6. Crowd Mining Applied to Preservation of Digital Cultural Heritage

Emma L. Tonkin*, Gregory J. L. Tourte, and Alastair Gill

Abstract

Accessible systems, in digital heritage as elsewhere, should 'speak the user's language'. However, over long time periods, this may change significantly, and the system must still keep track of it. Conceptualising and tracking change in a population may be achieved using a functional and computable model based on representative datasets. Such a model must encompass relevant characteristics in that population and support pre-defined functionality, such as the ability to track current trends in language use.

Individual published viewpoints on any given platform may be observed in aggregate by means of a large-scale text mining approach. We have made use of social media platforms such as Twitter and Tumblr to collect statistical information about anonymous users' perspectives on cultural heritage items and institutions. Through longitudinal studies, it is possible to identify indicators pointing to an evolution of discourse surrounding cultural heritage items, and provide an estimate of trends relating to represented items and creators.

We describe a functional approach to building useful models of shift in contemporary language use, using data collection across social networks. This approach is informed by existing theoretical approaches to modelling of semantic change.

As a case study, we present a means by which such ongoing user modelling processes drawing on contemporary resources can support 'just-in-time' pre-emptive review of material to be presented to the public. We also show that this approach can feed into enhancement of the data retrieval processes.

6.1 Introduction

It is a familiar observation that digital cultural heritage brings with it new challenges. One such challenge is the effect of long-term technical and societal change on our ability to access and make use of digital objects held within heritage databases. The array of materials that surround and support access to these resources, such as indexing metadata, also suffer from various forms of degradation.

Just as such materials, especially those intended for machine to machine (m2m) usage, may refer to or depend on unavailable or obsolete technologies, information intended for

*Contact email: e.l.tonkin@bristol.ac.uk

human consumption may suffer from processes of obsolescence. For example, resources or accompanying material may use terminology that is no longer current, or which has acquired unintended or problematic connotations, or which is in any other sense inaccessible to the reader. Hence, these materials are also a subject for digital preservation (Brunsmann, 2011). Similar issues and principles apply in contemporary information access contexts. The processes of information sharing between expert practitioners and non-expert members of the public are similarly dependent on the ability to bridge between communities. Consider for example the discussion in the work of Abu-Shumays and Leinhardt (2002) of the role of the *docent*, or volunteer guide, as intermediary between public and professional.

A great deal of existing work in knowledge and information management has focused on related problems, such as the development of knowledge structures designed to support user-focused changes made to information systems such as collection catalogues and search engines. Consider the use of adaptive indexing to allow indexing terms to converge towards user-supplied vocabulary (Furnas, 1985), the provision of facilities enabling user-supplied terms to be used, reused and combined flexibly with formal knowledge structures such as ontologies (Weller, 2007) user-centred evaluation of library indexes (Carlo Bertot, Snead, Jaeger, & McClure, 2006), and formal representations of user journeys through information spaces, enabling past user experience to be called on in future systems engineering (Canter, Rivers, & Storrs, 1985). Many institutions make use of ontologies or taxonomies designed to allow for the use of variant terminologies in describing technical aspects of cultural heritage objects; an example is the Getty Vocabularies published by the Getty Research Institute. Such structures are available for reuse by other institutions, leaving open the key problem of how these should evolve with different times and concepts. Several further examples of such works are concisely discussed in this chapter.

Digital preservation is a significant focus for cultural heritage organisations. In particular, museums, libraries and archives increasingly work with digital objects, as documented by the significant and growing literature on the challenges of digital preservation (Hedstrom, 1997). Long-term digital preservation refers to 'processes, strategies and tools used to store and access digital data for long periods of time', according to Factor et al. (2009, p. 6:1). The time period in question is sufficiently long that technologies, formats, hardware, software and technical communities are likely to change.

There is, we are told (Kuny, 1998), a risk of a 'digital dark age', in which data from the digital age is lost irrevocably. Digital preservation attempts to mitigate such risks, a process which often involves some sort of maintenance: for example, in the case of digital objects, format shifting can resolve problems of format, hardware and software obsolescence.

In this chapter, we explore the issues which can and do contribute to a 'digital dark age', and propose an approach—identifying promising related work in human-computer interaction and information retrieval—to enhance usability of digital data in the future. Whilst we acknowledge that this alone cannot mean that a digital dark age—or similar lack of usability or access—is averted, it can begin to help researchers consider some of the issues in this area.

Our research interest is not solely in preservation of access to data across a large temporal gulf, since similar issues of accessibility and usability arise in contemporary contexts; for example, Burrows, Gooberman-Hill, and Coyle (2015) describe the benefits of 'actively develop[ing] a shared language' between specialist system designers/implementers and end-users. We have a

keen interest in this area and hope to explore the adaptation of the methods in this chapter to support such developments. Cultural heritage catalogues, as large semi-structured datasets, do indeed offer an opportunity to evaluate the effects of cultural and contextual change over time.

In the following, we review some of the long-standing research which has contributed to our understanding of the usability and accessibility concerns presented by online/digital catalogue information. Given these problems, we then discuss in more detail how these and other related issues of digital preservation have an impact upon the cultural heritage sector. In particular, we focus upon the search and retrieval process relating to catalogues, as this is a core function of first- and second- generation cultural heritage organisation informatics as well as a key architectural component underlying implementation of additional functionality: education, study and enjoyment.

This leads us to our proposal, which we believe will provide a new insight in detecting the early signs of possