the items for themselves, meaning they can easily return to these records.

Similarly, the item's quality is changed by other tags which were meaningful and potentially used as search terms by a broader user community. The definition of discoverability is not dependent on the level of usage, i.e. how many times the added element of metadata helped someone to discover the item. Therefore, the relation between users' participation and discoverability was determined to be causal, and all tags analysed within this study were defined as contributing to discoverability.

Moreover, according to the findings of this study, the users' participation contributes to discoverability in four modes:

a) in invisible mode – only taggers can see their own tags, other users cannot see or use the tags as search terms;

b) in individual mode – attributed tags are public, but they are meaningful only for the taggers, e.g. marking up an item for returning to it later or adding specific technical tags for institutional workflows;

c) in restricted mode – everyone can use tags as search terms, but authentication for adding tags is restricted;

d) in public mode – tags are searchable for all and everyone can add tags upon signing up.

7.2. Contribution to Organisational Practice

According to Simon (2010), social tagging, which has been analysed in the examples of the case-study organisations, is mostly a contributory act – a simple way for a user to interact and see their contribution immediately.

Improving discoverability seems to remain a supplemental contribution for the institutions in general, but a necessary contribution for crowdsourcing projects like Playbills or LibCrowds. BL Labs competitions can be categorised as co-creative. Trust is an inevitable precondition to make co-creative projects happen – significantly discussed by both Simon and the manager of the BL Labs according to their experience. There is room for an educational approach in which teaching skills is slightly more important, although related, than the contribution itself. An educational approach was something also agreed on by visitors when asked whether they would join on-site or online workshops for increasing their digital skills in relation to collections. Simon also draws our attention to the need to scaffold participatory experiences in order to make them useful, not merely to submit a generic request or offer an opportunity to share a story or use a tool. An educational approach is also addressed by the notion of the digital divide by both, institutions and users: potential user groups with missing knowledge about the collections might need to be engaged by an educational approach.

The advantage of linear on-going engagement over crowdsourcing projects may rely also in risk aversion. As noted by Simon (2010), when a contributory project relies on visitors' contributions to succeed, there is a risk that visitors do not participate and the need for high institutional investment to avoid the failure. In the

catalogues, the feature of tagging was simply an existing feature that was switched on without institutional investment.

If we put Simon's concept of participatory organisations where visitors can create, share, and connect with each other around content into the context of social tagging, then there is a long way to go towards sharing and connecting. Even if particular users are hesitant to share because of the fear of losing credit for their findings, the interviews with visitors brought out the direction of thinking towards social – an experience turning from individual into social, as explained by the five stages stated by Simon. That would also *a posteriori* justify the choice of Engeström's approach to activity theory.

If collective action takes place online and tools are provided to connect individuals, it can be argued that this is another doomed attempt at "we build and they come," unless, a) it is the visitors who bring the social into the agenda, b) staff see the linking of content and compatibility of systems institutionally and occasionally also cross-institutionally. At the moment, the community of users tends to be formed of individuals who do not collaborate. But offline, the collaboration takes place in the form of events (e.g. thematic workshops, hackathons, describathons etc.)m transforming individual practices into collaborative ones and fostering interactions based not (only) around topics, but also around computational skills – a promising combination for improving discoverability.

This research project aimed to have a consultancy value for cultural organisations. Considerations were presented in chapters 5 and 6. The overall suggestions for consideration to the organisations include:

- Raising awareness of tagging among users, especially by articulating the value of the tags, once clarified within the organisations. Users are willing to participate *if* they are convinced that other community members can use their contributions. While not evident from the document analysis, but revealed in the interview phase, the tags in Discovery are visible only to the tagger who attributed them and remain invisible to other users.
- Communicating the objectives internally and synchronising working standards accordingly would also help avoid scanning the same items multiple times, and processing the same files instead for using the data for several purposes.

- Some interviewees among the staff saw a problem in the small number of contributors. But Simon sees it as natural so that the crowdsourced content can be managed, maintained, and useful to spectators. And perhaps the digital mind-set should be focused on on-going integrated solutions enabling participation and welcoming newcomers who bring along new ideas and leave behind traces to suggest further changes in institutional systems.
- Launching an API for the system would not only increase the number of tags significantly, but would also engage new participating audiences who are not related to the domain of the content. It would also pave the way to applying a more educational model of participation, which was welcomed by the visitors.
- The content analysis of tags suggests that tags are contributing to discoverability, but more sophisticated ways are needed to expose the different types of tags in order to decrease noise concerning the visibility of tags.
- Cooperation between different types of institutions is encouraged, especially due to approaches to information retrieval in the digital humanities where access to the items is not claimed one by one, but entire collections are processed instead.
- Following the example of the Flickr pages and the catalogue of the National Archives, The British Library is encouraged to authorise all online users to register on the catalogue and therefore be able to tag.
- Derived from users' motivation to tag to help others, the National Archives is encouraged to make social tags searchable.
- Recording time of tag attribution would also be an asset because it gives valuable information about the user behaviour, as was the case with the catalogues of the Library.
- The research findings on tags give clear suggestions what to include into the guidelines for users. After interviews with visitors, it became clear that these recommendations should be incorporated into the system, linked from every item, and suggested automatically while typing a tag rather than presented in a section of help articles on the website. Analysis like

that presented in section 5.2. gives sufficient data to work out the automated suggestions. The findings relate to the notion that nearly every person in the UK has had some experience of Google, Amazon, eBay or Yahoo, and many of them are now beginning to demand that these experiences are replicated in other areas (Baker 2005). Additionally, different syntax rules for multiple tags or multi-word tags seem confusing and more work is needed to find the best solution.

- If 'top tags' are presented for information or as inspiration for users, then it would be more representative and telling in terms of content to display the tags which are attributed by more people, not most attributed in total and possibly by just a few or even by a single person.
- Stepping into a dialogue with top taggers has led to many collaborations between the Library and some volunteers, which may have extended the period of being active for these people.

The issue of adding tags in private modes was raised for two main reasons. First, researchers with individualistic goals do not tag at all if only public modes of tags are facilitated, because they do not want to reveal their findings before having published it in an academic domain. It can be considered as a matter of values and self-determination to an individual rather than collective culture, which may change over time according to general trends of both society and scholarly communication. Secondly, users and staff mark up records or items in order to have an easy option to return to them later on, or staff use specific tags for other purposes. This category of tags is meaningless for the rest of the users. If we look at all tags, this category is noise, but only from the point of view of visibility– if tags are visible, some of them may be considered noise. But this is not how discoverability works. Leaning on the definition of discoverability, appropriate users find the items from appropriate infrastructure through the item's quality of being found. It is the item, not whole collection of tags, which is the focus of discoverability. Thus if a user looks for a map by typing in a search term 'map,' all that matters is that one of the keywords or tags is 'map' and that they are not distracted by all other terms, whether meaningful or meaningless. So from discoverability's point of view, the private mode is not a necessity.

"Doing big data" was brought to the agenda by experiments in BL Labs. In principle, it is in line with the definition of big data as a cultural, technological, and scholarly phenomenon that rests on the interplay of (1) technology: maximizing computation power and algorithmic accuracy to gather, analyse, link, and compare large data sets; (2) analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims; and (3) mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy (boyd and Crawford 2012). But it does not correspond to the five V-s, which are seen as characteristics of big data at the same scale as in other sectors: volume (scale of data), velocity (analysis of streaming data), variety (different forms of data), value, and veracity (uncertainty of data). It can be that it never will and does not have to. Hitzler and Janowicz (2013) claim that different (scientific) disciplines highlight certain dimensions and neglect others, e.g. super computing is mostly interested in volume, the internet of things in velocity, the semantic web in variety, and the social sciences and humanities in value and veracity. In order to put digital collections on other agendas than social sciences or humanities, the cycle must grow and continue. If users report back what they have done with the original data, make available their outcomes as processed datasets, and mechanisms are created to follow how these are used, it might make a case for big data. But the prerequisite is to license openly the organisation's data as was mentioned and recently done by the British Library.

This project revealed several aspects which are not common in the literature reporting the crowdsourcing outcomes. Given that staff and users noted that the future could be much more cross-institutional, it then becomes relevant to provide data and articulate on these aspects, which are often taken as self-evident within one type of memory organisation and are therefore not described. For instance, the correlation analysis of the total and unique tags and the number of tagged items reflected the perception of users about the distinct nature of the published collections of library and unique items of archival collections.

Furthermore, the comparative approach including two different types of memory institutions revealed the direction of travel towards archival practice. The Library is experimenting by putting into spotlight the items belonging to a volume – images from public domain books in Flickr or playbills bound into a single volume and

catalogued as a volume by the Library. Emerging technology could enable users to submit queries about items of interest across volumes if there is correspondingly structured data. And considering the rise of digital humanities, there is no doubt of the interest for users in doing so. The outcomes of social interaction can be input for organisations to create structures for ingesting descriptive data. Given that the number of illustrations per public domain book in the Library's Flickr collection counted up to 903 images in the case of a cyclopaedia, it could lead to the case of big data, if made discoverable, accessible, put into use by users, use monitored by organisations.

Some findings, which were published during this work (Mets and Kippar 2017) and simultaneously discussed with the case-study organisations, were already put into effect by the organisations before completing this dissertation. Likewise, bringing into the spotlight the bare fact of the monumental number of 18,000 users of the Archives who have added a tag in Discovery was claimed to bring the topic of social tagging back on to the agenda of the organisation.

Chapter 6 and Annex 1 with the direct quotes by the staff represents both a contribution to the literature for future researchers, as well as a source for other memory organisations to learn from the experiences of the case-study organisations.

7.3. Contribution to Methodology

The approach of running thorough document analysis prior to user data analysis gave essential context to interpret user behaviour. For instance, there is not less interest in tagging the Library's catalogue records, but the feature to register online in order to sign in is just not available.

As importantly, the document analysis and special inquiries enable to compare the data in long term if needed. The platforms change and so do the institutional affordances concerning rules, tools etc. That is why we have to provide the context in order to understand the direct impact of the tools and rules on user behaviour.

That is also relevant for institutions internally. Many interviewees referred to the change of employees and knowledge of the change of rules may disappear with them. It would weaken the future studies dramatically if user data would be analysed as operating under the same conditions. It remains to be explored if there will be more taggers once the Library opens up the option of signing up online to everyone, or whether the users of Discovery will be happier with the search results once the tags are included in the search.

The presentation of the study enables following the narrative of activity systems in the core part of the document, and referring to the annex for scrutiny in regards to the interpretation of the interviews with staff.

The main research question about the relation between users' participation and discoverability is addressed throughout the study. The comparative approach illustrated the similarity of the two case-study institutions as regards to the actors contributing to discoverability. Objective, Tools and Outcomes stand out for both institutions and users as the actors most related to discoverability. Additionally, the user community is a relevant actor for institutions. Therefore, more thorough inquiry about discoverability may focus around these topics. Significantly, all themes for Tools mentioned by staff were evaluated as contributing to discoverability. This would additionally justify applying the human-artifact model.

7.4. Contribution to Activity Theory

The central artefact in document and user data analysis was the platform, which was an institutional affordance to engage users. The analysis of interviews placed these mediating tools into the context of other actors, studied their relations, and determined their relevance to discoverability.

The views of the interviewees were categorised roughly according to the seven actors of the activity system, and some more frequently mentioned topics in the interplay of the actors were categorised as such. The biggest challenge to categorise an actor concerned the collections of the two institutions. Keeping an institutional activity system in mind, are collections a Subject or a Tool?

Neither of the options felt right, but both actors are most relevant to collections. Collections as Subject refers to the notion that they are a characteristic of the organisation. The type of the collection seemed to have impact on outcomes. For instance, more unique tags are attributed to archival than library collections. So collections could be considered as Subject in addition to the staff of the organisations, unless collections do not have the agency on their own. Similarly Balnaves and Willson (2011) have claimed that information has no agency in its own right. But a Subject must have an agency in order to take action towards the Object.

That would lead us to classify collections as Tools. That does not feel appropriate either, because neither organisations nor users can take actions, which are mediated by collections. The actions are mediated by the systems of catalogues or social

network sites, which present the collections (either images or the collection of records). The community takes action on those platforms, i.e. acts by being mediated by respective tools. Analysis of documents, users' tagging data, and interviews illustrated that different tools can be applied to the collections either one-by-one or as a system of tools which communicates within itself and may also occur as the Internet of Things (e.g. tags added by Flickr friendly robots to the images, based on the text, which was linked to from Flickr next to the image).

So the collections themselves cannot be considered as a Subject or Tools: they sit in the middle and are rather a resource – an actor that the subject needs as an input for running a tool. In addition to collections, other resources also needed by the subject for using the tools were mentioned by the staff: time, funding, human work power, including knowledge (see Subject-Tools in section 6.1.3. Tools). The visitors pointed out these resources as well. Thus there is a case in this research to place an actor Resources on the scheme of activity system.

This development of Engeström's conceptualisation is visualised in Figure 7.1. In order to keep the visualisation useful and not overcrowd it by naming all actors as equal, the initial central actors are made distinct.

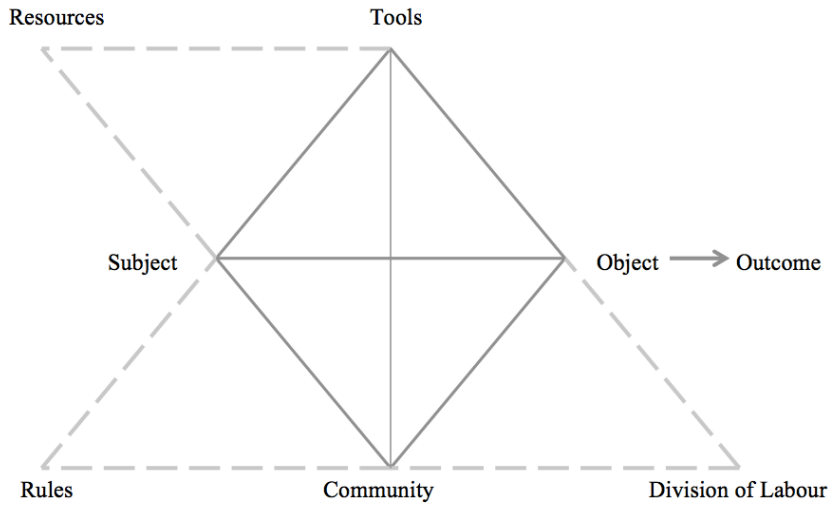

Figure 7.1. Development of Engeström's framework.

In addition to the proposal of the new actor, the contribution to Activity theory is testing the Engeström's framework as an analytical tool for an abstract activity: users' participation. It enabled working with large quantities of qualitative data from the in-depth interviews in a very systematic way. The method also revealed an important

aspect that was not addressed: the strong similarity between the two organisations belonging to different types of memory institutions.

7.5. Limitations and Contribution to Future Research

The approach of the research project was chosen to be mostly qualitative and explorative. Therefore, it ends up with suggestions rather than grounded conclusions and provides open questions for other similar studies.

Qualitative enquiries could run similar comparative studies to see if the outcomes for a library and an archive in another country context are similar. What might be other affecting factors of outcomes and views on applying activity theory? The exploration of the context can also be taken to a more detailed level by structural equation modelling and trying to determine the structural relationship between the actors and/or related concepts to digital collections, like visibility, discoverability, usability etc. By analysing related activity systems (like in Annex 2), it would be useful to know which features are the cause of others, and are they mutually beneficial or are they needed for the emergence of the next features? That knowledge could help the organisations to prioritise their activities.

A separate research question can be the one seemingly burning for both organisations and users: what is the value of opening up collections for reuse and remix, and what is the value of user contributions? An answer informed by activity theory is evolution, as explained in the conclusions of this dissertation, but application of different analytical tools may provide more distinct answers.

This research included only tags as indicators to users' participation, because tags were believed to be directly contributing to discoverability. The National Archives referred to the valuable nature of longer and more descriptive user annotations in previously run wiki Your Archives. This kind of interaction is not currently available in the case-study organisations of this study, but it would make a case to include user comments in Flickr for a comparative content analysis.

Alternatively, running social network analysis on the followers of the Flickr pages of the two institutions would help to learn more about the taggers and non-taggers. How similar are the people in the two groups by their interaction and the pages they follow?

Contributors are valuable and unique. If we focus on digital collections, then one possible further method to understand the users is netnography. It became evident that

there is a story behind every dedicated user: why, how, when, with whom etc. they take action regarding the digital resources of memory institutions. Netnography, possibly mixed with other methods, would make a significant contribution to understanding that particular community of people and the deeper connection between their contributions and their stories of life.

Also it is believed that the same study can be continued for forming the networks of activity systems. The activity networks between institutions, research labs, private sector, collaborators, users, and spectators may shed light on the question of value in addition to mapping workflows and conflicts of interests.

More in-depth insight into the interaction of users on the platforms would fall into the human-computer interaction (HCI) domain. It would then be useful to apply the human-artefact model (see section 3.1). Information for creating personas can be extracted from the user data in information systems, being complemented by the interview findings. Information for creating techsonas is mostly provided by the document analysis of the interviews, complemented by the interview findings.

Finally, it could also be relevant to abandon the qualitative approach and study folksonomies with quantitative methods. Separate inquiries for libraries and archives can confirm if the results of a quantitative study match the qualitative ones.

These suggestions imply the limitations of this work. Most importantly, the results of two case studies cannot be generalised to the whole sector. Comparative research is needed. Moreover, some findings, like the occurrence of taggers in a short timeframe, is based only on one case and an additional external case to this study due to absence of time-factor in the other case study of this project. The sample of interviewed users of the organisations represents mostly non-taggers. Additional study is needed to confirm that online taggers from any location also have similar preferences. In order to raise the consultancy value, there is a need to look beyond the definition of discoverability and compare the data about social tags to the data about users' actual search terms on those platforms.

The method of using Engeström's framework as an analytical tool leads to the subjective categorisation of the mentioned themes under the defined actors. This is overcome by presenting all themes per actor at a time, enabling the comparison of the organisations along the way and eventually creating the generic activity systems including information about both organisations.

8. Conclusion

The aim. The current research project aimed to investigate the relation between users' participation and the discoverability of content in digital collections. The aim was derived from the reviews, which expressed the need for methods bringing a social science perspective into digital library research and interpreting the findings back on a societal level. The project also had the enhanced ambition of contributing to comparative research across different types of memory institutions, and to suggest practice for improving user engagement.

The approach. The case study setting was chosen for answering the research question, how are users' participation and the discoverability of digital collections related? Two national institutions were included: the British Library and the National Archives of the United Kingdom. Five of their platforms were selected for analysis of the closest phenomenon to users' participation in relation to discoverability: social tagging. The following platforms were considered: the main catalogues, Explore and Discovery, of each organisation; the specialised catalogue Archives and Manuscripts of the British Library; and the Flickr pages of both organisations.

The findings. The research provides two groups of findings:

First, the findings illustrate a paradigm shift: organisations on an analogue-digital scale, and users on an individual-collaborative scale. From the organisations' point of view, when the user engagement and crowdsourcing projects date far back, organisations are now experimenting with emerging technologies for incorporating the interactive mode into everyday processes. The staff experience many challenges along the way, from digitisation peculiarities, to trust, to the user community using the digital content, and to the tolerance of risk in receiving unexpected outcomes. Institutional practice regarding the platforms seems to follow the trend of most social web initiatives of being in "perpetual beta" (Simon 2010) – not fully designed before engaging users. But what *has* changed is that, if Simon considered it a "particularly radical case" to share the digital collection content and software coding openly with external programmers, staff members, who were interviewed and related to the Flickr

initiatives, now find it essential rather than radical. Thus the paradigm shift is happening, albeit slowly.

From the users' point of view, integrated approaches cross-institutionally and with third parties, and multi-purpose use of the platforms are addressed by the staff and appreciated by the users, many of whom had ended up at the catalogue by being referred from elsewhere. Researchers, who are the majority of users of the case study institutions, fall into three groups according to this study: a) those, who follow the individualistic approach to research and prefer not to reveal their findings before having published them first; b) those who in principle agree with sharing knowledge, but who do not have the habit of participating in this domain even if they participate in other areas; c) those who do participate already for helping others with similar interests. The rise of open science and advances in the digital humanities give reason to believe that the shift from individual to collective is slowly taking place in the sharing of knowledge via users' participation regarding the collections of memory institutions.

Secondly, social tags analysed in this study contribute to discoverability through its definition of being the item's quality of being found by appropriate users through appropriate infrastructure. As a result of this research, four modes were detected, in which social tagging contributes to discoverability:

a) in invisible mode – other users of the platform cannot use the tags as search terms, only taggers can see their own tags;

b) in individual mode – attributed tags are meaningful only for the taggers, e.g. marking up an item for returning to it later or adding specific technical tags for institutional workflows;

c) in restricted mode – everyone can use tags as search terms, but authentication for adding tags is restricted;

d) in public mode – tags are searchable for all and everyone can add tags upon signing up.

The type of collection seems to influence the tagging outcomes: archival materials, which are generally unique in content, get more unique tags than the published materials in the collections of libraries. The type of platform was assumed to play a role in the number of taggers in favour of the social network sites because of their social role to engage new audiences, but that turned out not to be the case. The advantage of Flickr was its API, which was used by just a few people but enabled

mass tagging. The content on social network sites is sometimes believed to be messy and noisy compared to a structured and verified content in catalogues, but individualistic type of tags, meaningless to others were found in both catalogues but not in Flickr. It was not found out whether such tags were added because the attributor did not realise they would become public or because the user simply did not care. The distinctly hierarchical nature of the archival catalogue may have caused adding tags to the wrong level item. Mistakes against syntactical rules for adding multiple tags at a time and multi-word tags were noticed in all platforms, turning useful multi-word tags into noise (e.g. counting first and last names as separate tags).

Implications. The further implications of the study rest on detecting the development or change both in organisational and user behaviour. Activity theory, which is used as a conceptual framework for this study, aims to understand how human activity unfolds over time. This thesis was a snapshot to capture the pains of both – organisations and users changing their behaviour, including rules and division of labour, for operating under the changing paradigm. At the same time, we witnessed the creativity of both, which implies behavioural change – to development compared to traditional library or archival practices. Taking into account the history of the practices, what is built upon it, and how is it used refers to a cycle that is essential to activity theoretic approach.

From that respect, it is acceptable if users turn out to be creative and use the affordances facilitated by institutions in unexpected ways. Just like primates saw rocks and wood as useful tools, some taggers add tags like 'to read' or 'granddad' because those are useful for them. Organisations are considering to introduce the function of private tags for such behaviour. But the study illustrated that from discoverability's point of view it is not a necessity. Furthermore, leaning on an activity theoretical approach, these kind of examples should not be seen as failure, malfunction or misuse, but as opportunities on which to build. Users' participation with digital collections, either with library or archival collections, is not only valuable for discoverability – the common good – but also contributes to the greater goal of developing the services in collaboration with users.

The collaborators are not only the taggers. The other users of the platforms, the spectators are just as useful collaborators as smart the organisation behind the system is. Thus the small proportion of taggers out of all users of the platform should not diminish the value of the contributions or the opportunity of users' participation in the

first place. The study also found that the contributors are not here to stay, they participate during a very short time period. But according to the essence of collective culture, the on-going occurrence of the phenomenon of participation is what matters, not who specifically are contributing.

Weinberger (2010) defined information overload as a cultural phenomenon. Together with the evolution of technology, which enables us to apply filters and be smart in a different way, Weinberger's notion may imply to the raise of tolerance of noise in users. Noise or uncontrolled content is something that makes the organisations to doubt in the value of users' participation and impedes them to scale the experiments.

While organisations struggle with the question of the value of users' participation – and users wait to take action until its value is articulated to them, the right time to act goes by. It may well be that before the shift to a digital mind-set has fully taken place in organisations, the machine learning has become so advanced that text is recognised to the level that social tags are not needed and commercial companies are filling in the gaps to serve different user groups according to their specialised needs.

Collective culture is not a project. Therefore this study concludes with an encouraging notion: to use emerging technologies in favour of development supporting societal change without necessarily knowing the destination.

Future research. It is acknowledged that the approach of the research is not representative, but can be a solid basis for future studies to do the following: a) to compare the trends between libraries and archives, and see, if the trends stay; furthermore, this study unwittingly exposed the similarity of the two types of organisations – asking whether this could be a meaningful direction of travel for them; b) to compensate for the limitations of this study, take a step forward, and determine the usefulness of social tags as search terms; c) to repeat the unique methodology of using the activity system as an analytical tool for mapping the context around the technology-mediated phenomenon of users' participation. It may confirm or deny the placement of an additional actor *Resources* to the scheme of activity system.

Bibliography

Aleksić-Maslać, K., Magzan, M. and Jurić, V. (2009). Social phenomenon of Community on Online Learning: Digital Interaction and Collaborative Learning Experience. *WSEAS Transactions on Information Science and Applications,* vol. 6(8), pp. 1423–1432.

Auer, S. (2013). Digital Scholarship and Digital Libraries. *17th International Conference on Theory and Practice of Digital Libraries*. 22.–26.09.2013. Valletta, Malta.

Baker, D. (2005). Supporting the next generation of applications for delivering rich, library content and services. A White Paper. *The Talis Platform*. https://www.immagic.com/eLibrary/ARCHIVES/GENERAL/TALIS_UK/T051114B.pdf. Accessed 10.11.2017.

Balnaves, M. and Willson, M. A. (2011). *A New Theory of Information & the Internet: Public Sphere meets Protocol*. Series: Digital Formations, vol. 66. New York [etc.]: Peter Lang.

Boast, R., Bravo, M. and Srinivasan, R. (2007). Return to Babel: Emergent Diversity, Digital Resources, and Local Knowledge. *The Information Society: An International Journal,* vol. 23(5), pp. 395–403.

Bødker, S. and Klokmose, C.N. (2011). The Human–Artifact Model: An Activity Theoretical Approach to Artifact Ecologies. *Human-Computer Interaction*, vol. 26, pp. 315–371.

Bødker, S., and Klokmose, C. N. (2012). Preparing students for (inter)-action with activity theory. *International Journal of Design*, vol. 6(3), pp. 99–111.

Bødker, S., and Klokmose, C. N. (2013). From Persona to Techsona. IN: P. Kotzé et al. (Eds.): *Human-Computer Interaction – INTERACT 2013*. Series: Lecture Notes in Computer Science. Vol. 8120, pp. 342–349. IFIP International Federation for Information Processing.

Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., and Wilderman, C. C. (2009). Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. *A CAISE Inquiry Group Report.* Washington, D.C.: Center for Advancement of Informal Science Education (CAISE). http://www.birds.cornell.edu/citscitoolkit/publications/CAISE-PPSR-report-2009.pdf. Accessed 24.01.2017.

Borgman, C. L. (2010). *Scholarship in the digital age: information, infrastructure and the internet*. Cambridge (Mass.): London: The MIT press.

boyd, d. and Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, vol. 15(5), pp. 662–679, doi: 10.1080/1369118X.2012.678878.

Brazier, C. (2016). The British Library and its international collections. Paper presented at: *IFLA WLIC 2016 – Columbus, OH – Connections. Collaboration. Community*. http://library.ifla.org/1444. Accessed 8.06.2016.

The British Library. https://www.bl.uk. Accessed 13.11. 2017.

The British Library. Get a Reader Pass. http://www.bl.uk/help/how-to-get-a-reader-pass. Accessed 24.01.2017.

The British Library. Document Supply Services. http://www.bl.uk/reshelp/atyourdesk/docsupply/help/register/regularcustomers/index.html. Accessed 24.01.2017.

The British Library. Sound map – UK Sound Map. http://sounds.bl.uk/sound-maps/uk-soundmap. Accessed 12.01.2016.

The British Library. (2014a). Explore the British Library. http://www.bl.uk/catalogues/search/pdf/ tags.pdf

The British Library. (2014b). Explore the British Library. http://www.bl.uk/catalogues/search/pdf/ notes.pdf

The British Library. (2015). *Annual Report and Accounts 2014/15*. UK: Williams Lea Group on behalf of the Controller of Her Majesty's Stationery Office. http://www.bl.uk/aboutus/annrep/2014to2015/annual-report2014-15.pdf. Accessed 8.06.2016.

Calhoun, K. (2014). *Exploring Digital Libraries. Foundations, practice, prospects*. London: Facet Publishing.

Carletti, L., Giannachi, G., Price, D., McAuley, D. and Benford, S. (2013). Digital humanities and crowdsourcing: an exploration. MW2013: Museums and the Web, 2013-04-17 – 2013-04-20, Portland, OR. http://hdl.handle.net/10871/17763.

Causer, T. and Terras, M. (2017). Many hands make light work. many hands together make merry work: transcribe bentham and crowdsourcing manuscript collections. In: Ridge, M. (ed.) *Crowdsourcing Our Cultural Heritage*. Routledge, London.

Chan, S. (2007). Tagging and searching – serendipity and museum collection databases. In: Trant, J., Bearman, D. (eds.) *Museums and the Web 2007: Proceedings, Toronto, Archives and Museum Informatics*. http://www.archimuse.com/mw2007/papers/chan/chan. html. Accessed 24.03.2017.

Daly, E. K., Ballantyne, N. (2009). Ensuring the discoverability of digital images for social work education: an online tagging survey to test controlled vocabularies. Webology 6(2), Article 69. http://www.webology.org/2009/v6n2/a69.html.

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.

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 on 18.11.2016.

Dutton, W. H. (2013). The Internet and Democratic Accountability. The Rise of the Fifth Estate. *Frontiers in New Media Research.* New York: Routledge, pp. 39–54.

Engeström, Y. (1990). *Learning, Working, and Imagining: Twelve Studies in Activity Theory*. Helsinki: Orienta-Konsultit Oy.

European Commission. (2017). *Digital Economy and Society Index 2017 - Estonia*. ec.europa.eu/newsroom/document.cfm?doc_id=43003. Accessed 13.11.2017.

Feynman, R. P. (1963). The Meaning of It All: Thoughts of a Citizen-Scientist, http://95.76.157.166/astroclub/biblioteca_online/The%20Meaning%20Of%20It%20All%20-%20Feynman.pdf. Accessed 5.02.2015.

Feynman, R.P. (1974). Cargo Cult Science. http://calteches.library.caltech.edu/51/2/CargoCult.pdf. Accessed 5.02.2015.

Flickr (a). http://help.yahoo.com/kb/flickr/tag-keywords-flickr-sln7455.html. Accessed 24.01.2017.

Flickr (b). https://www.flickr.com/services/apps/about. Accessed 24.01.2017.

Fullerton, L., and Rarey, M. (2012). Virtual Materiality: Collectors and Collection in the Brazilian Music Blogosphere. *Communication, Culture & Critique,* vol. 5, pp. 1–19.

Galloway, E., DellaCorte, C. (2014). Increasing the Discoverability of Digital Collections Using Wikipedia: The Pitt Experience. *Pennsylvania Libraries: Research & Practice,* vol. 2(1), 84-96. doi:http://dx.doi.org/10.5195/palrap.2014.60

Grannum, G. (2011). Harnessing user knowledge: the national archives' your archives Wiki. In: Theimered, K. (ed.) *A Different Kind of Web: New Connections Between Archives and Our Users*. Society of American Archivists, Chicago.

Geismar, H. (2012). *Digital anthropology*. London; New York: Berg.

Hadnagy, C. (2011). *Social engineering: the art of human hacking*. Indianapolis (Ind.): Wiley.

Haythornthwaite, C. (2001). The Internet in Everyday Life. *American Behavioral Scientist*, vol. 45(3), pp. 363–382.

Higgins, S. (2011). Digital curation: the emergence of a new discipline. *International Journal of Digital Curation,* vol. 6(2), pp. 78–88.

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(3), pp. 246–259. DOI: 10.1002/(SICI)1097-4571(2000)51:3<246::AID-ASI4>3.0.CO;2-6.

Hitzler, P. and Janowicz, K. (2013). Linked Data, Big Data, and the 4th Paradigm. *Semantic web,* vol. 4(3), pp. 233–235.

Kani-Zabihi, E., Ghinea, G., Chen, S. Y. (2006). Digital libraries: what do users want? *Online Information Review*, vol. 30(4), pp. 395–412. DOI: http://dx.doi.org/10.1108/14684520610686292.

Kaptelinin, V., Nardi, B. (2006/2009). *Acting with Technology: Activity Theory and Interaction Design*. The MIT Press, Cambridge.

King, G., Keohane, R.O. and Verba, S. (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton NJ, Princeton University Press.

Kowalczyk, S., Shankar, K. (2011). Data Sharing in the Sciences. *Annual Review of Information Science and Technology*, vol. 45(1), pp. 247–294. doi:10.1002/aris.2011.1440450113.

Kreijns, K., Kirschner, P. A. and Wim J. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Computers in Human Behavior*, vol 19(3), pp. 335–353, https://doi.org/10.1016/S0747-5632(02)00057-2.

Kuutti, K. (1996). Activity Theory as a Potential Framework for Human-Computer Interaction Research. In B. Nardi (ed.), *Context and Consciousness: Activity Theory and Human–Computer Interaction*, pp. 17–44. Cambridge, Mass.: MIT Press.

Lee, F. L. F., Leung, L., Qiu, J. L., Chu, D. S. C. (2013). Introduction: Challenges for New Media Research. *Frontiers in New Media Research.* New York: Routledge, pp. 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:10.1108/10650750610706998.

Macgregor, G., McCulloch, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, vol. 55(5), pp. 291–300.

McMartin, F. (2006). MERLOT: A Model for User Involvement in Digital Library Design and Implementation. *Journal Of Digital Information, 5*(3). Retrieved from https://journals.tdl.org/jodi/index.php/jodi/article/view/143 on 7.12.2015.

Mets, Õ., Gstrein, S., Gründhammer, V. (2014). Increasing the visibility of library records via consortial search engine. *Proceedings of the 14th ACM/IEEE-CS Joint*

Conference on Digital Libraries, pp. 169–172. IEEE Press, Piscataway. doi:10.1109/JCDL.2014. 6970164.

Mets, Õ. and Kippar, J. (2017). Social Tagging: Implications from Studying User Behavior and Institutional Practice. In: Kamps J., Tsakonas G., Manolopoulos Y., Iliadis L., Karydis I. (eds) *Research and Advanced Technology for Digital Libraries. TPDL 2017. Lecture Notes in Computer Science*, vol. 10450. Springer, Cham. doi: 10.1007/978-3-319-67008-9_33.

Miller, D. and Horst, H. A. (2012). The Digital and the Human. In: Horst, H. A. and Miller, D. (eds.) *Digital Anthropology*. London, New York: Berg Publications, pp. 3–35.

Recker, M. M., Dorward, J. and Nelson. L. M. (2004). Discovery and Use of Online Learning Resources: Case Study Findings. *Journal of Educational Technology & Society*, *7*(2), 93–104.

Moffat, M. (2006). Marketing with Metadata – How Metadata Can Increase Exposure and Visibility of Online Content. *New Review of Information Networking*, vol. 12(1-2), pp. 23–40. doi:10.1080/13614570601133039.

Nardi, B. (1996). Activity Theory and Human–Computer Interaction. In B. Nardi (ed.), *Context and Consciousness: Activity Theory and Human–Computer Interaction*, pp. 7–16. Cambridge, Mass.: MIT Press.

The National Archives. How to tag records. http://discovery.nationalarchives.gov.uk/tags/index/howtotag. Accessed 24.01.2017.

The National Archives. Our history. http://www.nationalarchives.gov.uk/about/our-role/what-we-do/our-history. Accessed 2.11.2017.

The National Archives. Prisoner 4100. https://www.flickr.com/photos/nationalarchives/2978689272/#comment72157623848519008. Accessed 2.11.2017.

The National Archives. Social media use. http://nationalarchives.gov.uk/about/get-involved/social-media. Accessed 24.01.2017.

The National Archives. *Volunteering at The National Archives: The National Archives' approach to user participation*. Retrieved from http://www.nationalarchives.gov.uk/about on 2.11.2017.

The National Archives. (2017). *Archives Inspire: The National Archives plans and priorities 2015−19*. Retrieved from http://www.nationalarchives.gov.uk/about on 2.11.2017.

O'steen, B. (2013). A million first steps. *Digital scholarship blog*. http://blogs.bl.uk/digital-scholarship/2013/12/a-million-first-steps.html. Accessed 24.01.2017.

O'steen, B. (2016). BL Flickr image dataset: User Submitted Tags (till March 2016). Figshare. doi:10.6084/m9.figshare.3126481.v1. https://figshare.com/articles/BL_Flickr_image_dataset_User_Submitted_Tags_til_March_2016_/3126481. Accessed 15.10.2016.

Okasha, S. (2002). *Philosophy of Science: A Very Short Introduction*, Oxford, Oxford University Press.

Oomen, J., Aroyo, L. (2011). Crowdsourcing in the cultural heritage domain: opportunities and challenges. In: *Proceedings of the 5th International Conference on Communities and Technologies (C&T 2011)*, pp. 138–149. ACM, New York. doi:http://dx.doi.org/10. 1145/2103354.2103373.

R Core Team. (2016). R. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

Ridge, M. (ed.) (2017). *Crowdsourcing Our Cultural Heritage*. Series: Digital Research in the Arts and Humanities. London; New York: Routledge.

Simon, N. (2010). *The Participatory Museum*. Museum 2.0.

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). http://collaborativelibrarianship.org/index.php/jocl/article/view/240/181. Accessed 18.11.2016.

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.

Springer, M., Dulabahn, B., Michel, P., Natanson, B., Reser, D., Woodward, D. and Zinkham, H. (2008). *For the Common Good: The Library of Congress Flickr Pilot Project, Final Report*. http://www.loc.gov/rr/print/flickr_report_final.pdf. Accessed 8.06.2016.

Szajewski, M. (2013). Using Wikipedia to Enhance the Visibility of Digitized Archival Assets. *D-Lib Magazine,* vol. 19(3/4). doi:10.1045/march2013-szajewski.

Trant, J. (2009). Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information,* vol. 10(1). Retrieved from https://journals.tdl.org/jodi/index.php/jodi/article/view/269.

Trant, J., Wyman, B. (2006). Investigating social tagging and folksonomy in art museums with steve. museum. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*.

Tredinnick, L. (2008). *Digital information culture: the individual and society in the digital age*. Oxford, Chandos.

Van House, N. A., Bishop, A. P. and Buttenfield, B.P. (2003). Introduction: Digital Libraries as Sociotechnical Systems, pp. 1–21. In: Buttenfield, B. P., Van House, N. A., and Bishop, A. P. *Digital Library Use: Social Practice in Design and Evaluation*. Cambridge, Mass: The MIT Press.

Van Hooland, S., Méndez Rodríguez, E. M. and Boydens, I. (2011). Between Commodification and Engagement: On the Double-Edged Impact of User-Generated Metadata within the Cultural Heritage Sector. *Library Trends*, vol. 59(4), pp. 707– 720.

Verheul, I., Tammaro, A. M. and Witt, S. (2010). Foreword. *Digital Library Futures: User perspectives and institutional strategies. IFLA Publications Series,* vol. 146. Berlin/Munich, De Gruyter Saur.

Weinberger, D. (2010). *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.

Westbrook, R. N., Johnson, D., Carter, K., Lockwood, A. (2012). Metadata clean sweep: a digital library audit project. *D-Lib Magazin*, vol. 18(5–6). doi:10.1045/may2012-westbrook.

Yang, A., and Taylor, M. (2015). Looking Over, Looking Out, and Moving Forward: Positioning Public Relations in Theorizing Organizational Network Ecologies. *Communication Theory*, vol. 25, pp. 91–115.

Annexes

Annex 1. Quotes by the Staff

Symbols used:

[text] - text added by the researcher for giving the context of the quote
? - unclear text in the recording
[...] - skipped text, not relevant for the quote or included personal details
… - unfinished sentence by the interviewee
text - question or comment said by the researcher

The quotes may not fully express the institutional formal statements. The interviewees were asked about their personal perceptions around the topic of users' participation, which they kindly provided with the consent to be cited. Themes, which belong to several subsections (actors of the activity system), are hereby mostly listed only once, under the subsection where they were mentioned first. Quotations in this annex are grouped by themes in which they appear as groups of quotes per person. The interviewees are referred to in the end of each group.

Subject

1. The library, this physical location, urban location of this building is King's Cross. King's Cross has a very high concentration of information industries, British Museum, Wellcome Collection, [...] Google, all sorts of information industries are within literally a mile of this place.
2. We have a community engagement manager, who tries to get the local community to engage with the library.
3. I wouldn't say the whole institution [is behind the activity of enabling social interaction]. I think we had to get funding for Labs, which funded basically 2 people. And for a small time 2,5 people. It's now back to 2 people. We had to get external funding.
4. That [Wiki Food Editahon] was started three years ago by somebody called Polly Russell, who is the lead curator of contemporary British and she does a lot of research in history of society in modern Britain, and things like voting for women lots of things around women's issues and politics of domestic situation within the UK. She has actually appeared in television quite a lot. Also she's organising an event. [...] It's just something that started very small and is growing and getting bigger and bigger.
5. Actually we have a really important part of our strategy that we need to engage with our local community. There's only one person doing that, but

luckily I know her really well and I've worked on some projects with her, it's just because it's my own personal passion to engage.

6. Stella is very good in history of the Library, the culture and politics, why things are the way they are. I've been here only 4 years. Stella is 10 years, working on different posts in the library.
7. With handwriting, there is a project called Transcriptorium. It's a EU funded project, which we're a partner [...]. So, the idea is you have to give it a set of handwriting and then you also have to give it the transcripts for those pages. So, it's a bit like you are giving a training set to the computer and say, here's some writing, here's what it is in reality. So human has to do that and then the computer learns and then you unleash that programme on to all the other data.
8. One of the philosophies that we decided early on to the project was that if we are telling people to do this, we should eat our own dog food – if we ask people to do things, we should do it ourselves. It's easy just to kind of tell people to do it, but if we don't do it ourselves, we don't understand their challenges, their issues and so forth.
9. So what we're doing at the moment is, the Flickr thing was an experiment and we're probably conducting a lot of experiments.
10. The challenge we have is that this is a kind of a paradigm shift for the library. [...] So internally the problem we have is that everything is focused on our digital collections, so the culture of the library is very much focused historically around what we have physically. There is a cultural shift that needs to happen to move that into the place, where people are familiar with the fact actually the library isn't just about physical items, it's also about digital.
11. There are actually more and more public domain images out there. I think the problem we have, if we want to make things available, and I think you've identified this, is we need to understand and articulate, what is the value to the organisation by doing that. If we release things openly, make them public domain, make them accessible, what do we get. And I think that's a problem, and I don't think there's easy answers.
12. We use external evaluator to see how Labs is done and what impact it had.
13. My colleagues in Digital Research [...] have roles, which are called digital curators. And part of their role is very much focused on that internal agenda. So they run a very extensive training program called the Digital Scholarship Training Program. It's led by Nora McGregor, who is a digital curator. Digital curators are connected to collection areas within the library. [...] Nora e.g. is connected to Asian and African, I think, Mia is Western heritage... And what they do, they go to the public museums and then they try to understand what challenges there are in those areas, what training is required, what awareness they have of digital. The Digital Scholarship Training program offers training internally to staff, to maybe more traditional staff, who need to understand... learning how to use tools, to be able to interact with digital collections and data. So e.g. our one of the most popular training things that we've done is [...?] a tool that was developed by Google I think originally. It's a tool where

you can bring in for example metadata, you can clean it up. E.g. if the data format isn't consistent, you can make it consistent and things like that.

14. We've had lots of failures, we've learned lots of lessons. I think the dream of having everything as is on Flickr internally to make that external is probably maybe 5 years away in my opinion for a lot of our things I think.
15. We're not just national institution for books, maps and stamps. It's also a place for digital things.
16. The Flickr account[89] was set up years before we [labs] joined, August 2007 by marketing and press people. [...] We converted it into Flickr Commons account. It's free, it gives you a TB free.
17. Digital humanities has lead path to the philosophical origins of the project, also many Advisory Board members come from the area.[90]
18. We have systems, we have data, we have APIs, but they are not public. It would be lovely if they were, we have examples, but I think we need to kind of make more cases and the library and institution has to have a higher tolerance for risk. I think a lot of organisations like this don't have that tolerance for risk. I think when you're in experimental context it's okay to take risks. But for a large institution it's much more difficult to make those decisions.
19. *But in 5 years maybe you won't go out with these APIs any more?*
    Exactly, it's something else, there'll be something else. And you start all over. I would say the minimal is... e.g. to create data.bl[91] which is just file share, it took 2 years. Just to have that.
20. And in our new section, we've been looking at how we can open up archives, for example the radio programmes, TV programmes, where we don't have subtitles.
21. The whole space, we're trying to change the paradigm a little bit, so the largest data set that we have here is about half terabyte. Yes, it's half terabyte, which is large images from all the books. This is a very nice data set. So this is Hebrew Manuscripts. We have the metadata for the manuscripts so people can then download... for example these are 25 digitised Hebrew manuscripts from 1200 to 1599. So people can download them. They can do anything they want to do with them.
22. We have a collection of 109 000 individual sheets. 90 000 of them are playbills. So if you think about this historically, these are individual sheets, sometimes maybe newspaper cuttings or sheet music or a poster from a play, and these were originally stored in boxes.
    (Mahey, 2016, 2017)

[89] https://www.flickr.com/photos/britishlibrary

[90] http://labs.bl.uk/About

[91] http://data.bl.uk

23. I joined the library in 2015, a lot of projects had already been running, so things like Georeferencer, it has been going for 5 years, that was started by the member of the staff that's no longer here and then taken up by Philip Hatfield when he changed roles.
24. We're working together, the institution is behind it.
25. *Does the urban factor play any role here that you are located right here and surround by these different memory institutions or information industries?* Not really, no…
26. In the UK there's a kind of sense that London gets all the attention. So we're not putting London volumes up straight away. So people with a connection to London wouldn't necessarily come to this... that wouldn't come to this project, that connection to London.
27. People from ? collections to insure that we're correct in what we're saying about the playbills.
28. We're talking to people in metadata services again to make sure that from the start we can contribute the records back into Explore.
29. It's really hard to know what happens in the library, because it's so huge.
30. So it's quite deliberately run to events where people can try out interfaces we're developing. So really early feasibility testing with the idea that people will know about it and a bit of involvement and a bit like they're a stakeholder in the project and that would help it to succeed internally.
31. That hasn't been like a former process, where we had to go to some project board and say we're gonna do this. We just started doing it.
32. I think having people comment and enhance our collections is fantastic, but I think that there's a lot of collections... that historical collections come down through families and they're all tiny, they're all catchy and if we don't as a country look after them in some way they'll be lost. And there's no institution that really has the agreement to do that. And then like real Commons style platforms allow people to enhance... but without us being proprietary about things, that shouldn't matter really whether it's hold by the British Library or the National Archives or a local record office. I'm not that protective of the collections.
33. The collection of Playbills dates from roughly 1730–1950s. It's about 200 volumes about 270 000 pages. They were digitised as part of the single sheet digitisation project. So that's flyers that might be handed before performance or given out on the street... Basically one-sided sheets of paper. So the project didn't include things like programs, because they tended to be folded or statened? or whatever. They're quite rich in lots of detail, visually interesting sometimes, they tell us a lot about... because you sense what people do for entertainment. So some collections let themselves more?, they're not as amazing as say the maps collections, which are like gorgeous and visually interesting and instantly relevant.
(Ridge, 2017)

34. For Discovery it'll be me. Sometimes I'm quite neutral, I don't have a strong view one way or another, then I just get a steer, I just talk to... We do liaise, we're not an isolated bubble. I've talked to business and the business might have an idea from the professional...
35. We just haven't had the space, been focused on projects, so we haven't had the space to step back and think about... But I think this year we've been starting thinking about.. I'm hoping that as we're gonna not have any more projects, to create new features, new functions for Discovery. Enhancing existing function, but not big brand new ones.
36. A research team just launched with digital research within that team to work with digital humanities community.
37. Interpretation team, they did work with external communities to say something about it [the collections].
38. Vice-records knowledge team gives advice and guidance to the public, they also do cataloguing projects, including collaborative projects.
39. So that's what they're looking at the moment, if catalogues is something we get funded, technology to go out there, work with... But we're not there yet, we're working on that.
40. They've probably only saved it as a PDF image. But that's historic. Can we encourage departments to go forward? If you start to save PDFs now, can you save it as a text file, save it as an image file…
    (Grannum, 2017)

41. I'm not involved with that project [Flickr] now and [...] I know that the person who was doing my job after me has moved on quite recently. So I think very likely that activity there will pick up again, it's just that that person in post is ? just at the moment. So I'm sure, if you come along in six months there is someone to talk about what's going on with that now, it's a little bit vanished at the moment.
42. In the first I don't believe it wouldn't have happened [publishing digitised photographs in Flickr], if I haven't done it.
43. I started with quite a small number of villages, it was very difficult at the time to get the agreement here.
44. Again it seems slightly ridiculous now, but when we set it up, I suppose it must be nearly 10 years ago. [...] and it was very difficult to make it happen.
45. This is a relatively straight-forward organisation in some respects, in terms of publishing content compared to the constraints that people operate often under hidden small local government owned organisations, where they ? don't control their own website, they may not be allowed to use social media. Here policies were emerging.
46. With the sense of the direction of travel, I knew the ones we've got, have been very heavily scrutinized within the first, within the short period of time. But actually people get bored of that, nothing goes wrong, the house doesn't fall down and then the people sort of move on and they worry about other things.

47. That shouldn't seem to be so exciting, but in an organisation like this, actually content can be locked down, especially in particular teams. It can be very difficult to get a large number of digital images of the collections. It should be very easy thing for me to get a hold of that, but it can be very difficult.
48. The kind of a dpi didn't hold up for the images that was small. And that's always a problem with digitisation. Digitisation that you do has to meet the aims of the project you've got. In this case we were interested in knowing, who is that person or where is that. But if you can't find details in an image you can't supply that information. So there were definitely cases where the digitisation was carried out in such a way as to defeat our own stated goals for doing it. And that I hope is the lesson that was learned by the organisation. At least I think there has to be match between those two things, otherwise you'll be in trouble. But it wasn't the majority of the collection... So it was all right. But it certainly meant that those images as they were wouldn't be blown up to large exhibition panels. There's an idea that people should perhaps choose things that they were particularly interested and might respond to them and that would form part of the exhibition. That did require new digitisation, because we couldn't provide the images at the size so they could be used at the exhibition. That went toward a range of institutions within the UK.
49. One of the problems we had was that the quality of the digitisation wasn't really high. It was done in a very... in terms in a page sort of the way. Not really because of concerns putting the content online in the way that I've talked about before. I don't think that was the reason... I think it was kind of misscoped, because it probably would have been more expensive to be constantly adjusting the camera according to the size of the... So they were taken not rather an image at a time, but they were taking page at a time. But obviously in a large book with small images a resolution that's perfectly acceptable, if the image is in A4 size... for the whole page is completely unacceptable, if the image is much smaller.
50. They were digital images of a range of items from a collection [exposed for Flickr]. Some of them were digital, digitised photographs, but some of them were maps, some of them were written documents and some of them were photograph of object from a collection, they were digitised objects in that sense. So there's a range of material. If you look now, because of an enormous number of images that came out of the... what was called works through a lens, there is photographs from a Foreigner Commonwealth Office, Photographic Library. That's overwhelmingly bulk of the content and overwhelmingly photographs, so what you see is photographs.
51. From ever you are from, I can find stuff in this collection, which has interest to you, which is really exciting. And the internet allowed us to do that.
52. Because they [coloured photos] just purely, in a traffic sense, they seem to catch people's eye better, not that people don't like the well-tuned black-and-white photograph. But a Second World War photograph in colour is always going to be in more interest to certain kind of person than black and white. [...]

That was what I focused on our comparative narrowness. And just wishing we had certain kinds of sports material, not that you can't find nothing here, but... we don't have the amazing collections of baseball stars or... It was that. All the institutions would run little online events around a theme for example. And I'd have to think, we can't really contribute to that. We don't have any material about that.
(Pugh, 2017)

Objective

53. Actually what should we be doing, we should work cross-institutionally. That's what I think we should be doing. I would love it, when people would be using an image from here, from here, a text from here and creating something altogether around all those things. I think that would be fantastic, especially public domain stuff. That would be amazing. I think it's still early days. The institutions themselves are still figuring things out.
54. It's really trying to address the fact that people who contribute to Wikimedia Commons and Wikipedia tend to be geeky white men and it's trying to encourage especially more women to get involved in editing Wikipedia entries or contributing into Wikimedia Commons.
55. I mean Labs' primary focus is to get more people discovering our digital collections and reusing them in meaningful ways. So anything that fits that agenda I really want to support.
56. We did have a project awhile ago called Pin the Tale and the idea was that... it was a kind of a georeferencing project, crowdsourcing and that was around trying to get the public to kind to talk about stories around the geographical location that might be connected to our collections. So, the idea was that they were supposed to say that this poem comes from this place and they may have read it out or they may have provided some text and so forth. That was a project that I was involved in, took place a few years ago.
57. There are so many of our records, which are not online. For example in Asia and Africa, a lot of the records are still in card catalogues, they are not being digitized. So LibCrowds' initial focus was to see, if we could crowd source the card-catalogues.
58. They were digitized because they were at threat of being damaged or stolen because they were fragile or people could walk away with them.
59. The problem is we don't have information about the individual sheets. And in Labs we tried to look at this problem of how we might solve this and a new public library, a division called ? Public Library Labs tried to solve this problem, too. How do you get those individual sheets and how can you begin to augment this information, the information that's on them, even in basic way.
60. Trying to address how we can use computational technology to recognise handwriting.

61. Mechanical Curator enabled experimenting in order to encourage users to experiment. [...] So the Mechanical Curator was the kind of dogfooding. So we run our own experiments.
62. Our project is trying to get people experiment, do experiments with our digital collections. And the reason we are doing this is because we want to understand as a national library how we should be supporting scholars, who want to use our digital collections.
63. So the core of our project is that we are trying to engage with scholars, who want to use our digital collections and experiment with them, so we can understand, do we have the services, the processes, the people to actually support them to do what they want to do. We do not want to do this in a theoretical way. We want to do it in a very real way by working with real researchers' problems. And the ideal of work with them is that we understand what they want.
64. I would not say we were the first ones to do it, I think there were a few examples beforehand, I think they were probably the exception and because we were very lucky, we were given the time and space to experiment, but actually the most important thing was that it wasn't really about us, it was about getting our users to experiment. We did our own experiments, because we felt hypocritical, if we are encouraging people to experiment, we should do this as well. [...] Every year we launch a competition, we say, come up with some ideas of what to do with our collections and then we will work with you.
65. So, for example, all the data from the card catalogues, that's on LibCrowds, 'cause we haven't finished it all... For example these are the Arabic card catalogues and people can download them [...] as to form 95 MB PDF for example. So, what we're hoping is that people run their own experiments. Somebody in the computer science department might be really interested in that as a difficult problem.
66. So what we are doing is, we are trying to change the way people discover our stuff. The traditional model of a library is: look a single item at a time. data.bl allows you to just view large chunks of data, so instead of looking at one book, you can look at 65,000 books. And what we're discovering is that more and more people want to be able to have alternative ways to see our things. I think the model in the library, the British Library, really comes from the 1820s, the British Library was a department of the British Museum, and before 1820s all the books were on display. People could come in and they could look at the books and then the librarian called Panizzi changed the way the library worked. And that is the same system today from the 1820s. The system is to put all the books away from the public, because they might do horrible things to the books /*laughter*/ and the idea is that you request the item, the individual item and then the item comes up to the reading room and then the people see the item. What we are trying to do is we are trying to say No, you can get the access to lots of the items at the same time. And at the moment our library catalogue isn't designed like that.

67. Basically, what we are doing is we are releasing datasets [at data.bl.uk]. So for example the Flickr data that we gave you is now on here. It's still early days but … [...] What we've done here, it's very basic actually, all we've done is packaged up collections of metadata, collections of images.
68. But for us it was all about increasing discoverability of the books themselves, but also kind of repurposing the content within the books to be used for so many other things.
69. The mapping, the georeferencing, we haven't completed it. It's a much more difficult task, because it's 54,000 maps, we've done about 18,000. That's quite a lot left to do… We've been trying to think of new ideas to get the remaining maps georeferenced… basically we want to geotag all the maps.
70. So we set a challenge and the challenge was, show us the value of opening our digital collections as open data. That was our challenge, show us the value of it. What that means is, if we release it, can you build something that will show us, what the value of that is.
71. Maybe offering donations to organisations, maybe working with commercial companies to sort of see, if they're like for example maybe make products for you, where you get commission. All sorts of different ideas you could do…
72. I don't have a problem with it [monetising the cultural heritage content], I think it's a good idea. We have tried a few times to do commercial things with for example Flickr images. We've never quite done it. We've been really close in doing it, but not done it.
73. They [commercial services of the library] gave us low-resolution versions of the images that they sell, and the reason why they did that was that they wanted to harness the traffic that we were getting [from Flickr]. So that people would see some of these images and maybe request these as high-resolution images. But if you notice from every single Flickr image from the books we provide a link back to our commercial service for rescanning that. [...] We don't know still today, how much money that has made for the library. [...]
    So, the idea is you go on that page and, that page looks very different from other Flickr pages. It's not from a book and it takes you to Images Online, so it's about 500 images there. And then there's I think we have got some World War I photographs, which are put on there.
    We wanted, if you went to this form, it was already filled in with information. What we feel is happening is that a lot of people are clicking on this form but they're getting put off by user experience. We know we've been increasing traffic to that form by 3000%, but we don't know how much income that has generated to the library. We know they had some orders through from people who wanted high-resolution images.
74. We've tried to make sure that anything that we release, we try to provide links back. So e.g. we put 9000 bookbindings on Wikimedia Commons and they're images on book covers and then we make sure we had links back to our library, to the original database.

75. We always wanted the images to end up in Wikimedia Commons, that was always our plan [with 1 million collection].
76. But what we've intended to do is that obviously we have presence on the internet.
77. But the idea is that we shouldn't just assume by putting things on our systems that people gonna discover them. Even as good the metadata might be, even if we let people to index them. We have adopted a philosophy on that: there's nothing wrong of putting copies of that elsewhere.
78. Remember, these are books [in Flickr] that hardly anybody saw in BL. They just sat there, not many people took them out. [...] Because those images, probably some of them have never, since we got them, nobody has ever seen them.
79. The idea was to get volunteers to provide recordings, sound recordings of the sounds of the sea where they lived. A project called the Sounds of Our Shores[92]. It's what did the UK coastline sound like in 2015 in summer. Those are recordings of nature and machines, animals and the wind and rain and also, recordings like cars and trains and machines, things like that. Our collections are not just books. I think it's interesting to think about the possibility of sort of other formats.
80. E.g. if we put it on Wikimedia or Flickr, the chances are more people will discover it then not. And I think that was the original motivation of doing that. I still think it's important to keep the connection. [...] Some people would feel we are giving our crown jewels away for nothing. I think we need to figure out how to create that value, how we can articulate it.
81. Very early when we started to publish images, people asked us to rotate the images, because they were in the wrong orientation, because they were landscape images, but published portrait.
    (Mahey 2016, 2017)

82. So lots of different activities either use the collections as sources of information or trying enhance information about the collections.
83. There's a lot of points, where you can get interested in [the collection] and start thinking of questions and I'd like to see that as a way helping people realise that we knew that they're interested and suddenly they are and then we support them in learning.
84. They [playbills] are brilliant source for family historians, local historians, people working on theatre performance or whatever. They are not discoverable at the moment.
85. So we are looking a crowdsourcing project to specifically to get information we can put back into our catalogue, so that people could find actual individual playbills.

[92] http://sounds.bl.uk/sound-maps/uk-soundmap

86. We're working with people from ? collections to insure that we're correct in what we're saying about the playbills, the information that we create would be interesting and useful.
87. To show people that the things that they've done actually have an impact and actually help other people.
88. The library recently launched a IIIF, it's international interoperability framework, which means that you can share images across institutions, based on a common standard. So we're using IIIF images, and they're kind of designed to be easy to share and download and do things with. So we're planning to use that functionality [for Playbills] to make it easy for people to say... I found this really crazy theme or whatever.
89. I think the point is that when people look for information and you should get accurate, authoritative information and if we can contribute that to happen in Wikipedia, that's a good thing. It's not about us really. It's about helping people to get good information.
90. So I think it's very important to show people that the things that they've done actually have an impact and actually help other people. So by making sure that the material can go back into Explore relatively quickly, it means that we're able to actually to deliver on what we're promising.
91. To make the most of computationally enhanced records that you might get from the student projects where they're tagging, looking at the tags or using entity recognition software. So we need to be able to ingest that kind of materials.
92. We're looking at going volume by volume, but it depends on the interface, whether we can design projects so we can say, if you're interested in this region then here's all the tasks that you can do with this. So the micro tasks and the bigger tasks. Or some people won't care about the region and just care about the tasks. So we're trying to design a way of presenting a different micro projects to beat, but ideally we'll do volume by volume and then... just metadata that we can give to the catalogue people to ingest and then more volume is discoverable and then keep going like that, because otherwise that's so big that if you take.. if you spread out the effort across all the volumes it would take really long time to anyone…
93. Wikipedia obviously have some biases in terms of regional coverage, gender... pretty white-voiced... But we can work to change that.
94. Well. the mission is to help people to explore the world's knowledge. So the fact that someone on the other side of the world could access the material and would use it and do something really creative with it. And I love what the scholarly users have with the collections, I love the fact that the reading rooms are full of people working hard, but I also love the fact that it brings a joy to people as well. And I think that's a way to access... You need to do the hard stuff, and the easy stuff and the fun stuff and the challenging stuff.
(Ridge 2017)

95. We have 5 years cycles for corporate planning process. We did have online user collaboration in one of our former goals, but it's not in the current one.
96. What is the level of detail for formatting and describing needed to reach the objective?
97. There's quite a few academics interested in the data, I don't know where the data is, but no one's really got back to this is how I want to use it, this is how I want it to be improved, the value we have behind it. Or this is not really useful at all, so please, if you could do it this way. We haven't had that sort of steer. So there hasn't been a push to say as we were having to do something about it.
98. We are still playing with the idea about public-private tags. So people could use the private tags for their own areas of interest and there might be something on grandfather there, there might be something on grandfather there. It's almost like a wish list. What we don't know, what we haven't done is actually - what's the motivation behind it, why did they tag these? Is it because they found their grandfather there or is it that the grandfather might be there and I'll go back to that later. We don't know that. But also for the great scheme of things, how valuable is that data itself?
99. Or doing researching and then pulling different sources they had related to that topic or subject.
100. So what we've found with the record that a researcher might only look at a couple of pages, but saying something about those few pages still opens up that larger file to start people to get a hook.
101. I think that, if you get a hook in and that's what the researchers often want.
102. But anything that gets people a clue that there might be something and that's worth spending a bit time and effort to explore, to see what else there might be.
103. We brought out things you would necessarily consider.
104. One article of wiki was about roll names of muster, so lists of people on board that were on the ships. But they recruited people who were just prisons of war, passengers, and one of them that I looked at were Turkish slaves that swam abroad of the ship and they were listed because they were just getting food and drink and therefor they were listed on the ship's muster. You would not expect that. We didn't necessarily list everybody that were on that muster, but we did highlight something really unusual. So someone was looking Arabic slaves, 1720s, in Malta. What it does do, it opens up barriers of research that might not be considered before, because you've drawn out something that seems a bit unusual I suppose.
105. It's Zooniverse project at the moment for operation war diary where we uploaded images partly as PDFs for people to... I suppose to indicate people and places.
106. ...they then have the opportunity to say something about it.
107. People could say something about the record.

108. So the idea with Your Archives[93] was about people could say something about something they looked at.
109. But I think the interpretation team, they did work with external communities to say something about it [Flickr collection].
110. Which was trying to create a new front end for people to find stuff across multiple datasets.
111. So from the users point of view there's a single point rather than actually go through multiple points to get the same information about the same record for example.
112. How do you get the new ones, the one that's slightly different from all the others that are identical? So those are going to be the digital challenges for humans and machines to make sense of all this.
113. We also had historical manuscript commission, now archives' sector development [...] They have responsibility for also oversight and advocacy of the archive practice across the Wales. And part of their responsibility was actually to identify core collections related to British history, not just the British archives, but also archives around the world. But these were run in different datasets. So the view was actually... to do that for National Archives stuff, but you can't do that.. you gotta go to somewhere else to look at that database and that database and that database, so for most the idea was to fall? those in.
114. We don't have moderation tool. I suppose what we wanted was a large amount of data before we start thinking what we do next. And we haven't thought about what to do next yet.
115. Transcript palaeography. I think they worked with the Wellcome Institute on the... basically to teach the computer [recognize handwriting].
116. So we need to standardise things that way. [...] So after awhile you're free to have on-going continuing value, I think that does need to be managed for even crowd sourced information.
117. So that's why, what I meant, our organisation is structured that way, and increasingly digital. It's not humans looking at the data, it's machines looking at the data. So do we mark up the data in such a way that actually allows machines or people as well to do that? thing: this is the person, this is the place, this is subject. Can we annotate it in a way that allows machines and people to do more clever searches with it?
(Grannum, 2017)

118. At the time when we first experimented with that I think the case that we made, social media and Flickr seemed more like a social media platform back then than it does now. Seemed to offer an opportunity to really talk to users about content, share content, to kind of put stuff out there in a way that people could respond to and share.

[93] http://yourarchives.nationalarchives.gov.uk

119. We didn't really know where it was gonna lead. Well, I always thought that the risk was very low, I wasn't really expecting it to seriously blow up. But it did feel like a kind of experiment at a time, but I think it's been broadly indicated as an approach.

120. Something like Africa through a lens, which was the first large collection from the FCO to go up there. We knew we had images that we didn't know what they were. We knew we had images where we know perhaps to put it very bloodily, who the white people in the image were and we didn't know who the Africans were. And we knew that the people out there were gonna know more than we did, because that's the way the world works. Collectively, of course, they were gonna know. So that was one kind of engagement that we were looking for and we hoped for.

121. I'm not sure I was ever quite that strategic about it. I think that I saw it as a two-way conversation that we were trying to facilitate. And not conversation with us, but the conversation between the images and the people who were interested in them. And we were around to help with that. And that to me, that fostering the engagement, consisted of... was trying to do our best to put things in front of people that really interested and engaged them. And then looking to see what the results of that were. That's what I found exciting about this was when people did come and say, oh yes, we used to go..., it's the house of some important local official in town and area that don't even... I can remember when I was a child going... And they can tell us a fantastic story about that. And well, that's the point. [...] There was a lot of press interest, there was lot of traffic. We were having conversations with some of the people, too. But most exciting was people talking to each other.

122. So, it's never ideal, but in terms of kind of initial plan of with interest and putting things up and engaging with the online community, I think it was very successful.

123. The other kind was, just the people would see themselves or see something interesting and talk about that. So wouldn't necessarily be just about detective work in order to improve our catalogue. It was about just fostering a genuine engagement with those images.

124. We had a goal run our own online community a few years ago I believe. The member of staff that run that project isn't here any longer. At the time I personally thought it was a bad idea because for the same reason that I thought that the same kind of viability problems seemed have devilled Your Archives seemed inevitable there. Where the people want to have discussions about the family history, they want to have them in the place they already have the discussions about the family. They don't want to have them on... there are already very successful family history forums. I thought the logical thing to do was to set up there and talk to people.

125. We still could have that with Discovery, the ability to say, oh you're interested in cheese, here are main kind of record about cheese and for someone else to come back and go, oh I found some more stuff actually and

we have nothing like that now. And with my researcher hat on, that would be incredibly useful, because we can't meet researchers halfway. We can't say, oh from the query you've typed in I understand your information need is X. Let's bring our expertise to there to answer that question for you.

126. The other one I was involved in was what we called “This is how it was”, Wikimedia UK funded the digitisation of a collection of war art. We moved that onto Commons in order to find out more about that collection. And with the aim of enriching Wikipedia obviously, Wikimedia's aim was to provide high quality content for Wikipedia articles. So we did learn a bunch about…

127. We are having at the moment a project called Manage Your Collections, where we're looking at... we're working with other archives to bring in data that they've collected on their stuff. Some of them have got.. they don't have digital catalogues, they just use Excel to keep records of what they've got. We're gonna bring those in and move them in our system. And obviously that's great for them, it's useful for researchers, who wouldn't otherwise have known what is that they've got.

128. I think, the best place for cultural content is where lots of people can see it and then you get a range of interesting reactions and you let them to decide, what's appropriate or even inappropriate response to that. You just let people play.

(Pugh, 2017)

Subject-Objective

129. We looked home for these 1 million images. [...] We started looking at these images and we saw that they were really beautiful, especially when you juxtapose them next to each other. Because when you take the images out of context... They were just languishing in the library. So we thought we need to put a hung? for them.

130. So when we released the Flicker collection, we actually had other people approach us from the library, who want to put more collections on there.

131. That’s because we haven’t done really huge emphasis on promoting this [asking users what they use the BL data for]. But we now want to do a lot more promotion on this. And next what we’re doing a big staff talk to promote this, to get people to say... Essentially what we’re offering here to the library and library staff is like a huge Dropbox for library stuff, so they have got stuff that they want to make available and we can give them options to do this by putting it here and I think we're changing the paradigm of the way the library maybe operates.

132. For example maybe in some ways this is [data.bl.uk] our attempt to do some big data. But in other domains it’s quite difficult for them to understand that six volumes of pub signs is a big data search for a library. For a scientist,

big data is like data from buses and timetables and temperatures and things like that.

133. And I think the fundamental point I would make for the Library to learn is that you need to space for library, for any institution actually, for any memory institution to be able to experiment. I don't think that mentality exists in a lot of organisations. They don't have time or the resource. So for example and easy way to solve this problem would be to give staff one day a week to do experiments. That's what they do in Google for example. People can work on their own little projects. But I think that is a very big cultural shift.

134. I believe we have the mechanisms [to take users' outcomes to other users], but we don't have the people, who can make the decisions to make that happen yet. I think we have done some small experiments to do that. We would love that to happen. So our catalogue system is Explore, Primo. I think it's technologically possible, but I think there would be issues around data quality, which would stop it. And I think that's why we haven't done it up to now. I think there is a fear. There has been discussion about putting the Flickr data into another discovery layer. That's never happened. The only thing that's really happened is the example I gave you with 3000 maps. It's a start.
(Mahey, 2016, 2017)

135. We weren't just looking for doing a crowdsourcing project. And then looking for a collection. But this collection needed, people wanted it to be used, discovered and to…

136. Playbills is one of those collections that a lot of people think has a potential and so the idea has been discussed since... I think another digital curator Stella Wisdom and Tanya Kirk, who is from printed heritage first began to discuss crowdsourcing around here about 2013/14. It's one of those collections that people always think what should be done with it, they're quite rich in lots of detail, visually interesting sometimes. [...] So there's certainly been a lot of people, who seem to be keen on something happening to that collection. [...] People from the reference, people from the reading rooms, a lot of people working or have contributed, have worked in the reading rooms. They have a bit of sense, how people receive different kinds of collections, which might have formed their enthusiasm to the playbills by seeing other people's reception of them.

137. It became possible to do the project finally, because Alex's role is pressed on software development results in the team and people who were previously resistant to the idea of the project left. So lot of things aligned to make it possible.

138. Sometimes it's clearly some people [talking in abstract way] doing it because the technology is there or it's trendy.
(Ridge, 2017)

139. Personally, I think we learned a lot from Your Archives. I was very, very, very disappointed that we weren't able to continue with it. I still see that is a future. I think we don't have any choice but to.
140. I genuinely don't know what the organisation's direction is when it comes to user collaboration. [...] My personal view is that we have no choice but to and I'd be really keen to push that agenda, but I don't know what the organisation's direction of the travel is.
141. And eventually we go full digital route. How do we expose digital content in itself? Not digitised, but digital content. Is it still: I want that, order now. It comes to me, do you want machines to harvest it? Do they need to go through order now or question now button to make it physically available?
(Grannum, 2017)

142. But copyright was a concern, not copyright as such, but control that rights. We were at that time as we are now licensing digital images, we were generating revenue by selling digital images. And as it was also giving the images away and it made people in certain departments actually quite anxious. So we had to agree about the quality of the images that we would be giving away at that stage and say they would be sized in a certain way and we wouldn't be giving away the crown jewels at that point.
143. I wouldn't necessarily subscribe to it now, but at the time it was about, I can remember, quoting and talking about idols or attention and to put that sticky content that people wanted... Or perhaps saying we used to want sticky content, now we want spreadable content, we want content that people gonna actually.. And I have to say now it all seems very cynical, hard-nosed. But I think we didn't know, what the outcome was gonna be either.
144. And then there was a project based on Commonwealth Photographic Library, which was accessioned here. So the photographs were just come in and it was decided to digitise the whole lot. Then we had to decide what to do with them.
145. If you look now, because of an enormous number of images that came out of the... what was called works through a lens, there is photographs from a Foreigner Commonwealth Office, Photographic Library. That's overwhelmingly bulk of the content and overwhelmingly photographs, so what you see is photographs, but that wasn't the original plan. The original plan was to show the range of things of that collection.
146. We talked about large-scale ingestion of user tags, but we never did it. Even when... now we do have our own tagging functionality attached to the catalogue. It just wasn't really pursued in the certain point.
147. But we haven't systematically collected Flickr tags and compared them in that sense, even though the API would probably make that pretty straight-forward.
(Pugh, 2017)

Tools

148. What we have done recently is... 50,000 of those images are maps, those maps are being georeferenced, and 3,000 of those maps are now... the latitude and longitude of those maps and a record is created in our library catalogue, so they are now there in our library catalogue as a sort of a sub-record, like a child record of a book. So to say this book has this map and there's the latitude and longitude of it. So we've started that process, but I think there are all sorts of challenges and there's a lot of resistance.

149. So you see that maps have quite specific tags, for example 'geolocation'. That's really important, because that's a different set.

150. That's because every map needs about 10-20 reference points to make it accurately located on the Earth and some of them are really hard maps as well. Some of them are a picture of a burning castle next to a river and that's a map. It's difficult, because it's historic. Some of these you could describe as plans not maps.

151. I think at the moment our priority is focusing on the service we've launched at data.bl.uk, which is a deliverable, which we said we would do as a part of our Labs project. [...] We've created data.bl. Some of our collections are on data.bl. By the way data.bl is just a very simple place to download ZIP files, there's no API, nothing. It's that small it is. It's just the data in raw format. We have collection guides, we have 172 collection guides as of to date. The majority cover our physical collections, some of them also include some things to our digital collections. But more work is needed.

152. Another way, that is something completely different is actually how we could unlock our archives through people reading them out and then using voice recognition. For example TV and radio don't have subtitles, and then we run them through speech recognition and to see if it generates the text. We've been using something called Microsoft product from Maven where you essentially give it to audio and then it tries to create the transcript. And again there's similar question to OCR, it's not perfect, it's not bad.

153. We also on every image, if you click on the viewer [in Flickr] it'll take you roughly approximately to the page in the book where the image came from. [...] The viewer was the way we decided to provide the access to the digitised books on site of the library initially. Then that viewer became public.

154. If it is a map, it goes back to Georeferencer[94]. You can see it's been georeferenced, the old map on a new map. [...] We used a company called Klokan, they're based in Switzerland. We gave them the images. They've got an interface, by which on the left hand side you put up the image on Flickr and on the right hand side you put up an image of the world. And you basically try to curve, it's a little bit like snap, you try to map the image of the old map with the modern day image of the map. When you think they're the same map or

[94] http://www.bl.uk/georeferencer/

the same place you plot points on each map to show that it's on the same location. So you're plotting like reference points. And then what it does is once there's enough agreement, that map has been georeferenced, placed on the earth. You can look at it now or in 1850 to see and have an overlay. It's quite useful.

155. The Flickr account was set up years before we [labs] joined, August 2007 by marketing and press people. For many years there was a handful of photographs about BL.

156. Wiki Food Editahon, it's on site. The idea is to work with Wikimedia and Wikipedia UK. So they often come and the idea is to create... more and more people who create entries... Part of it is about collecting images and things, which can be popped to Wikipedia and other part is to create Wikipedia entries. Every year there is an event called Oxford Food Symposium and it's all about getting world experts in food, food writing, food critics, chefs, and they share knowledge about food. So historic knowledge, contemporary knowledge and I think there's food as well [*laughter*].

157. Another event we did was similar kind of idea finding things in messy things.

158. We ran a hack event in July where we took along the digitised books on a big storage device.

159. We've been trying to think of new ideas to get the remaining maps georeferenced and what we decided on, we're organising an event for teachers, geography teachers and we're gonna teach them how to tag the maps, geotag the maps. [The Way Ahead? Mapmaking and Digital Skills for Geography Teaching 12/11/2016 9:45-13:30][95]. And they are going to teach their students how to geotag the maps. So hopefully we'll have UK school children finishing off the job for us.

160. We were organising a hackathon at the BL, because some of the volunteers were asking us, they wanted to find all the maps. So we thought okay, let's organise an event and it was on Halloween 2014[96]. About 2 weeks before the event we got an e-mail from Mario, saying I don't think you need the event, because I found all your maps. So Mario tagged about 22,000 maps just using computational techniques. He is the second tagger.

161. So they were digitalized and 90,000 of them are playbills and these single sheets were bound into volumes, 323 volumes and a few years ago catalogued to the volumes. The problem is we don't have information about the individual sheets. And in Labs we tried to look at this problem of how we might solve this and a new public library, a division called ? Public Library Labs tried to solve this problem, too. How do you get those individual sheets and how can you begin to augment this information, the information that's on them, even in basic way. So they developed a platform where you can draw

[95] https://www.edcentral.uk/component/vikevents/?view=event&itid=101

[96] https://wikimedia.org.uk/wiki/Digital_maps_Halloween_tagathon,_October_2014

rectangles around the title for example and say what the title is. They developed their own platform to do this and we did some experiments with that and for various reasons we didn't implement it. But now Mia is working with a project with Zooniverse platform.

162. But he [a winner of a competition by a government organisation, where BL set the challenge] built a tool and the tool was designed to show what happened to our images, when we've released them. So the problem you have is, say you've got your own personal blog, you take one of our Flickr images and you crop it, you may print it, you put it on your T-shirt and there's a photograph of you with the T-shirt and there's a picture from Flickr on the T-shirt, right. Now, unless you tell us you've done that, we don't know. And the problem is that the original image has been changed, so how can you find this thing. So he developed a tool, called Visibility and it uses... best way to describe is like fingerprint technology or DNA type of technology to look at DNA of a digital image and then it looks for those similar images across the internet. And then there's already a problem, because the biggest image index in the world is owned by Google, Google has a large collection of images, they index them. They don't really make their API available for that, so he had to use a different API from a smaller company that only has an index of 10 billion images, which still sounds a lot, but it's not actually, company called TinEye. And what they do, they use something called reverse image search, you take your image, you drag it onto their interface and they try to find the same image. Even if it's been transformed, that's the key, because most people transform the images. And what he did, he put those there, it worked. We worked with him and what we discovered even in TineEye smaller index of 10 billion images, which is growing, we were still able to get some hints on what happened to our images.

163. When we did the digitisation of the books in the first place, the digitisation was done deliberately to ultimately scan the books for the text. That was always the idea, because the contract with the Microsoft was always around the OCR in the books. So when we were capturing the images in the first place, the capture and the post-processing of images was very much focused around an OCR. The original images you can see were yellow, the yellowing of the books, like in the original, they were also post-processed. They were made black and white, purely to improve OCR recognition. What we did is not immediately but a few months later, if you notice on every image we were also able to link to the full OCR text of the book. As a JSON file. It's not obvious. We added this afterwards. It doesn't work at the moment, because we got a grant from Microsoft research for cloud computing. So the service was kind of the back end of this was hosted there, it was called Mechanical Curator. And we were then able to provide links to the texts. This services was given to us on a trial basis by Microsoft research and we don't have it any more.

164. E.g. one collection where we seem to develop a bit of expertise are our newspapers, digitised newspapers. They were digitised and then the papers are scanned and OCRed. And that data from the OCR process is messy. So if the paper is dirty, has been folded, if there's dirt on a scanner, the words that are created by OCR are not always exactly what they were on a newspaper. And that means it's messy and dirty. We have millions of pages of newspapers that have been digitised. And although we've been working with commercial companies to develop products to build interfaces to search them, what people don't realise is that a lot of the data is not accessible. So e.g. if I was looking for pet food for dogs from the 19th century, those words 'pet', 'food' and 'dog', if I try to search, the word 'dog' may have been wrongly transcribed by a computer. So a word 'dog' may be 'd?g'. When I try to search for that it's invisible. I would say at least 40-50% of our data is like that.

165. But I think what we discovered, we did an experiment, where we tried a re-OCR newspapers, 1,200 pages as an experiment to see, if the quality... Because when that was OCRed in 2010 and obviously that was a long time ago, the software has improved. We wanted to see, if there was an improvement. Same images and there was a big improvement, not perfect, but much better improvement.

166. It's not for every book where you can download the PDF, we mean OCR, so it may not be accurate - it is full text, but it is full text generated by the OCR software. We use this Russian software ABBYFine reader.

167. We sent them [Sherlocknet team][97] the million Flickr images on hard drive. And they were starting using convolutional neural networks. It's artificial intelligence. Using a Google framework called TensorFlow. And basically they are using artificial intelligence to automatically tag all the images, which would significantly improve discoverability of the images. The second they want to do, which is really ambitious, is they want to capture every image, e.g. 'man with a dog on a horse by the river'. They want to do it computationally also. They claim their tagging accuracy for the images would be 86%, caption accuracy probably maybe 40%, still not bad. I think it's going to have a big impact on the Flickr collection. That's why we went to this project. What they're doing, is they're going to build an interface.

168. [Flickr dataset] doesn't always give us a reliable data sometimes. It's the best attempt. We've been trying to grab data from the Flickr API and it's not always there, there are sometimes gaps, but you'll see in the data. It's not our fault, it's Flickr's fault.

169. #bldigital – everything published on digital resources has this hashtag.

170. So we snipped out the images, then we did some face recognitions, experiments, we found because these were illustrations, the software library was really good at recognising female faces but not really recognising men.

[97] http://blogs.bl.uk/digital-scholarship/2016/08/sherlocknet-tagging-and-captioning-the-british-librarys-flickr-images.html

Through that experiment, when we were snipping out the images we thought maybe we should make a home for the images, so we created Mechanical Curator[98]. And that posts an image every 30 minutes to the thumblr. If you look down here, these are the tags that Mechanical Curator has been adding to these images. E.g. that hashtag is a date, a publication date, but also if you look at these tags, these are the most interesting two tags: 'similar to'; and 'new train of thought'. This means that the Mechanical Curator actually has it's own brain. It is using software libraries, which are around image recognition and it is analysing the image and saying "Aha, this image is similar to this image" based on a software library. So e.g. one software library might be about looking if the image contains lots of circles or lots of squares in it or the image is slanted. Or it could be the metadata of the image, find an image from the same place of publication or the same date of publication. Once it finds a similar image, it posts a similar different image. It's like a human curator but it's a computer doing it. We didn't decide to put these on Flickr. But we have these tags on Mechanical Curator, which is a virtual machine, which runs this service. If you look this one here... There's a software library, that looks for images the have kind of circles in it. Also it looks for slantyness. It's important to know, that curator curates the collections, it's a computed curator, so it's making curatorial decisions based on algorithms. It is tagging but from a computer's perspective. [...] Users cannot add tags here. This is only a kind of experiment for computer added tags. [...]

Some of it is taking data from metadata, some of it is taking data from an algorithm, which then analysis the image and decided okay, this image has this slanty. Whereas 'new train of thought' means that computer has got bored and is now trying a different algorithm. Sometimes it looks for faces, sometimes for slantyness, but it's really dependent on these software algorithms, the libraries who determine, what choice the computer is going to make. So it was a little experiment. The software libraries are open source libraries. They're based on Python. Remember that it was computer posting these images in every 30 minutes based on algorithmic method.

171. We were slowly putting the images up over about 2 months period from August 2013 up until the final millionth image (1,27 million or something like that) was uploaded. When we uploaded a final image we wrote a blog post[99]. [...] But I think, if you look at that blog post, about the origins and how people found stuff, it's probably quite important kinda starter. [...] I think this blog post has helped a lot in terms of discoverability of those images. If you google British Library Flickr commons, I think the blog post comes up second. We did that on the 12th of December [...]. This is the only announcement we made. And it just kind of went crazy.

---

98 http://mechanicalcurator.tumblr.com

99 http://blogs.bl.uk/digital-scholarship/2013/12/a-million-first-steps.html

172. [...] was the first person I spoke to get the images onto Wikimedia Commons. And he said it's unfortunately not probably a good idea. Simply because it's undescribed data. There is not information about the images. If you go to Wikimedia Commons, a lot of images have ended up on Wikimedia Commons, because now there is enough data to actually start to add more information about them. E.g. some of the maps are now in Wikimedia Commons and some of the images. We always wanted the images to end up in Wikimedia Commons, that was always our plan. But the strategy we ended up doing, was let's get them tagged first.

173. We're getting external users to add and enrich our content. And I think there is some way, how can we bring it back to our systems, a discovery.

174. This is all about linking back to our systems [links to approximate page of the book were added to the Flickr images]. The plain text.. We thought of adding this as a link to this page we noticed that discoverability has improved significantly because Flickr was using it to create it's index. [...] And it's not just in English it's in lots of languages, because the books cover not only English but different languages. [...] We probably need to think again about how to provide that link back again, because what we found was, when we provided this information, Flickr was using it to help to build its index, it was crawling it. So it was using the OCR text to help improving discoverability. It wasn't just the tags that were helping to improve the discoverability. It was also this.
(Mahey, 2016, 2017)

175. I've seen some unnecessarily complex projects and some task-flows. And talking to Douglas ?, the theatre curator here at the library, the first version was just crowdsourcing playbills or programs was quite complex. So it's mark an area, mark the role, mark the person and it was like lots of different steps and some of them required you to understand a lot of cognitive overhead that got kind of obsessed with outreach? team.

176. I think the catalogue is one thing and the objects is another.

177. [Playbills transcription project, online] It's basically either mark up an area and then transcribe it. We were talking about tags... But we're probably like make it a generic comment field and then see, how it's used.

178. I think the bigger challenge is machine-learning technologies. That we're working on another project, the Transcribers project.

179. *Do you plan to cooperate with Georeferencer anyhow?*
The way that the playbills are collected, it tends to be one theatre for entire volume, so there's really only one location to be georeferenced. And it's an address in text. So it's uneven? kind of information.

180. Editahons. So, getting together to improve things like Wikipedia entries, food history is in this week I think. They are tagging particular kinds of collections of animals and things like that.

181. We had long conversations with Yale's library using the software they were developing, based on Zooniverse software. But we found that it would be too much work to adapt the code for our sources. And we already had LibCrowds to hand so we used that.
182. I don't know if something's gonna be around. [Flickr] It was a fantastic platform, but clearly I worry constantly that it's not being maintained, search is getting really bad.
183. Flickr Commons isn't... I personally don't think that Flickr Commons is the place for illustrations, because it's for photographs. But if we had another Commons platform then that would be a really appropriate place.
184. We talked about having it [links to social media from Playbills platform] so that you can share.. if you find anything interesting you can share it on social media.
185. You can sign up and get an account and then you can see the things you've contributed to and there's a leaderboard, whatever they like leaderboards, but some people do. Or you can do it anonymously. There's a type you see, there is not too much chance of spam or anything like that. And there's a forum that's built into the ? software.
(Ridge, 2017)

186. We actually gonna form now – Wikipedia did that – forming categorisation tree. If it's a new category, fine, but first of all you should be assigning it to one of the existing ones as it brings all the related content together. We don't do that.
187. Well, we had online catalogue, online database. Quite as traditional archival library type of catalogue. But we had intentions to... it was quite old, I think even then it must have been 8-10 years old. And technology moved on, it was unsupported application. We could fix it, but we could not develop it. So new features and functions that we wanted, we just could not do that with the existing catalogue as it was. So there were intentions to try to rebuild it. We put the energy to build, to create a new front end, but what we didn't do, is actually to build the whole... that's actually that we're planning at the moment, to build the whole new thing. We just rebuild what the user sees. And then added features and functions to do faceted search and all that sort of aspect we did, which wasn't possible with the original one.
188. And also we had an idea, we wanted to bring in other data sources as well. So before Discovery there probably were... must have been 8 or 9 different datasets. And we brought, Discovery has now I think 5 or 6 of them instead of few other datasets out on the website. But most of them are now in the Discovery.
189. The user account is stored at different server than the tag one, so they don't talk to each other. [...] That's stored in different things, they're not brought together, so I could see who has tagged what. So say in spam point of view, I could go back to them and say, stop doing this now, and suppress their

account, if necessary, which was quite easy to do with wiki. 'cause everything was linked to someone and I had control over what that someone could do. We haven't built that sort of functionality into the tagging. I suppose what we wanted to do, we wanted to grow, but we didn't, so we came out with a sort of this...

190. When we scan stuff, it comes as JPEGs. The only ones we've done with the full text search so far are the Cabinet papers. And we started that project about 10 years ago, disc?-funded. The quality of OCR then was not as good as the quality of the OCR is now, but there's enough in it to give you an idea.

191. Everything is full-text except the image files. The videos aren't full text. Audio is not full text. Pictures aren't. And I think that PDFs coming from most departments aren't probably, where they've scanned or saved a document as a PDF and they've probably only saved it as a PDF image.

192. Manuscript records, hoist of our records, even if they're typed they'll be annotated. Very few of our records are... probably 20th century ones are typed. But before that, they're all gonna be handwritten. It'd be nice, if we can scan it in a way that even if the machine can't read it now, machine can read it in the future. But if they can read it in the future, and there's an image file..

193. The idea behind that [Flickr collection] is actually, we didn't have a space to promote these, mainly what they called “Through the lens” projects, “Africans through the lens”, “Caribbeans through the lens” etc. So we had this rich photographic collections, but we didn't have a natural place to promote it, to share it even though we knew there's a value in it. So Flickr at that time was seen as… We had these valuable, visual things, but no natural, we didn't have our own technology stat to promote it to the Flickr route.

194. We have experiments, we tried Zooniverse, Flickr, we've tried wikisource, we've tried various things.

195. The channels are more mature now that they were. There are new technologies out to handwriting recognition. It wasn't even conceivable even 5 years ago. So, we can only automate so much.

196. We've got a project at the moment that's kicking off. We've been trying to work with transcribers. I think it's German software company, who do transcript paleogeography. I think they worked with the Wellcome Institute on the... basically to teach the computer, you do several pages and you teach the computer what those... But they're not doing it word by word, they're learning blobs. So the person writes hopefully that blob will be consistent across all their writing. For records which are using a regular hand, usually pre-19 century, where they're written in a common style and all the clerks wrote in a common style, you can probably teach the computer quite nicely to recognise those blobs of text. As soon as you start getting universal education and free form, that's gonna get a bit more difficult. I think that with transcribers, using wills, proback? records, registered wills, otherwise written by the clerks... You do testimony, you die, your executive takes it over. So you get the individual

document, the file by the individual, but then you can pay for a clerk, who went into these registers. So there's tend to be a fairly common form. Cause you often might get the same clerk, who might have been in these registers over many years, but also clerks written in the same style. So you got several clerks writing it, but the handwriting is fairly similar, uniform. I think mid 18 century might be a good example, we might want to use that.

197. But what we haven't done again is wait tags. When I first looked into this was - that record tagged, say Jamaica, that person tags it Jamaica, that person tags it Jamaica, so increasingly it might not say so in the description, increasingly the most important tags shout at you almost: This record is about Jamaica! Doesn't matter what anyone says about it.

198. I've used delicious in the past. Delicious was a social tagging - it allowed basically you to tag any website, a bit like stumble upon it and read it. So you could tag any website and saved your space. But you could also mark tag as public. So they're private by default, but you could mark it as public. So people could collaborate.. Websites about project management... So people doing markup for these websites doing project management, then you can go to delicious an see the folder, these websites are that people have indicated as useful for project management. Weather delicious still exists or not, I don't know, but it must have been 8 or 9 years ago since I used that. We used to have tags in Amazon, tag books. I don't know, if you still can, but again I think by default they were all public..

199. Our own catalogue referred to the material that were public records, but not held by the Public Record Office in the National Archives. So there was stuff that was Post Office Archives, National Art Museum even within our own catalogue, within historical manuscripts catalogue. There was stuff that were not public records necessarily, records that were actually held by us - deeds?, Crown estate records, private collections. So there's tens of thousands of records, which actually overlap between the different systems. But you have to go to different systems, there was a project to bring those in. We're increasingly finding that..

200. Some features are as they are rather by accident than design.

201. We do have APIs, it's quite a basic API. I don't think tags are part of API yet. There's no reason, why they shouldn't be. We've just come out with the new one. The other one will be retired, because we completely redid our schema. There were data elements... The old API was purely for National Archives' data, so even though we had this finding archives and other archives' data brought into it, it didn't use the same schema. Why? Don't know. In the new API everything is brought into the single schema. And therefore the new API will work for any data source that's been added to Discovery. And what we've done is just tag data for it. But there's nothing that stops us doing it.

202. But are there ways to go forward with new cataloguing? Can we actually introduce best practices for people to annotate, as they catalogue can they load things up in a way that allows things machine readable. And if its

machine readable, can we extract it out? So the data in the moment the, the data has loaded in Discovery that's pulled out by the API is a raw data, so includes all the markup and the labels, so they could scrape everything and identify those that have been marked up with people or places and then do something more with it.

203. We do have a space to play. Tags as example, it probably was an experiment to play.

204. And if you think of the WebArchive, if you try searching for it, what you haven't got is... these are unique hits, if you want all hits click here. So the moment you get all hits, these are all the same pages from my point of view, because this page didn't change over 5 snapshots. So even for a machine, how do you make sense of all this clutter even with digital records. An e-mail chain? that gets saved a dozen of times, you do a search, I get dozen of hits, they all look the same to me, how do you get the new ones, the one that's slightly different from all the others that are identical. So those are going to be the digital challenges for humans and machines to make sense of all this. (Grannum, 2017)

205. But that seemed a way of? us not really having to move that material linked, which again would have caused enormous headaches around the building about... the propriety of having that content alongside the content of archivists, which is really ludicrous debate to have, because you can obviously have them in the same system next to each other and it's not difficult for people to distinguish between a rather ? description by archivist and something that's written by somebody else. Those two pieces of content can go perfectly happily, but at a time that was felt to be a rather stressful situation.

206. There's a sense that the tags that have been gathered are not useful, because they're too personal in that sense that they are, too unique to the preoccupations of every individual the researcher. I find that very difficult to believe basically on mass, I know there are a lots of guys like that, but there are also lots of other valuable tags in there. I think we should be ? them and using them as subject descriptions. We have talked about that more recently, but no one's done that work. So I think we're missing the trick on both cats... And if these two have been brought together, it probably becomes a bit circular, because the tags were perceived to be of low quality. That removed any inputters towards actually let's think hard now, how we do move them into the system. Which is as I say is a source of some controversy, where the user-generated content should be appearing. So obviously they appear sort of side by side in that way. But the fact that the systems aren't properly integrated really means that they appear closer than they really are and I would like them see promoted into a.. I like a parallel description a sort of say. After all in some cases the tags are being produced by someone who read the document from cover to cover and the catalogue description may well be written by somebody who is advanced at it, they catalogued it and on they go. That

person may well know more than the person who has written the description, maybe on the position to take those terms as we can. And once they appear in alongside, I think you would expect to see more of a kind of Flickr type of approach to tagging. Because they'll be integrated in a way that they're a bit more familiar to people, oh yes I see that document of spoons is tagged with a lots of different kind of spoon and next time I read a document about spoons maybe I'll...

207. I applied joining Flickr Commons, which again at the time was a big deal. It was guaranteed to massively increase your traffic. It was popular and posed for a while and Flickr was going through some problems, because it had been bought by Yahoo... [...] But we find an agreement with them, with the Commons and then the traffic went very significantly higher, which I was expecting, because there was so much promotion then done through the Flickr website about that.

208. And then there was a project based on Commonwealth Photographic Library, which was accessioned here. So the photographs were just come in and it was decided to digitise the whole lot. Then we had to decide what to do with them. And it didn't seem sensible or I don't remember the resource being offered actually. I think it was very much, where we would put them, obviously we're not gonna build something. I don't remember anybody suggesting that we should build something. But looking at the alternatives, even by that stage Flickr wasn't the draw that it had been. And today, if you might say where to put, if you're doing a large scale digitisation project and you want to put them online somewhere, it's hard to know where you would put them. Because I would thought that strong competitive Flickr would sort of come up the reason with anything quite like that. At the same time Flickr doesn't have the same traffic and interest and cash that they had when it seemed like a good idea back in sort of 2000...6/7

209. I've run projects on Wikicommons as well, done some nice collaborations with Wikimedia UK, I still do quite a lot of work with them, with Wikimedia. But the images did find their way there in the end, quite a lot of them actually are on Commons. First tranche was the African tranche that with interest in the Wikipedia community. And again we signed an agreement with the Africa project and a lot of them are on there.

210. And as I said it wasn't fairly? satisfactory from the point of view of the people doing that community engagement. It is alright, but it wasn't actually ideal for them. But I'm still not sure what we could have done that would have worked better that didn't involve building a new system from a scratch.

211. What we then had was a large amount of content, it's always there, we can always go to it. If I'm looking for an image, I might well remember that it's live on there and I can go and find it. We don't have an internal system that works as well as that. Or it doesn't work well for me, maybe it works well for some people in other parts of the building, but…

212. The people who were involved in Flickr at other organisations. I think there was a group actually... there was a Flickr Commons mailinglist.

213. And the conversations that we then had what was gonna happen with the FCO Library content... From my point of view those were very technical conversations. I was mainly involved in a sort of... they have decided they were running projects in a particular way with the community volunteers and they just wanted a technological solution that the participants and those groups could see the stuff. And what did I have to make it possible, I just went well, I've looked around and I think this is gonna be the easiest for us.
(Pugh, 2017)

Subject-Tools

214. We did a lot of prep work before we uploaded the images. What we felt was, it would be important to add tags to the images we already had about the books. So when we uploaded an image, we also uploaded a lot of tags to every image. That's we we have 95% of tags attributed to us. So the tags are taken from the metadata from the books. … So the images came with the tags but the tags were not about the image, they were about the book that the image came from.

215. There are many organisations that have collections but don't have time or money to pay for cataloguers to catalogue them. That could help a lot in terms of improving discoverability.

216. What that means is that what we've learned from that project is that perhaps it would be worth investing in re-OCRign those images, because that would probably help discoverability. But the problem with that is that it costs money. We need to understand where that money is gonna come from, because we don't know where that's gonna come from. It's gonna be significant amounts of money.

217. What we also did, we worked with an Australian company. And they don't re-OCR. They do something called OCR correction. And what they do, they look at sort of typical mistakes around that computers make and they have libraries of like spell checking. They got very good results as well. So we now have a dilemma. My gut feeling is, just logically effort to OCR first and then clean up again with OCR correction.

218. I think this is a problem for all institutions and where we are right now is that we know humans are still the best at doing this. The issue is the time, resources and money to do this and I think there are certain special situations where you can use a combination of computational and human efforts. But I think you need both, you are probably going to need both for a long long time.

219. The outcomes were, we were able to throw lassos around the content, but what we realised was that we had some initial findings, but it's almost like another PhD to work on, to actually go through that data. We have some promising result, but I think what we've realised is that we don't want to mislead people. What we realise is that there's still a lot of human effort

required. I think what we've done, we've thrown the lasso, and researcher is still looking through that data. Some of it are speeches, a lot of it is not. So what we would probably have to do is to go back again and write more code, which we haven't got time to do. We would need to get funding, it would be like a proper funded project.

220. So because we got funded we were able to do these experiments. And importantly we were given the opportunity to fail. Because actually that's when you learn your biggest lessons. So experimentations are all about, not just about success or about fails, but not being afraid to try. I think that culture maybe doesn't exist so much. I think people kind of have a focus on perfectionism in libraries. They're not going to release things unless they're perfect. Of course we realise that nothing is perfect in life. So actually just our philosophy has been trying to create the examples.

221. They are iterative actually, but they are in a very short time window. The philosophy we've developed is really fail fast. Ben calls it building bridges, but not computably engineered bridges, but bridges made of stone and wood. So what we are able to do is very quickly understand and look at the problem, but we deliberately haven't got much time, we don't have that much resource to be honest. What we're able to do, we're able to understand, if a problem is a big problem or a small problem. That's probably, what we can do. And if it's a big problem, it's like OK, that's a big problem. Where's the money gonna come from. I think because we were externally funded, we are in luxurious position, to change that in the library, we would need to put core money into the library's budget to make this as a business as usual. We give small amount of our time to do these things. I think it's absolutely critical that libraries experiment, that all GLAM sector experiment. But I think you should experiment internally, you should experiment externally as well.

222. We've tried to educate some of our staff to say, look, actually it's better to get the data out there. It'll be great to build the infrastructure, but that will take time. Problem that we had in our project, we didn't have that time. So we had to make a critical time-based decisions on... The philosophy we adopt is let's just get there as much as we can get out there, doesn't matter, if it's not perfect, just get it out there and that's what we've done. So we've used things like embrace dirty data, just accept it, don't worry too much about it.

223. We initially approached Wikimedia, they refused, because there was not enough metadata about the images themselves. We did have catalogue records for the books - i.e. where the images came from - but not about the images themselves. When Wikimedia said no, we looked at the possibility of whether our IT infrastructure would support something like this. And because of the limitations of the time of the project, we felt like .. plugging into the infrastructure the of BL would not match with our own timescales. We thought committee would meet for a long time, maybe many years and maybe not come to ultimate conclusions, as we needed a fast solution.

224. But we don't have that any more. The links don't work any more because we got that virtual infrastructure from Microsoft, we got a grant from MS research. We actually have now a service on a different platform. Microsoft gave it for free, they wanted us to buy it [...]. We had it about a year but we didn't have [...] dollars. So what we have done is, we just bought our own virtual machine from a company called Mythic Beasts, who are based in Cambridge. And actually we had issues around their technology and Microsoft research, so we kind of just bought much smaller scale virtual machine that does all the codes that Ben writes, it's kind of sitting on a new machine. What we haven't had time to do is to change these links. We could provide new links, update these. … Sometimes we have a link to the viewer.
225. We are not naive to think that we should invest all of our efforts on Flickr, because Flickr is owned by Yahoo that's been taken over by Verizon and that could disappear tomorrow. So we have a copy of everything for our own preservation reasons.
226. We took the images from the books by algorithmically snipping them out from the digitised books. And we put them there [Mechanical Curator; ...]. Some people, some book curators might feel a little bit perturbs by the idea we are taking images, we are almost like ripping books apart and putting them on the internet. I think some curators might feel a little bit anxious about taking these images out of context from a book curation point of view.
227. The biggest problem of the project is that infrastructure is just slow to develop. And we are very agile, rapid projects. So we have to harness other technologies, posted elsewhere to be able to do what we want to do, because of the time. What we can do is using other systems. We can then feedback to our committees to say, here is a really good example why we should be investing into an infrastructure…
228. *During the conversation we had problems in changing files, because Skype usage is not favourited in the library. Alternative suggestion to use Google Hangouts couldn't be accepted either, because it's blocked, too.*
229. What we were trying to do because of the timescale of the [1 million] project... We actually spoke to our IT department and we soon very quickly realised that the timescales are not gonna be quick enough for us. They would probably set up a committee that would have taken up to couple of years and we decided not to do it. We needed very quickly an option to be able to put these images somewhere where we could also show the potential of what happens when you open up contents. What amazing things can happen in terms of getting the crowd involved, people doing great things with it. So this is a really really important story to our project.
(Mahey, 2016, 2017)

230. And like all the datasets at data.bl.uk are.., because they're based on collections, they're always catchy and visional and they might be from a particular digitisation project that for logistical reasons did only the easy stuff

or the stuff that the funder was interested in. [...] We've got a mechanism for sharing tutorial and for explaining the gaps and the scope and the context for the creation of that dataset.

231. We are ready to launch a single volume, so that would be about 200 items. [...] It became possible to do the project finally, because Alex's role is pressed on software development results in the team.

232. It's on set of items that you make available?

It's hundreds of them. So yeah, it'll take a long time to get through them.

233. I think it should be iterative. Internally to be able to look at crowdsourcing, organisationally you need to do work to make sure that it has institution's support and for the project. So I took that lesson and that's probably particular to the library in some ways... Also people are busy, so you need to make sure you are allowing to that. But I think iteration is important because a lot of institutional projects got/cut? a deadline, when the money runs out and you got project and the everyone goes on and does other things and then... just say, hey, would be great if you could do this, and like Yeah would be great, if we've got any money. So I think you should, a nice thing about working with Alex, is working and saying do this, do this, so we've already got like a post-launched list of things that we'd like keep working on. The things you've got to have done to make it good enough to launch and then things we would like working on to make it even better.

234. With the idea that we will evaluate after that first volume, see how it needs to be faced, technical comments of uses, ideas of uses, about participants, about what they can do. Things we might wanna put forward.

235. As software projects are also hard, if you don't need to maintain software for it's going out of date and ... It is maintaining projects can be a big consumer of time. But hopefully if you keep maintaining it oftenly it takes rather than doing a big jump every once in awhile.

236. We're not asking much from other people, so it's easier for them to let go with it, because it's not we have to say put in funding or anything like that. We're using it to help the library to think about workflow and like really boring back end infrastructure stuff, but to make the work with crowdsource started, to make the most of computationally enhanced records that you might get from the student projects where they're tagging, looking at the tags or using entity recognition software. So we need to be able to ingest that kind of materials.

237. If people can't reliably search in a field in the Explore interface, it feels wrong to ask to spend some time adding that information. So we're focusing on things that we know can be usefully put into Explore or there will be consisted datasets that researcher or historian could use.

238. But one of the issues is the software is designed for modern books and these aren't really like, they're bound volumes, but they're not books, they weren't published as set. They were collected over time and then put into a set. So we'd like to be able to point to individual page and the catalogue isn't by

default set up for that kind of thing... So it's one of the impacts of digitisation is that we want to say things that aren't about individual regions of the page, like with the Flickr images, an image might have hundreds of tags... there's not somewhere to store that information about that page. So for the moment I think the information will go back into big record for the volume and not for the page. We'll have information for the page, so they could be done right. But it means ? to be able to find playbills about ? that had there different place. Volumes have different things in there.

239. Because all the volumes are genuinely linked to a specific place and over specific time period. So we're starting with regional locations, if we can find the connections and maybe local library might like to run an event to add tags or describathon or something. We don't have any resources to organise something.
(Ridge, 2017)

240. Move away from Mediawiki platform to one or integrate Mediawiki platform, because it was still quite a stand alone environment, which actually took a lot of resources from a technical point of view than a moderation point of view. [...] There was a business decision to close it, because it actually used more resource than was available. There were technological issues that were occurring, which many people think about benefits realisation? of it. And those resources, who were basically trying to keep it working were being asked to work on other projects. It wasn't their core work, but it creased the point as taking significant chunk of their time and that needs to be diverged to other projects. As it was coming to close.. at the same time Discovery was being created..

241. We've created the new front end and then we have to create those reciprocal links between the old catalogue, the new front end and the external system. So it ended up being.. we had great intentions to do so, but logistically we had other priorities at that time and it just did fit.

242. Now the world has moved on, expectations moved on, technology's moved on, can we change, should we refresh that aspect of it? I'm quite keen to do that and tagging will be one of those to look at.

243. *Why don't you let people tag the other items that are not located here?* That is probably by mistake rather than by design. So Discovery went online with the National Archives data, and then we brought in some other national archives datasets into it. After couple of years, there was a view, a project to bring in other archives' catalogues. So the access to archives with the project we set up and run, probably about 12-15 years ago, where we went to other archives, which didn't have online catalogues, online presence. So we went to about other 400 archives to catalogue their records and make them available, gave them access through our archive portal.
I suppose that part of the project was done independently of what was there already, there were different labels, different fields of science, they have

different functionality. Why? Data was folded in, but I don't think that anyone took a step back to actually looking holistically, what features or functions that we've got alive have we not brought across the new data, so as using new style sheets and everything like that basically. So we were working at to create a single schema and a single star sheet for the main details page, but again we have not got those extra widgets we haven't built-in yet.

244. But what we haven't done, we haven't really got to retirement of the environment. We've put things alive, but we've never ask those questions, if there's a value in it. Because you put energy in something, you get things going rather than no, it was great, we've learned from it, let's stop doing it. We've starting to get to that thought now.

245. But I think the interpretation team, they did work with external communities to say something about it [Flickr collection]. But [...] I don't think they had the time and resource that actually follow it to do anything more with it.

246. But our priorities changed, so Your Archives was closed, the Flickr channels.. which I think... much to that, but they didn't go exactly the way we had anticipated, they were actually quite resource extensive.

247. Issue with digital records is that they are - not from the database point of view, but a digital records' point of view - most the data coming through is closed, so there's not much they can harvest it. Most of the data coming through at the moment is what I call digital paper, it's PDFs, documents, JPEGs. It's not multi channel, multi format data yet. It's not far off, but it's not there at the moment. The datasets we've got from years ago, they just reconstructed CSVs. So you can reconstruct the database and use it, but each field, each form has been created as a separate CSV. So you can rebuild it, but it's not emulated as a database. It's just technology used for done it. And we transferred them 15 years ago or whenever it was.

248. But I still think so much will not be automated, because they have just not been digitised, they have not been put to environment, they're still manual, people still have to physically look at them and they then have the opportunity to say something about it.

(Grannum, 2017)

249. And I'm not sure what the lessons of that are now for today in the sense that it's a little bit difficult to know what you do, if you do have 10 000 images of Africa circa 1860-1960. I'm not sure quite what I would do with them today. I'm fairly clear I wouldn't put them there, but it's not immediately clear to me where I would put them.

250. And there are some reasons, which make that not technically complicated, but sort of archivally complicated. Which I don't get into it as it's in any interest at all... But that was one of the reasons, the tags were kind of proliferated so that we could try and view those two bits back together and

say, this is the place where we are keeping this content. That sort of canonical way.

251. So that was always very powerful and it could be reused for a bunch of different events. We if can just rocking up to other people's hack days and going, oh we've got this great data. We know it was public. And Twitter worked in the same way, ironically we were live tweeting II World War and linking to documents that were already alive again because the documents were published online, it was possible to run a thing and linking to them. If the documents aren't there you can't link to them. The core of digitised content is the necessary precursor to building something exciting and interactive on top of it. We haven't got that, we haven't got anything. You go to a hack day and people are like, what have you got, some stuff in the building, but I can't say that's ... yeah, the publishing have to come first. You have to do something else with it. [...] Like Smithsonian, that collection is so diverse compared to National Archives that is largely governmental organisation. But it's also superficial things like, if we only had that many coloured photographs.

252. It always felt slightly strange, the postings that we.. it was institutional account and I was always very aware, whatever I wrote was posted next to a large logo of the National Archives.

253. What's if we're communicating online as an organisation, how much do we sound like ourselves, to wonder how much as an organisation. So that was a bit of there [on cross-institutional Flickr Commons mailing list].

254. I think TNA built a photo platform, we were just like, why?! Why on earth would anybody do this when this alternative seems to be available? And there were other ill-fated projects in the UK around content sharing that failed. That is something we learned that especially if you are trying things out a light weight free alternative is a much better idea than building your own system.

255. It may simply not be enough of a kind of organic list of active users to make it a viable community in that sense. It doesn't mean that it's not useful and valuable. [...] Your Archives was a brilliant project. The plug was pulled I would say just at the point that the organisation started to wake up to how useful it was to have somewhere a content could be organised that way.

256. In terms of digital engagement I think [...] we're very active across social media, but I think [...] that has become quite routine, we're not actually reaching out perhaps the way that we were a few years ago. That's my perception, that might not be the case.
(Pugh, 2017)

Tools - Objective

257. Then we had an event for it. And because of that event all the people that were interested kind of got to meet each other and then started to kind of obsessively decide they were gonna try to find the images of maps.

258. We have an online shop, would be great we could do something and then use that money to digitise more. That would be great. We have not quite figured out how to do that. I think that would be a fantastic thing to do.

259. I think just a sort of say that we are not the only institution that has this problem, we have so many things in physical and digital form might from the digital side... the information in them is locked away for lots of different reasons and how we can use different types of technology to unlock those, it represents different kinds of problems.

260. And of course in an ideal world we would love to have everything as we have put on Flickr, where people can just access things, every object has a unique URL, it allows different contributions, allows people to be able to tag like they do on Flickr, has an API, so people can write their own programs. We are some way away from that yet to do more of that. But for us what was really important about Flickr was that we have the example. And a really strong example of what is the possibility. We're not saying it's perfect.

261. Maybe people are asked every time they download the image, who they are and what they use it for. Because once it's out there it's very difficult... [...] One of the things that we have feedback here and I think the organisation is called Te Papa in New Zealand. And what they do, before you download you have to fill in the form, say who you are. We're thinking of doing this, because one of the problems we have is that, if we want to inspire people to use our collections, we have to tell them, what we have, which actually isn't straightforward. Secondly, we can then say, how can you find it and then thirdly, what people have done with it already. So we can give them clear examples. What we've tried to do. We've tried to give those examples, because you're right, we expose things on the web and sometimes we don't know what people are doing with it. [..] Probably not everyone would fill that in, but at least you're trying to do something.

262. So I think what we've done before we were doing this, we were sending data to people via hard drives. So this is just a better way to do. The problem we have with this at the moment is that we don’t ask people to tell us that they’ve used it. We ask them it would be great if you could. Since we launched this it we don’t know.

263. I think there's real possibilities to use computers to help discovery, I don't think we're long way before those can be trusted enough to be able to really-really improve discovery.

264. But the problem we have with that scenario again, kind of computational approach is that we need good data at the beginning. We call it ground truth, is there 100% data first. The problem was that we really didn't have 100% data.

265. For me the most important thing is get the data out there, don't worry too much about fancy interfaces for now, because that takes a long time. Just get it out there, as long as there are systems within the library to facilitate that, that's the most important thing. If there are nice platforms like Flickr or

whatever, and they're able to do it very quickly then yes, use them. But I wouldn't suggest trusting on your whole infrastructure on those things. I don't really know the answers to this, all I can suggest is that better to try, also accept that there will be failures in those things.

266. What happened interestingly is... some of the records started on Flickr and some have moved to Wikimedia and some of them are within Wikipedia articles. E.g. you talked about Synoptic Index in your presentation, if you go to Wikimedia, there are pages created, which are all about trying to find the maps in our collections, in the Flickr collections. That was all done with Wikimedia. What we feel is, for us just in our project, we felt that was the better home ultimately, for collaborative crowdsourcing. What's been interesting is even though they refused initially, some of the images have ended back on Wikimedia.

267. We needed a unique URL for every single image and we needed a possibility for every image to be tagged. That was the purpose of putting them on online platform. That was the whole idea to make them more discoverable.

268. These I can see have been automatically created. They were created by the Mechanical Curator. So the Mechanical Curator is the account for probably for the 12 million tags [added by BL while uploading the 1 million images]. These two [gray] are added by humans or computationally by someone else. These [white] are tags added by Flickr itself. The white tags we noticed were using computational methods to tag the images. And their tags are different because they are in white. So these two tags here ‘romance’ and ‘love’ have come outside. Tags from the Mechanical Curator have a colon in somewhere. This tag ‘large’, when we algorithmically cut out images, we had small, medium, large. So we have that information. All of this is only really from the book. For example, if I click on a system number that should give me all the images from that book. These are all the images that we found in that book. It was kind of easy way for us to add all the tags for the images. So if we published this we noticed that the discoverability improved quite dramatically. When you searched on Flickr on this collection, providing links to the plain texts for the book, actually improved the discoverability quite a lot. I think it was available till Sept 2015. We got a grant from Microsoft and we were using a platform Azure. And we put the Mechanical Curator, which is the engine, we used it as a way to capture tags, to put text up, we used it to manage the whole project really.
(Mahey, 2016, 2017)

269. I'm writing at the moment the help pages and we've got kind of 'about' text where it's a type what you see kind of project. We've been through a lot of debates of how to describe a lot of things. Because it is the way that playbills work. It will often advertise a lot of performances on one day and one ticket might let you see it two different plays and exert? to another play and maybe some songs and maybe a lecture. So there's a lot of ... They're not like modern

plays, you have your ticket and it starts at 19:30 or it finishes 10:30 and you're done, so lots of different things on one day. So we're trying to work out how to describe what a performance is, what an event is. So that people know what they're marking up. So I wanna make sure that the instructions are really clear so that people aren't kind of like not sure they're doing the right thing.

270. I think that Flickr is more for photos and actually sometimes I'm like damn you British Library, damn you Internet Archives, because I don't want... because if there's a lot of like?, and illustrative, decorative, I'm actually trying to find photograph and you're just in a way.
(Ridge, 2017)

271. We created links [in Wiki] so an article about a particular record had a link there to take you to our external catalogue. Our external catalogue had a link too back to the Wiki. We had that with several external catalogues, retrospective links. What we had hoped is that as we phased out the old catalogue we would integrate the media wiki to come to seen as more integrated service. [...] But the content, where's the narrative we left.

272. Your Archives has been archived, but we didn't do anything more with that, with the content itself. It's still there if anyone wants it, but it's difficult to search, because it's a web archive, but technology has moved on, search has become better, so maybe one day someone says that yes, let's bring it back and copy it and put it somewhere else, but not at the moment. We wanted to try to preserve some elements of it, we just knew that we can't preserve all elements of it within Discovery.

273. It's not used often [flag a tag function]. It's more used by index spots than by people. So everything's been flagged as inappropriate. Because every time Google or Bing uses their index spot, it's as truthful as flag. We do look at it, but it's not useful.

274. It could be that there's no value [in social tags], I don't know, what the value is. If there is no value, do we need it at all, do we want it at all. If it is mostly noise, so where is the value in that? If it's not used for the end-user to find information. Or e.g. browse, I'm gonna just switch off browse, I mean browse tags, it's in alphabetical order by first letter, folio, nr, page. But that's no good. You're not gonna find anything from browsing. And it used a lot memory by trying to process that page and display it. Let's just use search and that's how people use it.

275. Visit Britain or I'm really keen on getting historical geogazettere out there, because place names has changed, locus of a place moves as a town expands or decreases or whatever. That one has value. And people whether doing research or want to know more about the place - so can we do more about that. But that means, we're trying to do it retrospectively, that's just really challenging.

276. The online catalogue was basically a digital version of a paper catalogue. So for chemistry office ? you need to go there, there and there.

We've done that at least to put it online. You've missed that supported learning. And we've done to go through the different iterations of the paper catalogue. So we try more integrated experience. So the research guides and other useful web resources, education resources about themes and subjects to integrate these more with the records themselves and vice versa.

277. What we found is that majority of our users don't come through our front door, they're coming through third party Google, Bing, Ancestry, so they land in the middle of the site going.. very interesting, what's it telling me? What do I do now? So we want to put some effort over the landing page, where people come into. I think part of that.. unless you get consistency across different data sources, how we change the feel of the site, so people understand where they are, textualise a bit more, maybe a bit more supportive learning.

278. But what we had beforehand, we had a global search - a search across our datasets, including the OCR content and they were swamping in results. They were minority of the records, but the way relevance ranking works is that they were swamping the results. And insignificant file was coming up first, because you were trying to describe a full text as of catalogue description. Full text is always gonna get higher score that the catalogue description, so how do you... We've brought it into Discovery now, there's one we did a few months ago, there's a little tick-box under the advanced search to do a full-text. But it's only of the OCRed content.

279. Human is not going to be extensive several hundred, several thousand pages. What they want is a nice little summary. Humans are kind of curated, so is there a way of systems to make sense of this text to say, here's a wordle?, these are the key themes that came out of this word cloud, key themes that come out of this document. Is that the way we go forward? Because leave machines, they can do what they like with it, but ultimately it's still gonna be humans at this point of time who still wanna try to make sense of it. (Grannum, 2017)

280. Again there was more a kind of functionality problem. You can't comment on images on Wiki Commons. [...]
And the fact that people could talk about them and they could say what they felt about them, was really really important. Commons would have been a perfectly acceptable platform, but as I said Commons is not the place where you can collect what people think and how they feel about some images what they see. That's why we worked in parallel in the end. The content went to both places instead of it could be used in Wikipedia article straight-forwardly and it's indeed out there.

281. Again that's not why I signed up, but I became interested in Flickr API and we did experiments here. [...] We did generate a large number of tags here and again it's an experiment, partly a way to linking the catalogue directly to the images. Which we didn't have initially aware of doing.

282. One of the last things I did here was to run a hack day in... 5 years ago, 2012. And that again, what's so useful about those platforms is that once you got content there, we have got a large amount of content even reused. What we then had was a large amount of content, it's always there, we can always go to it. If I'm looking for an image, I might well remember that it's live on there and I can go and find it. We don't have an internal system that works as well as that. Or it doesn't work well for me, maybe it works well for some people in other parts of the building, but… So that was always very powerful and it could be reused for a bunch of different events. We if can just rocking up to other people's hack days and going, oh we've got this great data. We know it was public. And Twitter worked in the same way, ironically we were live tweeting II World War and linking to documents that were already alive again because the documents were published online, it was possible to run a thing and linking to them. If the documents aren't there you can't link to them. The core of digitised content is the necessary precursor to building something exciting and interactive on top of it.

283. I mean people talk great stories. But they weren't necessarily the kind of things you could move into a catalogue. And that in a way was one of the reasons for wanting to have strong links then between the digitised material, which has then got interesting user-generated content attached to it. You want the archival catalogue to link to all of that, because they all are quite interesting. Actually in some case for research might be very very interesting.

284. And I'm not sure to begin with we were sure that we did have the technical capability to just call that all in [stories by users]. But I'm sure the API now would make that increasingly straight-forward thing to do.

285. *So it's all about making materials visible first and then discoverable maybe?*

Yes, again in a very high traffic environment you don't have to worry so much about that. And to make something stand out and you don't have to worry about that then frankly.

286. They are very different [catalogue and Flickr] and the way, which tags are presented is different. You can't tag here again unless you have signed up for an account. And I think people are not necessarily aware that their tags can be seen by everybody. The tag functionality as it's presented more widely is quite kind of hidden away. So I think it's easy for people to think it's kind of personal thing.

287. There are two systems that have never been integrated together. So you still can't, if you carry out a search in Discovery you're not searching the tags, they've never brought together.

288. In terms of should we be more interested in user-generated content around our collections than we are, I think we definitely should. There's enormous opportunities there for us to work with all kinds of people, from professional historians to anybody, who's interested to learn more about our collections. But you need a lot of flexibility around presentation to do that.

And our systems are rather lumbering, which is why you then tend to put.. why I looked it's such a pain. It's the same again. We put the content somewhere else and see what happens.
(Pugh, 2017)

Community

289. They range from individuals, I don't know how to call them - independent scholars, artists, entrepreneurs, and to other memory organisations, commercial organisations, charities, small charities, large charities. It's quite a random group, I would say that even though our primary audience was universities, I would say there's probably more interest in these other... And I would say GLAM for sure - galleries, libraries, archives, museums, that's an obvious one.

290. I would say learn from all the different people, who are doing it. So Rijksmuseum is doing some interesting things about reselling high-resolution images.

291. Actually what should we be doing, we should work cross-institutionally. That's what I think we should be doing. I would love it, when people would be using an image from here, from here, a text from here and creating something altogether around all those things. I think that would be fantastic, especially public domain stuff. That would be amazing. I think it's still early days. The institutions themselves are still figuring things out.

292. What we've discovered is just in the UK, when we've gone out to people, about our primary target audiences are UK universities. Although I would say more organisations outside universities have engaged with us. In 2016 we had over 500 requests to use our data, doesn't sound like a big number, but if you look at the statistics by university, the average is about 5 requests by university.

293. We're research library, our primary audience is scholars, but actually a lot of examples come more from artists.

294. So they [Wikimedian in residence] are residents in a memory organisation and they help to promote Wikipedia, Wikimedia, they've also helped with materials online. We have quite a few things on... I think Wikimedia/Wikipedia would really make an interesting case study. It'd be a different platform from Flickr for example... You would find quite a few GLAM organisations using Wikimedia to put there some of their materials. And the dynamic in tagging is a little bit different.

295. With this case of food, you expected that the participants would come with a set of skills that they already have? Or you train them?
No, they don't. That's why we have connection with wikimedian and Wikipedia. So they come and they train. What happened, we started with a very small group, they came again, so they help some of the others. But we also had people from Wikimedia and they do the training. Anybody who was new, they would go away and get trained, the people already experienced, they

started to create new entries. It's just like sowing seed, you put it in, you look after it and it grows. And I think there was something in your presentation today. I think, institutions once they realise that, if you give people the option to do tagging, they're not going to vandalise and create silly entries. The reality is not a huge group of people.

296. People who contribute to Wikimedia Commons and Wikipedia tend to be geeky white men and it's trying to encourage especially more women to get involved. [...] She [library curator] has invited women who are like food writers, food critics to come to the British Library. It's not just, you know it's not like women only can come, but the majority of people are women.

297. To be honest, this whole concept of crowdsourcing I think is a bit rubbish. The reality is, for most people, they do it once and never come back again. Our experience of working with people is that that is more niche sourcing - a very very small group of people do the majority of the work. That has been our experience. People and libraries have been talking about crowdsourcing and things like this, but the reality is that most people do it once and never come back again. [...] That's a big problem actually.

298. The whole thing about the Flickr images kind of came from a number of angles actually. We ran a hack event in July, where we took along the digitised books on a big storage device. There was a guy, a researcher called Matt Prior, who actually came on the idea of doing interesting things with the images. [...] There were other researchers, there was a researcher from Cardiff, who wanted to do some of these things. I think it wasn't the first time the idea have been thought of, to be honest. Matt actually submitted an idea to our competition[100]. He didn't actually win, but we liked the idea. And I'm not sure exactly how it all happened. We decided to run our own experiments. So what Ben did, he was interested to see, if we take the images out of the books and whether we could run face recognition software on those images to see, if we could identify the faces.

299. I think the main contributors are quite special. It's not just their task to tag thousand images today.. There must be really interesting reason why they do it. [...] There's a whole network of people that were involved in. It's not transparent to you. We know the top taggers in Flickr. FlickrIDs are IDs for you, but humans for us who we work with.

300. Some of the smaller contributors, people who have done thousands, 4-5000, not sure how interesting those are, but specially the top 5 .. We have relations with these people and some of them don't mind actually being contacted and telling their story, how they have been tagging.

301. [...] he used manual techniques, that's quite impressive. [...] was the first person I spoke to get the images onto Wikimedia Commons.

302. []... used to be my old boss.

303. [...] sits about 5 meters from me.

[100] http://labs.bl.uk/Emergence+-+Distribution+Graphical+Content+in+19thC+Books

304. We know Mario[101] really well, he's been tagging most of the images using computational methods, he calls himself a code-artist, he writes code to create art works[102], but also to automatically classify some of our images. Mario Klingemann did quite a bit of work in the early days to identify maps. It's a funny story actually. We were organising a hackathon at the BL, because some of the volunteers were asking us, they wanted to find all the maps. So we thought okay, let's organise an event and it was on Halloween 2014[103]. About 2 weeks before the event we got an e-mail from Mario, saying I don't think you need the event, because I found all your maps. So Mario tagged about 22 000 maps just using computational techniques.

305. The second person [...] is a human like Mario, but he has done everything manually. He is 75 years old, he lives in [USA], he is disabled [...], he sent us a beautiful e-mail about 1,5 years ago. I know who he is, but he does not want any publicity. [He] has been doing everything manually and for about 1,5 years he was ahead of all the computational taggers. [...] I wanted to give him a gift and public acknowledgement, but he wants to remain anonymous.

I was sending [*him*] lots of e-mails by Flickr-mail, because that was the only way I could contact him and I gave up. All I wanted to do, our symposium was coming up in 2014. I could see the incredible work he had done and I wanted to kind of acknowledge him in some way. Maybe sending him a book token or something for the incredible work he has done. And the night before the symposium about 3 in the morning I got e-mail from him via Flickr mail. And it was really moving actually, reading it, because he told me about his personal story, but he also told he did not want acknowledgement.

306. The next guy called James Heald[104], he's been doing all the maps. And he used a combination of human and machine learning methods to tag maps. Not just to tag 'map', but location 'africa', 'mali' and so forth.

So he's been going back to Flickr API and adding map information. He is British, a physicist, he works at the university. He does this in his spare time. He is a brilliant guy actually and a pretty nice guy as well. James uses lot of computational techniques, as you can see by the tags you used. [...] Google Fusion table[105] has all the data for the georeferencing. What happens is that when people georeference the images, the maps that data is put into Google Fusion Table and from the fusion table that information is put back into Flickr. And all that work is done by James Heald and he's a volunteer. And that's

---

101 http://blogs.bl.uk/digital-scholarship/2015/11/british-library-labs-awards-2015-the-winners-and-runners-up-announced-at-the-symposium.html

102 https://www.flickr.com/photos/quasimondo/sets/72157638820730895

103 https://wikimedia.org.uk/wiki/Digital_maps_Halloween_tagathon,_October_2014

104 http://blogs.bl.uk/digital-scholarship/2015/11/british-library-labs-awards-2015-the-winners-and-runners-up-announced-at-the-symposium.html

105 https://fusiontables.google.com/DataSource?docid=1BMm0FeSsEBa40zgs3C3vySKC0gnPk-pSvrDqqnA7&pli=1#rows:id=1

why he won an award. I think there were couple of people involved in the discussions. But what's great is this… which is a really detailed way to discover content in the books by geographical location. It's called the synoptic index[106]. And I think it's a really nice way to discover stuff. Because you can go by place into the books. And we are still talking about images in books, map in a book. You see georeferencer information has been added and then that information is then being put back onto Flickr as tags.

307. Community is probably those engaged with maps. We've engaged with them. Some of them are community, a mapping community. Because they would have come to the [Halloween] event. [...] James will have an idea of community. I think it's fair to say that James led the community to find the maps.

308. We did recently some post with our top georeferencer [...] Maurice Nicholson[107]. Maurice worked with James. He's been helping to georeference the maps that we found on the Flickr site. He actually also helped to identify the maps in the first place. But I don't think he is very high on the top list.

309. I saw a really nice example actually amongst our illuminated manuscripts. That collection, the digitised collection is the most popular collection that we have of all our collections. Somebody told me about selfies with illuminated manuscripts, especially figures. So what people would do, they would take a picture of themselves in the same pose and sometimes wearing the same cloths as the original image. I just thought it was a really lovely way to sort of engage with our collections, if you really imagine to in a different way... It went through Facebook. [Initiated by] some manuscript geek, who is like really... It's really cool, it's very funny. I just think the library gets some publicity, it's fun, it makes people simile. Just a really different way to engage with our collections.

310. He is very interested to try to find emotion in images. Certain things like he found 19 tragic looking women in the collection. He trained the computer to find them. He also found something like 300 images where there is a hat on the ground. He kind of makes funny collections for discoverability. The hat on the ground means trouble, that's useful and so funny. He discovered that. Obviously in the 19th century people used illustrations as a way to communicate something. And hat on the ground tells trouble, meaning that every image that he found where there is a hat on the ground something bad is just about to happen. Somebody is about to get killed, a man is just about to leave his wife.

311. Last year I met an amazing woman and she has spent 30 years building up a huge database of food recipes in Europe. She was going to the Oxford

---

106 https://commons.wikimedia.org/wiki/Commons:British_Library/Mechanical_Curator_collection/Synoptic_index

107 http://blogs.bl.uk/digital-scholarship/2016/10/maurice-nicholson-british-library-flickr-commons-map-tagger-and-top-georeferencer.html

food symposium. So she has gone in [at Wiki Food Editahon...]. I think there's a 130 000 recipes and she's been creating this database for years and she became last year... she's in her 80s. She's American. She was asking some advice, how to make it available online. I think she started with spreadsheets, then went into Microsoft Access database, now it's got too big for Microsoft Access. The idea was how could we make that available.

312. A historian looking for speeches about slavery in the 19th century in our archives. That's super interesting as well. [...] There was a movement in 19th century called Black Abolitionists. These were black men and women who escaped slavery from America, came to Europe and they were trying to give speeches about why slavery was evil and why the British government should put pressure on Americans to stop it. And their speeches were captured in the newspapers. The idea was could we find the speeches, how can we find the speeches. Very difficult. [...] Well it's kind of different problem, looking how to find things in that muddy OCR text. Which is again using machine learning and computational techniques, but it's a really hard problem, because how do you find things in a "rubbish bin". We're having some success. The point is that when the computer makes the texts effectively invisible you are hiding history. It's history that's hidden away that people don't know about. Because if you try to find out on a computer on Google, it won't show. Because the word 'dog' may have been OCRd as a '?og' and that is not 'dog'. If I was a human and I saw it on screen I could see it was a word 'dog'.

313. We work with a group in Stanford. They are going to tag every single image using machine learning, they are using computer algorithms. They are going to tag every image with one of twelve tags. They have been working with our curators to decide which 12 tags. They are going to computationally tag every single image to improve discoverability.

314. For example we're working with somebody from Finland at the moment, who is looking to find images around Finland for the 100th anniversary of the independence which is next year.

315. But also the local community right next to us, the local community called summer span? - there's a high degree of poverty, in that community. And I think the library wasn't always great in engaging that community. That's changed recently. We have a community engagement manager, who tries to get the local community to engage with the library. The library is literally on their doorstep. The library took 20 years to build. So they would have had this headache to see this building been created. There's been much more work in trying to get the local community to engage with the library. And I think for many they see the library as a kind of snobbish institution that is not for them. Actually we have a really important part of our strategy that we need to engage with our local community. There's only one person doing that, but luckily I know her really well and I've worked on some projects with her, it's just because it's my own personal passion to engage. Because that's my own background, I come from sort of very poor background. So I completely

associate with that. So what I try to do is to develop this philosophy in Labs. Some of the work we do is quite technical, not always, but some of it. What we try to do is… so we are finding things in messy things, once we find them we want to celebrate them. So we've done quite a lot around performance, artistic things, cultural engagement.

316. There's kind of perception that crowdsourced data is poor. And actually that's not true, because the reality is that the whole group of people, who do the majority of the work produced really excellent quality tagging. I'm sure [...] was a librarian or something, because the quality of his tags are superb. They're really detailed. He knows a lot about bibles. So I don't know is he religious…

317. I think it's the topic, the map tagathon was maps, most people were obsessed with maps, food-food. What I would love to do and I still haven't done, I want to do is to engage with fashion industry. I would love to get fashion students to tag for example Flickr images, how they might be used in a fashion context. I think there's real potential there.
(Mahey, 2016, 2017)

318. We were hoping to connect with local libraries and local historical societies and maybe even the theatres as some of them are still standing.

319. What we're doing over the next few weeks is looking to community who already working or using the kind of sources. There's a few special collections at different universities that have collections of playbills. So we're going to talk to them about do they know anyone who would be interested, can we connect with them some way. We're talking to V&A?, because they now have the museum of theater, which was absorbed by V&A? So we knew that there was some really keen theatre historians, theatre curators, so we're talking to them. I'm assuming there are communities out there who would be interested.
(Ridge, 2017)

320. [Government departments] use Discovery in a different way than the researcher use it. They want to know, what was transferred, when, how it was described. [...] So they're not using it for research, they're trying to use it for a context point of view.

321. [Archives] are one of our users. Not just because you have other archival content in it, but also from best practice point of view, they can learn from us, we can learn from them.

322. But also the contributing thing for the end-user, they're not doing it because they love doing it, they're doing it, because they're paid doing it, they're doing it, because someone is benefitting from it out there.

323. But the idea of actually share that information was quite alien then, which I don't think is as alien now as it was then.

324. When we ran the Wiki, for academics, they were willing to scrape the information but they weren't willing to share information. [...] I suppose many, they wanna be the first to publish, they wanna be the first to say something, to find something new that no one else has necessarily found before. Or if finding things themselves, they don't want to share it yet, but they may share it later through dissertations or articles or whatever. But then going back to enter it into another system retrospectively is not gonna happen. People happen to do things forward, but to do things go back.. You just don't, your time is just too precious to go - I'll share that or I'll let other people to know about it. Unless it'll become live again, then actually people start to work collaboratively.

325. *Third parties, businesses, other archives? Would they trust archivists more?*

It's just archives more. If I use the example with the wiki. Even if you weren't be able to bring the data into the catalogue, the push was to have it in a separate domain. If this is curated content, this is crowdsourced content. It's different. This is been checked, validated, trusted and approved - tick! It's the Wikipedia argument, how true or authoritative is Wikipedia? The most of the articles there are truthful, trust-worthy. Still rings bells, we here academics to say, we tell the students do not use Wikipedia. Why? Surely the source you should encourage them to verify to double-check, to update or whatever. But "don't use that"...

From my experience with wiki, I don't see why you shouldn't trust the content there. Those people are honest and trustworthy, they may misunderstand, they may misinterpret, but that's no different than the professional, who has tried to summarize, what's in the document. A little new ones there, they might not understand what the author meant. Can have a completely different meaning when they write it, when they summarise it, when they produce a catalogue entry. So they themselves have to use temptation and judgment over it. How qualified, how trained are they any more than ordinary member of public, who wants to say something about that document. I know there are tensions there, but I don't think they've found it, personally.

326. *What do you think, where does it come from that you have so many taggers?*

No idea. We've never even studied it. It just happened. We just haven't had the space, been focused on projects, so we haven't had the space to step back and think about.. But I think this year we've been start thinking about. [...] Because we don't manage it, we just sort of left it. I hadn't realised that so many people have tagged, that's different matter.

(Grannum, 2017)

327. They were just people who were interested in the content from all over the world. So that was the great thing from the kind of outreach point of view and from the project point of view as well it was to be able to publish large

number of images from say Uganda and have people from Uganda come online and say, oh these are really interesting, oh I'm gonna show these to my grandmother. So that was really really fantastic for us.

328. It mixed in with the people, who used to live there and moved away, some of them were in UK, it [Flickr] was very international site.

329. A really significant proportion of our digital audiences is from abroad. And quite a bit of that comes from America, Australia, in a countries with whom the UK has got strong historical links. [...] There are always be significant interest globally in the materials for a bunch of different reasons.

330. All the contributors at that point, can't remember how many there were, but not that many, not many more than a dozen I think. And just... there were loads of people looking at all your stuff. I mean it really was as simple as that.

(Pugh, 2017)

Community - Subject

331. So actually just our philosophy has been trying to create the examples. And then inspire other staff to say, actually look, you can do this, too. Or we can help you do this.

332. What I try to do, I try to communicate, what we do internally, so we have a newsletter, events. E.g. today's event is all about what people are doing with our data. So you've been doing things with our data. So you're a perfect example actually. You've taken our Flickr data and other people's data and you work on that, it's part of your PhD. So you are a perfect example to show what people can do. That's why people are interested. We run a lot of internal events, but I think it's a large organisation, we have maybe 16 000 employees. It's a lot of work. We have 3 sites, this is the main site, we also have another site about 150 miles away, 200 km. Much more work is needed, I think.

333. And what happened because of this attention from the media, other people in the library wanted to put their collections next to ours.

334. The long term strategy, the most beautiful strategy would be to get that data back to our discovery systems. And that is a whole political discussion that needs to take place. We would love that, but the problem we had is that there is a lot of politics of the quality of cataloguing.. The problem is, a lot of cataloguers, if you tell them their data isn't that clean as they think it is, they don't really like it. If there's a spelling mistake or there's an extra space, if you try to do things computationally, it just won't work. E.g. if you have a date which is 1922? - computer would not understand that. A lot of the data is like that because it's created by humans. There's also sensitivities about the quality of the data.

335. What we have this project is a perfect story to be able to go to our management and say we should have open collections we should allow crowdsourcing of tags, every image should have an URL, we should have an API. These are all arguments that we are taking back to our management and

stay that this is the kinds of systems we should support. That was one of the reasons we did this.

336. There are gaps in the data we have collected via FlickrAPI. Flickr have analysis tools, they can do analysis for your account, but we've never gone down that route. It makes me feel, maybe it's time to talk to them.

337. We expose things on the web and sometimes we don't know what people are doing with it. [...] The way we try to solve that problem is we've... I don't say we've solved it, we've done two things actually. One, we've run something called British Library Labs Awards[108], where we say, if you've done something with our collections, tell us about it and you might get some money and some fame and glory.

338. Maybe this interaction with volunteers has been by a little bit of accident. But I think maybe we just need to start thinking a little bit smartly, how to make more of that. I'm not aware of any specific campaign initially to get people to... E.g. Discovery - how were people encouraged to tag? - *Not at all.*

Exactly, imagine, if they did. Imagine the difference. Imagine, if you gave people some kind of reward. I don't know what the award would be, but some acknowledgement even. There's so much more in that. What you're discovering is the early days. As organisations become more comfortable with users engaging this way, I think it's fantastic.

339. So random example, what we have to do in order to get this message out, we have to go out to people. These things are not available on Google, people can't find them, they're not easy to find. So we need to go out and tell people. And that's lot of hard work.

But why not make them accessible via...?

We are trying to make them accessible like this and I agree with you completely, we should make them more accessible, but..

You mean indexing by Google?

No, more and more that we have available, we are trying to put out there. But I still feel that for now a lot of people don't know, what we have. A lot of people in the library don't know, what we have.

340. I think that the dialogue that we have with academics and researchers, and I know this from personal experience, we go to a room like this with 20 academics in there and then we tell them about the library and what we have physically and digitally and we have to explain the following, that we have to manage their expectations, because they assume everything that we have is digitised. And it's not. A very small percentage is digitised. So the metaphor I use is like, if you want to use our physical things, it's a bit like going to a giant hypermarket. We have quite most things. But if you want to have our digital collections, it's more like a boutique sweet shop, because... If you're looking for something salty we might not have it, because it's a sweet shop, right. The

[108] http://labs.bl.uk/British+Library+Labs+Awards

problem we have that a lot of academics, unless we have something specifically in their area of their academic interest, they're not interested, especially, if it's digital. So if we're looking for 12th century manuscript digitised the chances are that we don't have it, because so little of our physical holdings are digitised, it's actually about 2%, a very small percentage. Still a big number.

And a problem we have with digital about 10% of our collections are actually online as of the digital. So we have about 687 collections as of today, 15% of them are openly licensed, 10% of them are online. So there's so much more work to do be done.

341. For example somebody found an image and they wanted to make a giant poster of that image, the resolution we released that 300dpi, probably wouldn't be good enough. So they could request to have it re-digitized. This was a very deliberate use. When we published images with tags from the books we also provided links back to our systems, links back to our catalogue, viewer.

342. Most of them [participants of Labs competition] are artists, businesses, they all pitched ideas. We then... they get some money, not huge amounts of money, but they also get our resources and for 5-6 months. So they get a lot of Ben's time, who you met today and basically we work with them to try to solve a problem.

343. What we are doing by putting the data out there is creating the potential for them to be reused. We may be lucky and people might do that anyway, but my experience tells me actually you have to tell people, you have to make some effort. So we go out a lot, we have to tell people, it's... In a group of 20 people, maybe 2 or 3 people will understand and think the idea of what to do with our collections. But that's how it starts.

344. Some of them we've contacted ourselves through Flickr, we've e-mailed them through Flickr-mail and said Great work! They've contacted us. So a lot of them that work has actually resulted from this, this has been through dialogue.

345. There's a psychology of working with volunteers, because you want them to feel special. We spend a long time with them, bringing them to the library, taking them to lunch, giving them presents, making some fuss of them. [...] Because it's important to recognize the efforts, these are just volunteers, we did not ask them to do it. They've done it from their own passion and it's important to reward them. I tried to reward [...], but he doesn't want anything. They all have different stories.

346. A lot of them [participants of Wiki hackathon], what they do is they just do that, take a picture and then do it afterwards. Not sure you're officially supposed to do that, but they do it.

347. We need to sort of articulate better, what's in it for them, if they contribute to. I think it's important to us, that if we use third party platforms that we provide links back to our original source in some way. I think people

do value that, but I think if we don't do that then we loose the connection to the library. I think people do perceive the quality, I'm not sure I've got the evidence to prove that to you, I'd say it's a guess. I'd say probably they do perceive there's better quality, but I think people do not necessarily care as long as they can have access to the content.
(Mahey, 2016, 2017)

348. *How do you reach the audiences?*
That's why we're talking to like libraries and like historical societies and... Trying to work out at the moment how to approach them [potential user communities].

349. Trying to work out at the moment how to approach them, so that they understand, what is it that we're offering or asking and what kind of relationship we wanna have with them. So we wanna get the basic interface sorted, so that they can see it and understand what this project is and sort of explaining this abstract concept to them.

350. So probably what I've done is look a bit of that project and thought about how to make sure that people are engaged and know what we're doing for and who might be interested in it. So that they could support us and keep people engaged. And I plan to spend my time to look out for updates and communication and... so it's not just hoping it's gonna happen, but making sure it's gonna happen.

351. So one of the challenges has been making sure that enough people in the library are engaged. If we have a question, the curator knows why and suddenly saying like can you tell me what happened in 1797. They'll know and they'll know it has to be any answers actually written purely.... So it's kind of carrying to lay the groundwork for you, wider engagement from the subject specialist in the library to support the community. They find something interesting or they have questions. I don't know how actively the form has been used, but it's only way to find out.

352. *Do you keep also your eye on some hosted projects or some external projects that uses your data or your materials?*
It's hard to, unless people tell us it's hard to know about. I know 19th century papers, 19th century books are really heavily used, partly because they're in good condition and relatively OCRable and not copyrighted and no data protection.

353. And I liked the focus of the Smithsonian transcription center, they tailor the interface to different sources or sets of different area. And then spend a lot of time in social media in encouraging people. Clearly value the expertise of the participants ? to the project. They don't mind that their time has split between other projects as well as on their own projects. Because it's more holistic view of what participants would do. It's quite generous.

354. But what I'm talking about crowdsourcing a lot of time the key is ethics. Asking people to spend their time doing something and we're all busy

and time is valuable. So we should be sure that if we're asking people spend time on something that we're taking care of the information that's been contributed... that it just doesn't just live in systems, because if someone leaves or some hosting expires and you lose that information or if you upgrade your systems you loose information or whatever.

355. *About playbills project here, do you plan to award top contributors anyhow?*
Well, they're partly going to be led by...? So if [...] their region of the area and we may be doing something with local institution. If they're international it can be trickier.

356. In a way Labs projects are kind of co-created when they do the competition and specific scholars can come and work with them. I think that competition structure, where someone has six months working with Ben and Mahendra is a realistic estimate of how much time co-creation takes.

357. And that also makes technical time and also community time and time to feedback into these, institutional feedback to the community about what's happened, to say, hey, first volume is in Explore! If you search this you find this stuff that you've done.

358. One thing I know is that as soon as you launch a project, people surprise you. You just see, what happens.
(Ridge, 2017)

359. [The data] can be used by different people on a different ways. Just because a end-user, who also is a bulk of our users, the public user, just because they don't use it, doesn't mean any other groups don't find value in it.

360. We never have the luxury to describe everything down to the level of detail that people might want. [...] So how you balance out the different needs for different users?

361. If you think of archives in a whole, most records aren't looked at, and most that are looked at, most of them probably are looked at one or two times. If you think of a will, someone who is doing a social history of Proback? records, someone looking at the social history of that village or that town or trying to understand academic? structure of tailors or sailors or whatever, might do a holistic ? of candle makers, who died through this period from that town. But who else is going to stick out all these candle makers. A genealogist might be interested in that candlestick maker, will you be interested in that one? Or what was the relationship? Did candlestick makers stayed altogether or they were married into each other then there might be different story coming out of it. But the chances of looking for a surname that's what I'm interested in, going there ..well, it's fairly slight. When you write a story about a particular record, how many other people necessarily going to be using it?

362. If not Zooniverse, but similar crowdsourcing project, because wills are full of social, economic and financial personal information, you can build up social hierarchy, social networks as well as the idea of the wealth of

individuals or communities, as well as occupations and everything like that. There's so much more you can tell from the will rather than this is the will of..., who died at... , which is how it's catalogued. To actually here's the names of friends or family relations, books that he bequeathed, property that he owned or whatever happened to be. So wills are with wells of information.

363. The same with archives. If you can get them to be not archivists, but researchers. They would look at the records in completely different way than archivists would look in it. An archivist would look at it in a trust and authentic point of view, but researcher's looking at different point of view. So if archivist would look at it as a researcher - actually I've got to use these records, I've created, can I use them, are they as descriptive as I want? Did anyone really care that there's a full stop in the end? That it's mostly formatted. Do they really care about that? All the energy you put in formatting, that's the thing I remember when we went first to computers. I had colleagues who spent seconds to write it and days to format it, just get the content out. Maybe that's the way I want things to turn, to become a user and then you might revisit some of your practices. OK, we might be expert super-users, so we might not be experiencing in a way that actually novice or a casual researcher might be, but we still have some idea of the ease of navigation, the ease of use, the quality of data. We're never gonna actually, we just can't describe the records to that degree. We can only describe the file, we can't describe the letters, we just cannot. That's just unrealistic to do. But user might want the letters, so is the file descriptive enough to give true indication what's in it?

364. The main challenge we have with going to academics ..ordinary researchers, especially with those, I think it's more probably might be a western thing than might be more of a European thing seem to be to value collaboration and seem to value mark-up tagging, say something about things, volunteering, about having citizen archivists.

365. So a volume, you might describe a few pages, it might be several years before someone else describes few other pages. The amount of time and effort is to manage it. The cost-benefit is quite difficult to justify I suppose.

366. Suppose we don't know who the users are and that's something I want from the management tool, so I can actually start a bit of a dialogue with. With the wiki there was.

367. ..a research team and there's digital research within that team and part of their aim is to go to digital humanities community to understand, what their needs are.

368. We've got to learn how people use current system and actually I know our wrinkles. At the same time we've got to hit the horizon scanning and what's coming along, what people experimenting.

369. But even when you talk to researchers, when you try to find what you want - that's what they want now, how do you find out what they want, you haven't come on stream yet. I don't know how you do that.

370. So we're doing more user research. So we're trying to take a step back from the technology, to accept that as an enabler, and the business again as an enabler to actually, what do the users need, what do they desire. And if they don't need it or they don't want, they don't use it then why we've got it? May be there's a business need for it, well that's fine, if there's a business need, there's certain information to get it really quickly, you can't get it from the back offices. Fine, we can keep that, if they're able to demonstrate that there's a business need for it, then we'll keep that. They might not be the end-user.

371. Then we'd be having a wide discussion across the organisation and possibly with other user or stakeholder groups to get a view, if we get rid of it, so what? If we stop doing that, does it matter? Who would be impacted on it? If someone says, no, this group of people would be impacted, okay, how significantly will they be impacted? Would it stop doing their work, would it stop systems doing the work that can possibly in the future or even in the short term of future stop systems using it properly. So these are sort of things we're gonna think about, but ultimately the end-user is in the focus, because otherwise they are actually using it day and day out. And if they're getting lost, they go somewhere else. Whereas the other ones, the specialist users might be willing to ? to figure out, once they've figured it out, they come back and they are learned.
(Grannum, 2017)

372. Some kind of advanced strategy for how we're going to make sure that the right people to see this content - it simply wasn't necessary in a digital sense. And the hard work, when it came to FCO materials was really done by outreach colleagues, who were reaching out to groups, whose countries of ancestry were represented in those collections. They were deliberately saying, alright what if we've got the car name? and the stuff in a particularly rich... One of the community groups out there that might be interested in that material, they would like to work with us. I was very ? about it, but actually we had a team, who was very focused in kind of good ? way. But actually let's go out and find those people, very proactively... see what their reactions to this stuff is, were... we run workshops. I was purely involved in it by a technical sense, but yeah, they produced physical exhibitions, they did a lot of stuff with the material.

373. That's what we've learned, I think again it seems a long time ago, but we're talking it this way, it's a definite kind of ‘we build it people would come’- a sort of a field of dreams, they might just like well publishing the content is the same as access. They're not. You have to go and digital engagement etc. But you have to go and do it. You can't just go, oh, it's on the internet, now I'm gonna go off and do something else. So were the things that we were trying to do to make sure that people were using the material and engaging the material.

374. The collections are global, so the interest is global.

375. That's a particular kind of digital engagement. I'm engaging with one section of community. It's narrow and deep, it's not broad and shallow, which social media can sometimes be. You make kind of strategic choices there.
376. My average colleagues might think the Flickr materials were running... what they wouldn't have called it digital engagement, because to them the kind of outreach we did all the time. But to me it was digital engagement, because we were engaging with that material online. They were doing it in a face to face in a very proactive way.
(Pugh, 2017)

Community-Tools

377. She was basically saying when Explore was launched in 2008, people just didn't get the whole idea of tagging and no one's really done anything to promote it properly, it's just a feature. There's never been like a focused effort to encourage people to do anything with it.
378. We planned launching crowdsourcing application, it implies to something that we were going to build [blogpost says: "we plan to launch crowdsourcing platform in the beginning of the next year"]. What actually happened. We didn't do that, other people did it for us. People like Mario Klingemann did things like automatic image analysis. We got busy with other things. And there was so much interest.
379. I think there are certain special situations where you can use a combination of computational and human efforts. But I think you need both, you are probably going to need both for a long long time.
380. Typical problems we've found were... the type of research question the people propose were around finding things in messy data. That was a typical problem we have. So we have a lot of data, part of it is messy, how do we find things in it?
381. Our technologies are getting better, so some of them don't always have to be physically on site, they can be remote, so they can be... we have a researcher in Australia right now, Finland, many different places in Europe, America, Canada and they are accessing remotely. But that relationship has to be based on trust, we have to trust the. I'm working with a psychiatrist at the moment, who's looking at our 19 century newspapers. He is particularly interested in, how immigrants... how mental illness was seen amongst immigrants in the UK that came to Victorian Britain. He's actually been doing some text mining, using R and he has now built a corpus on suicide data from our 19 century newspapers. He has done that all through this system of working strictly on site. He sometimes comes in, sometimes works remotely. It's always not straightforward.
382. We had an historian, who was looking for Victorian jokes in our archives. His idea was called the Victorian Meme Machine. First of all, how do we find Victorian jokes in our archives? Difficult question, there's no real catalogue records. We have books and newspapers digitised, so how do we

find them and the data is messy. In Victorian times, people didn't call jokes 'jokes', they called them other things. All sorts of problems. The idea was to take the jokes, build a database of them and then connect those jokes to images like a comic and release them over social media. [...]
We actually failed initially. The biggest problem we had at that time was access. How do we even get access? And then quite a few problems. We did have a tool, but it wasn't really based on computation, it was a transcription tool. So the idea was that you would find a column and a human would actually have to transcribe it, so it was using the Omeka platform. So we built a tool, but that was very human generated, very much humans did... So Bob, who did the research got his PhD students to do the work as a reward for their studies.

383. But the Black Abolitionists, what we did with that one was, we didn't really use the tool, we used machine learning. So we wrote code, not me, Ben. And that is all about trying to say, okay, how can I train a computer to find what I'm looking for in messy data? So it's a bit like looking through a rubbish bin, how can we find the good stuff? [..] Only digitised. So what we had to do there was, OK, she luckily had 1000 speeches transcribed, which she found just throughout years of going through newspapers. So we were able to start with that and train the computer and say, this is the kind of patterns you must look for. Then we looked for keywords, you know, slavery, black, abolitionism. We looked for possible variants of errors that computer might make. We were like digital cowboys, we were throwing lassos over the collections and trying to collect something and then a human has to look in that and say, oh yes, there's some cows there... So it's a kind of metaphor. So what we did with her help, we had 1000 speeches, some clues around... The problem with her work was we didn't know exactly where to look. Jokes usually appear in a column in a newspaper, speeches could be anywhere. So that was part of the biggest problem, to find where they were. We threw the lasso, we got some stuff and the human then has to go through them and say, speech-not speech. What we found was that the most relevant speech was about ants. And we were really puzzled, what?! That doesn't make sense. When you read the talk, it's actually about ant society, but it uses metaphors from slavery. A computer said 97% sure this is about slavery. The other problem was that at that time another movement called Chartists were campaigning for working men to get the vote. This is before the women got the vote. And a lot of their speeches were using words about slavery. So there was a lot of noise and actually trying to find things was difficult. What we had to do was to say, okay, Chartists movement was within this time period, we have to remove this. We had to remove so many variables, I think in the end we had 10 year window to find what we were looking for. But it still involves human efforts.

384. Do you know about Trove in Australia? This was a project on how users could correct OCR on digitised newspapers. So users could actually

change, improve the OCR, line by line. They got like a million entries, it was very big, very successful project in terms of user engagement.

385. In terms of human tagging or interaction some of those tags even have been done algorithmically, e.g. ‘maps’ have been identified algorithmically. We’ve probably about half of million of tags, which is approximately 10% of the images have been tagged by humans. If we ignore the 12 tags that we did, I’d say we’ve probably added half a million tags to about 10% of the images.

386. *As regards to LibCrowds I always wanted to ask you, why don’t you teach computer to understand this very nice librarian cataloguer handwriting?*

So, the problem is that they're in multiple languages, Asian, Africa, so it's not just a Roman script. So it'll be a lot of Indo-European languages, Hindi, ?, Nepali and so forth. So, we don’t believe that in some of those languages… like if it was an OCR, optical character recognition, some of the OCR technologies don't exist for those languages. In terms of handwriting, that’s a different type of problem. Some of the catalogue records, mostly are typed, so, yes, I agree, that a computer approach might be good and then it’s also intersperse with handwriting… So there may be Chinese characters on there. So it’s a difficult problem.

And also I think there is a trade-off, because there is not that much information on the card, OK. If you try to use a computer programme the data would not be 100 % perfect. So, it is like where do you draw the fresh hold. If it is over 80 % correct, that means yes, a human can go and fix those corrections. If it is like 50 %, it almost doesn’t make sense a human try to correct what a computer made. The computer might as well do it from the beginning. This is the kind of a problem. So we did experiment with Jane Austen’s original manuscript when she was a child, 13 year old child, so we took some of her handwriting, we took some of her text, the transcript and then we tried to run a computer programme to kind of see, if it would recognize the rest of it. And we've got to 67 % of success rate.

387. The process is probably more important. The tools are just using Mediawiki is the tool behind Wikipedia. And they just learn that and they create an account.

388. If you look at the agenda [of event for geography teachers], it’s not just about us, it’s about learning about digital mapping and things like that. They learn skills. So they learn about geotagging, but if they want their students to work on a project..

389. Wikimedia's tools are developed by volunteers and therefor they are very difficult to use, especially when you want to upload like hundreds of thousands images, so bulk upload. We found that really challenging, because we didn't know who contact... They're all like in clouds. We were really lucky that we had colleagues from the National Library of Scotland, colleagues from England and we had a colleague here, who is wikimedian.

390. I think out of all stories our maps story is probably the most successful, because we released the million images, we were hoping people would tag them, we didn't know how or why they would tag them and then interest towards to emerger? on our maps. Eventually they found all the maps, which I think that's enormous effort, was probably combination of... It was all volunteers. I think it's a fantastic case-study there about a small group of people, who did a fairly monumental task of actually finding 54 000 images. That was done through a combination of human effort and computational effort. People tried to automatically find maps and humans tried to... But flipped through a book, aah there's a map there, 3 maps, 4 maps there. I think that was the most successful story. That's not finished, that story, because those maps are being georeferenced and that's much harder. We had 54 000 maps, we've probably georeferenced about 22 000. It's hard to georeference a map. Some of the maps you can argue they're not even maps, like a painting, it's like is that even a map. So I would say that has the reason why that's the best story is because that's the story where the data has gone back to our systems. I would say out of all of them, that's the perfect circle. We put it out and then augment to discovery, it's not finished, but that's probably the most successful.

391. Sherlocknet was using computational methods to automatically tag the images.
But when they started we gave them all our tags that were ? from Flickr, so that they already had a set of data to help them with this process.
So they were using sophisticated different things. One of the problems we had, was because the Flickr images are mostly illustrations, we didn't really have any, we call it ground truth data to be able to say, this is a drawing of an elephant, this is a frog. We didn't have that data. What we did have were picture libraries, photographic libraries of... picture of frog, it's a frog. What they did was an initial experiment to use photographic library to see, if it worked with the drawings, not so good.
[...] The British Museum had a catalogue records of prints and drawings from the 19th century about half a million. And we were able to use that data as a training set, because the Flickr images were mainly prints and drawings, because photography wasn't invented. [...] They had made these records available as linked data. So with their permission, that was half a million records, which was catalogue records, which had descriptions. And then they used that data to help them tag the images, and also they did an experiment with captions. Captions is even more difficult then... They used the description from the prints and drawings to try to caption... The Microsoft Coco is a photographic collection and the British Museum's prints and drawings is the metadata for about half a million prints and drawings from the British Museum's catalogue and they use this. E.g. an experiment of a caption, a man standing in a field with a cow. That's pretty good. [...] There are some issues. I would say that was interesting collaboration, there's a lot more work needed.

But what they did do, is they tried to categorise all the images in terms of 12 categories. So that what you find, if you go to Sherlocknet. They also then looked for surrounding OCR text around the image to see, if it helped to give extra information to tag. So they did a few interesting things.

392. I was approached about the machine learning tags that Sherlocknet did on the Flickr images. So they provided about 20 million tags. I was approached by a data scientist, who is going to look at those tags and see, if they can be improved using maybe different machine learning techniques.

393. We preserve the data that they've generated. So we've got datasets.

394. Technically the libraries architecture and infrastructure is not suited for the kinds of things a lot of researchers want to do, especially with large collections. The library is very much based on giving access to one item at a time. The catalogue is designed that way. Whereas someone wants to access thousands or millions of items at the same time to conduct experiments across the corpus that .. we don't have that setup.
(Mahey, 2016, 2017)

395. We've done some usability testing so far. Not with any external people yet. So marking up seems relatively straightforward. Typing a text... If I think generally social media for what that teaches people.. So tagging projects got a lot more easy when the Facebook started using the tagging. So you used to have to explain it and then Facebook got photo tagging, suddenly everyone understood what it was and started doing it, it made it so much easier. So I try to use mental models that fit what people will already know.

396. Really interesting, the inactive handwritten text, that there's an intimacy of handwritten text, there's kind of trace of someone's, like actual presence that carries through the digital objects, that's quite instinct. And handwritten text is a good way into history for a lot of people, so you start transcribing things, get interested, maybe become more sort of more historically curious. So if computers can do all of that, then we've lost an interesting way.. Because that was really interesting crowdsourcing ? engagement where you'd start looking at playbills and then people like ... who own these playrights and what was it like being an actor, why does it always say presented for the pleasure of... There's a lot of points where you can get interested in and start thinking of questions.

397. So if machines get really good reading in handwriting, and really good at reading texts and really good at understanding semantics and then we loose crowdsourcing projects, that would be a shame, because we've lost that gentle way in. But then in the other hand, it's gonna happen at some point, so thinking of ways to either.. like human-computation systems and finding other ways to engage people.

398. And I would much rather see what they say about the collections that is more meaningful to me than what they say about the catalogue. I mean looking at the tags, it is... I'd always suspected that it's just people how they

use other catalogues, like ? 'to read' tags, 'check'... The catalogue isn't naturally a place where you do stuff, it's metadata, it's about discoverability, you don't really have good subject loops, it's a mechanism rather than an experience. And I rather see... there's more value for me in the knowledge that people generate as they use the actual collection items or as the experience whatever then…

399. A lot of the tags are given to modern books, so why would we need to be the ones to ? of those tags. If you look at GoodReads or sites like that, there's much more reader activity and much more discussion and richer intellectual resource of information than here looks in our catalogue is.

400. *Do you plan to gather there data.bl.uk also the experiences what others have been doing with the data so that if someone has the same interest wouldn't invent a wheel?*

Yeah, we... There's a mailing list, I think it's on Mahendra's to do list actually to get that going. We hope that people might... they'd clean up some data, they might ? say like this is how I cleaned it, this is how I tighten it up, main things consistent, but that's still work in progress.

401. That's why I get obsessed with workflow and infrastructure. So that the bits and pieces that you do get back that we kind of make the most of that.

402. We got information for transcription's project for various selected volumes that we need to check the copyright, we're not sure the project contributors, if they will allow us to ingest the materials we're trying to work that out, we get scholars e-mailing us that I've transcribed, I had my students transcribed like 60 volumes of correspondence. And we're like OK, let’s try to work out how we could include that. The library hasn't in the past thought about external people to deliver content. We do a lot of partnership projects, they tend to make little ? systems, so we're trying to be more flexible in how we ingest material. But it's a big challenge.

(Ridge, 2017)

403. Academics, who can use it in a different way than the lay user. They might be looking for the same information, but they might use it more in a macro level. They're be trying to get themes, collection-wide rather than lay user, who might be looking for a specific record and individual record. And they get frustrated because we haven't described it to that level. The academic will hopefully look it more holistically. So again they might want different tools than the public user. They often want contextual information, they want to know, when it was transferred, how it was transferred, who had the custodial information of provenance, trust, ?.

404. So there's a thing in it that.. the filters that the average user does not use, does not need to use, but the specialist user does use. So.. a press release.. there's records about UFOs that comes to National Archives, oh, I just heard BBC got ? of UFOs, you just put UFO and get thousands of results. But you just want to see what they were that came out today. Tick that box there, aaah!

Simple! Most people don't need that, but for a small group you can see it more as a business need.

405. There could be some features or functions that people don't use or don't realise are there, but we might still keep those. Tags can be example, that most people don't use them from a end-user point of view, they just use these for their own user account. Then we can think about actually all the different ways we can manage it. If they're using it for their own user account, so there's no value for the end-user. Why don't we just put it public-private?

406. But yeah, it must be 19 years ago when we set up a wiki, that wasn't a push then, it was still sort of unique, that you got your name on that article, which wasn't necessarily seen.

407. But also wiki is quite niche, there actually isn't a big market, there aren't a large number of users.

408. What I've found is that a lot of people tag at the high level not on the record level. It could be that the record level is not descriptive enough. Or they found that information elsewhere, I think that some of that has been tagged at the high level would be army service records - we haven't catalogued them down to individual soldiers' level, but they have been digitised and they're available through FindMyPast. Now it's possible that people found them there, that they know the record collection that they came from, but not what record they came from, so they've annotated that..

409. We're trying to put up a new one that actually, if it's flagged they can say something about it. Because the other thing is that there is no mechanism for being able to say we're happy with it, except by just e-mailing us and we've found that a few times where people have used this Suggest a correction form. So they use that to let us know. So we've done it occasions to remove tags that are inappropriate.

410. The challenge is actually finding people to use it [API], because it's old stuff, most modern... As far as I can tell, there's very few organisations that harvest historic stuff or has put it away so it allows...People ask us to use the API. We get back to them and they haven't. We've just come out with the new one. The other one will be retired, because we completely redid our schema. There were data elements... The old API was purely for National Archives' data.

411. Unless it'll become live again, then actually people start to work collaboratively. And I know nowadays it tends to be more collaborative than it tends to be more of a push.

412. It also stays when multiple people want to mark or tag the same record. So they keep it for their own accounts in order to go back to it. So you loose that element as well, so you're not getting the idea that lots of people are also interested in that particular record.

413. Because people might be using that means something to them. Because when they look at their own list of tags, it means something, when it doesn't

necessary mean anything to anyone else. So that's the private-public view again.

414. Thing is, if things are all public, you might be more resistant about what you say. You might be more ? about marking up things as say you're doing work collaboratively with other academics, but what you don't want yet is to make it available. So how you do that in a more closed environment? So you might not get descriptions that could be useful, but at the same time there is a lot of noise because we have the only option as public.
(Grannum, 2017)

415. In terms of community engagement, that's not the appropriate place of doing community engagement in there in that way. Actually the outreach team kind of came to the conclusion that Flickr wasn't either, because you needed to sign up for an account to contribute. And when they were doing actual "hello, let's have a cup of tea and meet", running sessions like that, actually that was a significant barrier.

416. At the time there were people who thought, well I can remember going to a conference about web scaling in the title, the suggestion being, that if you provided content in a large scale platform like that it generate automatically lots of traffic and that was not the case. Lots of institutions learned that we proceed to a particular route that got us into the inclusive club that actually for a short amount of time did guarantee lot of content with lovely... We were on a Flickr blog, it was big deal then to have these things. But I think we understood that just because the audience is very very large... it's been just like in 90s we were desperate to sell the new products to China, because there are lots of people in China, it just meant that... just get a box of ? into China, you would make unlimited amounts of money. And the idea seemed to be, just get the content on this popular website and then it will be gravy, and that obviously wasn't the case.

417. Maybe that doesn't matter [where the images are], because now you have social media platforms that you can use separately. The images just have to be somewhere and then you can talk about them somewhere else and you don't have to have the content and the conversation at the same place. As I thought but we didn't do then, because things like Facebook and Twitter were somewhere in their infancy?.

418. The kind of people's bad behaviour online seems to suggest that there is this kind of ... Because they're not present, in a way behind the screen they feel able to behave in a way that wouldn't in real life. And there seems to be some evidence, that if they're anonymous that strengthens their ability to do that. And obviously yes, the Commons or Flickr are account based, they're attributable, you can't see, who's made them. Maybe they've got a fully flashed out profile, maybe it's like a.. the egg as on the wall on Twitter, it's a little bit like a cartoon, a robotic face. I can't remember what the icon for somebody

who wasn't an institution, who was a person with a profile. Obviously you could see the difference between the two accounts.

419. I think it probably intimidated people slightly to be talking the National Archives. Again I'm not sure it would have helped, but might have seemed like a big deal. But it also made me slightly more formal in those interactions; well aware of that big logo is kind of next to whatever.

420. Yes, they're actually doing it here [adding individual, private tags to the catalogue]. I don't recall seeing this very much in the Flickr tags. My perception of the differences between the two sets of tags was that Flickr tags did tend to be more generally descriptive in that sense. They would say what they saw or they would be describing other context around the images, they wouldn't use the tags necessary to express personal things quite the same way. Not that those weren't represented at all, but they weren't represented to the same degree that we've seen them in the TNA tags.

421. We talked then, we talk less now, we should probably talk more about digital divide, about people who would be interested in this stuff and actually wouldn't be able to say whatever happened, because they don't have internet connection. And as I said, I learned that although the barrier to entry as it was for Flick seemed to be low, friendly, well-known system then. Actually there are loads of people who still haven't heard of it. And actually signing up for an account was a big deal for people in some cases.

422. There was a collaboration between the large number of national museums in the UK, Creative Spaces. Creative Spaces was a really embarrassing failure in around about 2010, where most of the national museums in UK had put content and a side lost it in a very very short amount of time, because it was based around the idea of commenting and sharing and so forth, but it was in a closed space. And of course what people wanted to do was to share content with the people that they know. They don't want to only share it with people who were also on that platform. If the platform was massive like Facebook, that doesn't matter, but if it's very small and it only got museum things in it, that's not a recipe for a success. And so they simply didn't get the buying they hoped for and it kind of failed. That can be very difficult. You know, Your Archives is another case where, we were trying to artificially from a scratch produce an online community. And that may not work. It may simply not be enough of a kind of organic list of active users to make it a viable community in that sense.

423. Your Archives was a kind of digital mediator, so that it did replicate much better the role of the archivist than a system that's just a search system. And the same is with any platform, which allows more of a kind of interaction between an owner and a user is obviously gonna be better elicitation and resolving... People didn't generally asked us a research question... People seemed to get on Flickr that it was about the stuff there and talking about that.

424. There have been other experiments... We had a goal run our own online community a few years ago I believe. The member of staff that run that

project isn't here any longer. At the time I personally thought it was a bad idea because for the same reason that I thought that the same kind of viability problems seemed have deviled Your Archives seemed inevitable there. Where the people want to have discussions about the family history, they want to have them in the place they already have the discussions about the family. They don't want to have them on.. there are already very successful family history forums. I thought the logical thing to do was to set up there and talk to people. That's what you wanted to do, why would you go and stand... You know, if you wanna have conversations to people you don't go to a room on your own. You go to a place where there are lots of other people. And that has always seemed to me the best approach to do the digital engagement, to operate in space where people are. But that was something that was tried, and I think they did manage to get some people there, but it did ? up, so it can't possibly have been that successful.

425. But I think what I learned more specifically and talking to colleagues who were doing the outreach on the wall stats, you wouldn't necessarily... I don't know, there are different kinds of engagements basically. Although we successfully engaged with one kind of community online by attributing the content in this way, we actually not necessarily engaged other kinds of communities.

426. If you use a global platform, the Flickr was a community of people who were interested in looking.. they were just interested in show me something visually appealing online. In that kind of broader sense. And museums and libraries and archives can do that. That was really all that audience had in common. It was pretty diverse. We have a wide range of stuff here and we could find something that was you know, interesting to them. And I'm proud that we did kind of experiment with that.
(Pugh, 2017)

Rules

427. Books in physical forms, which they can then use to create entries for Wikipedia. [...] In terms of Wikipedia entries we actually have to be very careful about the collection items, because they can’t leave the building. So we've set up a room where people can request items like books with information. [...] Because that's in the physical building, we can bring them in there, you are not allowed to bring any drinks or any food and then people would use that.

428. What we've tried to do, is to try to focus on the things that are easier. While recognising we still have a big significant issue around... so most of our digital collections, you can only access on site. We've tried to address that problem, so we've found a solution. It's not sustainable, but what we're tried to do, we try to get people, so they work with us, but they have to be security cleared and we effectively make them staff, that's how we do it. So they become security cleared, they become a project worker, they get a staff pass, they get a login and then they work with us and then they can start trying to do

things with our collections on site. [...]
Some of the things we have to go through lots of hurdles to give people access. E.g. that psychiatrist has to be security cleared, has to have a health and safety induction, has to... and go through quite a few things, and importantly we have to trust them. That they won't take all our things and run off. What we're trying to understand is how can we make that better and more sustainable in the future. Because at the moment we simply can't just anybody who wants to use our things on site, make them members of staff, we can't do that and it's not possible to do that for hundreds of…

429. Some of those digital collections... probably because of legislation and legal issues and copyright will always only be able to be used on site. This is a digital collection. So putting everything out there sounds great, but actually the reality is actually a lot more difficult. And some of the collections that we have would take a lifetime to try to make available openly. E.g. we have a famous person's laptop and when they died they gave it to the British Library and how do we give access to that laptop for our users, when that laptop might contain personal information about their relatives, about their e-mails, internet history. So every single one of those collections would come with their own story and it's own reasons why you can and can't use it.

430. But I think that in the future of course as many of those collections 687 that we can make them available openly. We should, but it's unfortunately for every collection there will be a different story as to why you cannot...

431. And we deliberately focused only on the things, which are openly licensed. Almost everything on data.bl is openly licensed.

432. It is always about ensuring the provenance of the images was always clear, always back to the books and the viewer.

433. E.g. with our newspapers, we have to communicate with our publishers, who maybe paid for digitisation, they're not all publishers, but some of them are publishers. So we communicate with them, say, we have a researcher, they're doing research for purely non-commercial reasons, they may publish some data or some results and can they do that. And that's a kind of negotiation on a case-by-case basis. Our long-term vision is that this should be just every day practice.
(Mahey, 2016, 2017)

434. Often they get in contact to ask for like an easy way to access the material. The thing about releasing materials with open license you don't necessarily need a patent? people.

435. Just metadata that we can give to the catalogue people to ingest and then more volume is discoverable and then keep going like that, because otherwise that's so big that if you take.. if you spread out the effort across all the volumes it would take really long time to anyone to...

436. Some of them required you to understand a lot of cognitive overhead that got kind of obsessed with outreach? team. Like how much hard thinking

are you gonna ask people to do and how much they can be nervous about not having enough the specialist domain knowledge to perform the task. If you're making people anxious, that's not fun for anyone. Which trace back to kind of ethics to care for the participants.

437. When a different institution made their co-created exhibition and there's so much kind of knowledge that people have in the exhibitions, feasibility standards, conservation standards. People are like how are we gonna do this, you're like oh yeah, no we can't do that. And that they allow enough time to commit, transmit their knowledge that these are the basics of having objects and people in a close proximity and allow interview groups and all these things. Because people didn't have the right framework to be creative in. But suggesting a bit of ? I think that wasn't a great dynamic... To me to do it well it takes a lot of time, a lot of commitment, a lot of planning.

438. I get annoyed something about Wikipedians, because they don't always make it easy to find that institution that holds the image. Because I think if you find an image online and you're intrigued by it, the collection that holds that probably has other things that you like, so you should be able to go back and find that and other images. So all images around the internet is fantastic. But in general it's everyone's heritage, we should make it available... as long as you maintain the provenance and the information about this is an image from a particular historical context, it was created by someone somewhere for some reason. And you should always be able to get back to that. It's actually history really in terms of which institution holds it. That's just my view.
(Ridge, 2017)

439. Maybe we should do it as a project rather than waiting people to .. for serendipity, where a lot of people sort of drop things in. [...] So a volume, you might describe a few pages, it might be several years before someone else describes few other pages.

440. Where we took Your Archives, we actually did try to agenda good practice and go back to the users and say, look can you rephrase it this way, or we could standardise our categorisation for example, so it was going down particular route. After a couple of years we saw we have to start standardising the categories. [...] I think from a professional point of view, what you want to do, you want to standardize. So Wikipedia start standardising, look here is a page, the headings on a page, if you're right about a person, this is the sort of information.. So they quite quickly realised that they need to help people to write in a particular way. There are people, the editors that actually go standardising the text. So the fact that this is crowdsourced doesn't stop it's being managed. So I think crowd-sourcing does need, if I use the example with the tagging, cause it's unmanaged, it's a bit too free-form. Whereas if there's a way managing it and folding people in "great that's your grandfather, can you give me his name, can you say when he was born and died" as it starts to get a bit more value for other users than just yourself. Use 'grandfather',

don't use 'grandad' or 'gandpa', so we need to standardise things that way. But the fact that we haven't managed it, is a bit…

441. By default, if you sign in, you can automatically tag. We haven't got to the point to block people. So we haven't different levels of authority. You've got an account - you can tag. We haven't got to say, when people might be using it or using it badly or completely inappropriate. We don't have: have an account but not able to tag. We have not come to that level yet. It hasn't really been that much of an issue.
(Grannum, 2017)

442. There's a lot of risk conversion, and somebody has to be the one going: Oh no, it'll be fine. It's important to have someone going what if X happened. But they do need to be counter-balanced by going: OK but the risk of that is very very low. So I would usually be that person to say: Well, look, if there is a problem we'll simply remove the content. And it's fine. There have never been significant problems as far as I'm aware. At the time I was managing the account I probably deleted, I've probably deleted, moderated one comment I think, which over that number of images is ? impressive. These concerns are out there and we see all sorts of problems on social media now. At the time those risks were largely theoretical I think. We didn't see those problems, even with a relatively sensitive, in a sense that the whole collection is a colonial collection - overwhelming is what it is. But I think people understood that really. They understood where it had come from, they understood what is represented and a sort of what it didn't represent. And we made a little bit of an effort to try to contextualise it a bit, but obviously we couldn't... the descriptions were individual items were simply taken from the catalogue entry and it might be very very vague. Or people brought their own context and that was OK.

443. The usual concern that you don't hear so much now, people saying Oh, we don't know what content people will put next to our content, we don't know what purposes they would make of that content. If you say, yes, that's right, that's the point. There were all sets of concerns about ownership and copyright. [...]
But we were in a way tipping a toe in the water. It required a lot of talking to different teams and internal lobbing and so on so forth to get there in the first place. But as I said, once it was set up, almost immediately it became relatively uncontroversial.

444. So we thought a little bit about content and we obviously had... I mean the main thing would have been the take-down policy. I mean I was very keen on the idea that you have to respond to people, but you don't need to be too panicking about these things.

445. We looked quite carefully the images and I'm not sure were they all, they probably are all posted up there?, my suspicion is that some of them are not. We did decide that some of them weren't directly appropriate. At the time

I thought that is was a bad idea. I think I thought that we were better off putting them, it was better to put them all up. I thought the risks of us being selective out waved the risks of content being seen that subject by other people that shouldn't have been seen. But I was wrong about that. There was no recursion at all. They were both theoretical risks, they both remained theoretical risks. I don't know no one ever said what happened to X, all of the other stuff has been online, but this subset is missing. Maybe they're not missing, but my suspicion is that they probably are still not up.

446. And they tend to share experiences rather than saying I'm researching my family history, can you tell me about that, which we obviously do on social media now, but I have to say, well we have that lot somewhere else. But I don't remember too much of that on Flickr. Again it was a kind of... I think there were quite strong community norms around the kind of interactions that were going on and I don't know kind of accepted ways how to talk about the content. That's my memory, is it still the case, probably not.
(Pugh, 2017)

Division of Labor

447. But the point is that actually it's our users, who tell us what they want to do. And then they're the ones to experiment actually and when they come with their ideas, they're the ones to set the boundaries of what can and can't be done. So they come to us with a question and we say, oh that's too difficult, we don't have the systems we don't have the processes, we don't have procedures to do this. And then other person comes along.

448. I have wanted to promote the people, who have been doing an excellent work, volunteers. Cause I know there are not many. But if you promote them, they will may be a seed for others. I may have told you about georeferencing. A guy called Maurice Nicholson, he was our top georeferencer, retired last fall, obviously had time maybe. I invited him in, I asked him to do a blog post.

449. We noticed e.g. Mario was tagging a lot of images and we were convinced he was using automated methods to do it. And I just e-mailed him and found out who he was and I just said hi, hello and we started conversation. And we ended up doing a lot of work together. The same with James, Maurice. This is kind of lovely way to show our work keeps continuing.

450. What we develop is determined by our users. I think other organisations sometimes are the other way around. It's like maybe some archives are more about: No, we start with us first and look what our skills are and go out, so we make sure that we have the skills first. I think there are some differences, but I can't say for sure. I've never worked in an archive. But I know that there are similarities and some differences as well.

451. Through that you learn a lot about your systems, you learn about what demand there is, and it gives you the possibility to then think, OK, these are all

the things that we've learned, what can we prioritise, because we can't do everything.

452. [in LibCrowds] the idea was that people would check the records against OCLC reference and if it didn't exist on the OCLC then people could either recreate the record from scratch or flag the item that needs to be done. So the part of it was like to look up to see if the data was existent already somewhere else or to recreate it. [...] LibCrowds initial focus was to see if we could crowd source the card-catalogues. And the idea was to scan them all in, they were all microfilmed.

453. Every year they [participants of Wiki Food Editahon] create more titles, more entries, more records, it's based in the British Library, because we always use our collection items. Actually it's quite traditional in some ways, because we're using... A lot of information is taken from physical books and then recreated, that's kind of how it works. [...] Mostly physical [books]. What happens typically, before the event starts, people suggest, I'm going to create records about Estonian cakes, that's I would like to get this book ready, so that when I come in, I'll make some records. That's how it works.

454. [The first two contributors next to BL in 2014 database] were absolutely related to BL. The first (Mario Klingemann) - we've been working with him a lot. He has been using computational techniques to tag the images. He has been working with the Stanford guys, he has been kind of mentor to them to help them.

455. Maurice is going to run a tutorial, showing UK teachers how to geotag the maps. [...] And they are going to teach their students how to geotag the maps. So hopefully we'll have UK school children finishing off the job for us.

456. So we want to work with researchers, who want to do that kinds of things, and see where the challenges are, what systems and processes we have already, what we don't have, what we need to develop, where the gaps are, have we build those bridges. It's all about learning how the library supports the new reader-researchers, especially those we call digital scholars. (Mahey, 2016, 2017)

457. I look at things like the activities in the Labs has been a way in getting some input from the public, scholarly communities or creative maker communities.

458. To get information we can put back into our catalogue, so that people could find actual individual playbills. [...] Because they are bounded into volumes and the volumes are catalogued instead? actual items being catalogued.

459. And then assuming that we get in questions and clues from participants, and we'll add those to the list of things we wanna do.

460. I'd love to see as do the same persons again with the books, I don't know, if that's legally possible because of some digitisation things.

461. We've spend a lot of time exploring the different models of the tasks. And we've got some real micro tasks, like mark the title, mark the date, few others. And then this different variations of that. We don't know, if people want a lot of micro tasks to ? get it very quickly, or they want to do the whole thing. Some of them are quite dense in terms of amount of information on them. It would take a lot of time to transcribe everything on the page. So we have a kind of kitchen sink task, like everything is available to you and that's people who like the sense that they've done everything on the page and it's all complete, correct. [...] We're trying to find a way to put together different kinds of tasks so we can say, if you've got two minutes to go and do this thing, it's really easy, if you want to feel like you've completed everything on this one, but it might take you half an hour to do a page... Some of the typographies are quite dense.

462. I think there's a big difference between information that you gather from people that is either the type of what you see, the thing is in front of you and you type directly copy the text there, getting into describe what you see, which is.. can happen in many levels, like image, tags. Then with archives it tends to be more experimental knowledge or subject descriptions that might be like place names and then getting back to more type of what you see. You don't have to make a judgment about... if there's you new castle, you just type a new castle, you don't have to necessarily trying to work out which of the many new castles it might be. And this is in the context. Images are a lot more easy, to say a lot of different things about them, unless they're complex or technical images. Obviously transcription is pretty easy and transcription is a nice way to get people into this…
(Ridge, 2017)

463. But anything that gets people a clue that there might be something and that's worth spending a bit time and effort to explore, to see what else there might be, that to give people a chance to say something about it would be great.

464. So when we went to Discovery, we found the categorisation of where the pages were categorised was useful, where people could select a particular topic, had listed a lot of records related to that topic, that was useful.

465. Or you put in energy as they do it with [Nara?] and a few others is where we are going to concentrate, we are going to focus on our resources to concentrate on a particular area, to get all our volunteers to get to work on that area for a common goal.

466. Most Zooniverse type of projects where you are focusing on very narrow set of work, but anyone knows what they're working to and in the end you get something in the end of it, which is tangible, manageable, you can actually show evidence of it. Trying to describe a thousands years of history, set of miles of records, it's quite.. not narcy, but it's quite unmanageable.

467. Annotations, but also the opportunity to say something more. No just sort of highlighting things, not to say a few keywords, but to be more descriptive, transcribe. Ideally I think transcriptions is everything. Brilliant. And marked up appropriately to allow people to ? to use it. It's my future. But I don't know…
(Grannum, 2017)

468. Because of some local reporting, that went on and the content was picked up in some regional newspapers on the continent.
(Pugh, 2017)

Outcomes

469. Our maps story is probably the nicest story in terms of crowdsourcing. It's the most complete. We found all the maps and all the maps were georeferenced.

470. There's lot of tagging going on from different angles.

471. [The Sherlocknet team members] are just about to add 20 tags to every single image. They are going to present their results to us on 7 November. One of the thing they were saying they do for the project was that they would tag every single image. And they've identified 20 tags and they're going to tag an image with one of 20 tags basically. That would make a big impact to the stats. [...] So e.g. the Sherlocknet people, what we haven't done, we don't know how accurate their tags are, we know just by looking at some examples, a lot of them are not very good, some of them are quite good, good with animals. They are going to build a web interface where you can actually edit the tags, if they are wrong to improve them, you can edit the captions, if they are wrong to improve them. So I think it's a really important project.

472. What we know anecdotally or by accident, what some people have done with our collections. So artists, people have made skateboards, they've made board games, pop signs, T-shirts. You know what the most use of our images was? They were resold on commercial sites. The argument is, you can say that's terrible, because they're public domain images and people are reselling them. But you can also argue that we are helping the economy. /*laughter*/ The problem with that is how do you then assess the value.

473. I think for example the thing we wish we had done with the Flickr images was to put some kind of donate button, and I might have mentioned this to you. We didn't do that. We should have said, if you like these images, please donate one dollar to the library. We've had probably... Our stats are broken on Flickr in terms of use. We think it's probably about 700 million views now. But if you imagine half a percent of 700 million, that would be financial value to the library. It's much more difficult to articulate value, when it doesn't equate to money, to say, ok this has helped to improve my life or made me happy or I was able to create an art show or I made a T-shirt for my mom, who was really happy. All these... Those are much more difficult to

articulate, but I think this is the problem, the dilemma that a lot of memory organisations have. If they make things available in the public domain, how can they demonstrate the value of this. I think lots of people have tried to solve this problem.

474. Basically every year they [participants of Wiki Food Editahon] have contributed great things on Wikimedia Commons, connected to our collections and Wikipedia.

475. What we do is, we probably have 60 of those projects and some of them are not all successes, some of them are failures for lots of reasons. But we try, where we can legally do it we then give them the option to publish.

476. What we haven't done, we should be doing better, we haven't documented properly all the examples. We need to do some more work. So the idea is that when you got to a page for a dataset, it should also say a link to how this data was used to give people ideas. We haven't done that yet.

477. Then people will see the jokes and think they are funny or they are not funny and things like this. So that particular one we were looking at how can we make engagements with community, local community and wider community. So for that one, we did a comedy night. So we got comedians to come in, quite famous comedians. And we gave them Victorian jokes that they had to make funny again, because a lot of them were created at that historical moment in time. If you look at them now, they don't make sense. So how could they be... In that particular event I worked with Emma and we made sure for example that local community got free tickets to come to us.

478. I saw a really nice example actually amongst our illuminated manuscripts. That collection, the digitised collection is the most popular collection that we have of all our collections. Somebody told me about selfies with illuminated manuscripts, especially figures. So what people would do, they would take a picture of themselves in the same pose and sometimes wearing the same cloths as the original image. I just thought it was a really lovely way to sort of engage with our collections, if you really imagine to in a different way.

479. The Flickr collection inspired artworks for the Burning man festival in Nevada desert[109], which were later on display at the British Library[110]. [*rephrased due to missing recording*]

480. But one of the things we thought was actually along with the Flickr images, people are using them to make products. You know, plates, mugs, T-shirts, skateboards, card games, all sorts of things.

481. One of the ideas for our Flickr collection was an idea to take the images, cut them out, put them on transparent backgrounds and sell them. Because people in graphic design might want to use them. We thought it was a

[109] http://blogs.bl.uk/digital-scholarship/2014/08/the-british-library-meets-burning-man.html

[110] https://www.bl.uk/press-releases/2015/june/burning-man-installation-is-unveiled-at-the-british-library

great idea, but the idea would be, you have to pay someone to do the cutting out by putting it to Photoshop. But once they've done that, the image is made available, for free or for small subscription. We didn't accept that idea. But that idea got funding through Europeana. Europeana design challenge, I think it was. The project was called Public Domain City. They would fund it, and she's from Latvia. Unfortunately the project, they experimented with it, but it didn't finish successfully. But the idea was simple to get public domain images, to snip them out and then make them available. But the business model, I think they didn't understand completely, how they could sustain it, it think that was what they wanted to do. I think she was one of the winners, she had a project, she was an entrepreneur, start-up, they kind of helped her. But she's not doing it now.

482. And what happened was that it became clear that it was receiving a lot of attention. Lots of people were looking at them. We were getting a lot of internet traffic.

483. That was the only announcement we made to the world. That blog post is the most read blog post of the entire BL blogs. Every month we have a marketing meeting where we look at statistics and that blog post is always in top of everything. It's fifth now. [...] It was to go a little bit viral across the internet. We got a lot of press and publicity, national newspapers.

484. Up until now we've received over 500 million views. So it's much more than probably 2014. So nearly half a billion. Every image has been seen at least 30 times on average.

485. We gave Europeana a lot of our images that were tagged, I think we gave them 60000 or something.

486. We also have a set from images online, which is like a commercial arm of the library. A lot of them are from the books, but we also have other collections too in amongst this. So there's actually a little bit of a ecosystem opening in this collection.

487. Internet Archive did something similar by cutting out images from books and putting them on Flickr[111]. They copied us. - *Did they refer to you?* No, we weren't very happy about that.
(Mahey, 2016, 2017)

488. We get projects that generate information in a lot of different ways and we try to cooperate into back to our systems.

489. So we are going to release the datasets. So that you can use whatever uses they might like, whether it is well grounded research or data visualisation or analysis of text or linking ? materials.

490. LibCrowds project before, which is retroconversion of card catalogues, basically helping the material that isn't in an electronic catalogue to catalogue it.

[111] http://www.bbc.com/news/technology-28976849

491. I think probably the pop band that used the Flickr images. Cause it's just they sound so modern and they were using these images so not modern. So they made a video-clip and they ... It was all like under sea, it was really cute. They sort of integrated images from the Flickr from the collection of 19th century book images into their video over there. So it's just a really lovely use of historical imagery. And they didn't? care that we're the British Library, but they did write to us to let us know.
492. I love the British Library Labs competition, the Awards, the one they ask people to tell them have you done an interesting commercial use, creative use, whatever. Cause that meant that people were actually motivated to send in information about how they used the collections. That's really good ? to use and to say, this is how a scholar in Birmingham has used, this is how a ? pop-band is used that.
(Ridge, 2017)

493. So when we went to Discovery, we found [from wiki] the categorisation of where the pages were categorised was useful, where people could select a particular topic, had listed a lot of records related to that topic, that was useful. We moved those across as tags.
494. They are unique tags. There's no structure, there's no hierarchy and there's no importance to associate it with them. They are just quite flapped. And I don't know if that's valuable.
(Grannum, 2017)

495. I'm trying to remember, we definitely made cataloguing changes based on working with Wikimedia. We found out significantly more than we knew about some of the objects. I can't remember, if we ever made corrections to the catalogue based on Flickr.
496. And that [Flickr Commons mailinglist] must have been the way we coordinated things like, oh we're gonna have football theme for the World Cup or whatever.
497. We [...] have started talking again about subject tagging and other kinds of tagging in Discovery. And trying to understand why, because it's kind of academic articles written about Flickr tagging and kind of photograph tagging. And why the tagging that we've got in our catalogue is so different from that kind of tagging. [...] and it's probably also about visibility and a few different things there. But what the literature says about tagging doesn't seem to apply terribly well to the tags that we generate on our own content.
498. Some of the catalogue corrections was very clear as we starting putting stuff up. But lot of the works were by artists we didn't know who they were and we were able to identify new artists based on the suggestions of the members of the public.
(Pugh, 2017)

Annex 2. Themes for Semi-Structured Interviews with Visitors

General interaction

- Are you a user of the Library/Archives?
- Do you use any of its catalogues or social media sites to engage with the physical or digital collections?
- How? What do you do?
- Have you taken part in the events in the Library/Archives?
- Have you taken part in voluntary activities elsewhere?

Interaction related to discoverability

- Have you ever added a tag, comment, suggestion for correction of data on any of those platforms?

– Yes:

- Goal?
- Tools? Any missing features?
- Community? Who affects your motives or actions when using the catalogue? Who benefits from your actions?
- Rules? What are the main principles, guidelines or procedural rules that you keep in mind while interacting with the catalogue?
- Division of labour? Role of the organisation? How important is acknowledgement of volunteers to you? Private or public, material or not, by the institution or other users?
- Outcomes? How frequent activity is it? Compared to other businesses/platforms?

– No:

- Why? For which purpose do you see that you would do that?
- Tools you would use? Features you would need? Public-private tagging? Anonymous or named tags?
- Community that could help you or benefits or plays other role?
- Rules perceived important?
- Division of labor? Role of the organisation? How important is acknowledgement of volunteers to you? Private or public, material or not, by the institution or other users etc.?
- What would be the outcomes, if you were involved in tagging or benefit from the activity of others?

Demographic

- What is your occupation?
- Level of computational skills?
- Gender, age – *observed.*

Annex 3. The Ecosystem of the British Library Flickr Collection

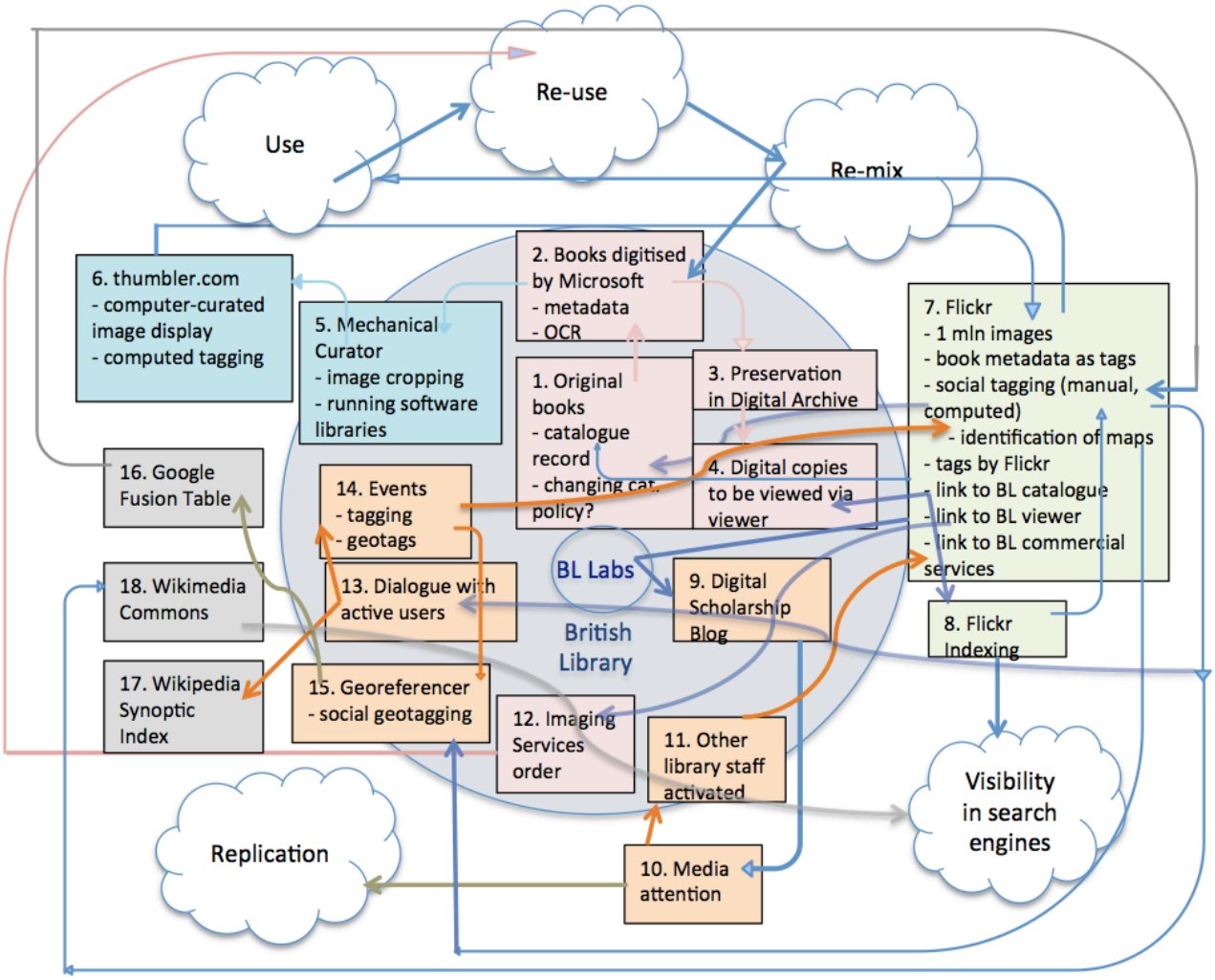

Figure A.3.1. The Ecosystem of the British Library Flickr Collection.

Acknowledgements

This work was made possible through interdisciplinary and international efforts. The author is grateful to Dr. Marco Gui (University of Milano-Bicocca) and Prof. David Lamas (Tallinn University) for supervising the thesis, and for always giving the author much deeper knowledge than her questions involved.

The project was kicked off through consultations with Prof. Serena Vicari and Prof. Alberta Andreotti, supported by Marianna D'Ovidio, Guido Anselmi, many doctoral students in the Department of Sociology and Social Research, and Maurizio Di Girolamo and Ilaria Moroni from the Library of the University of Milano-Bicocca. Many scholars in Tallinn University also shared their expertise: Jaagup Kippar and Mati Mõttus in R, Ilja Šmorgun in human-artefact model and technical support in the lab, and Baseer Baheer in extracting the data of the National Archives via Flickr API. The core part of the thesis was kindly edited by Pippa Brush Chappell.

This project would have been impossible without collaboration with the British Library and the National Archives of the UK, especially Mahendra Mahey's devotion and the input and trust by all the contacts there, including Mia Ridge, Guy Grannum and Jo Pugh – who have turned this individual research project into a co-creative experience of its own.

The author would like to express her words of gratitude also to her family and friends for their support and reflections throughout these three years of the doctoral programme The City and the Society of Information (URBEUR_QUASI).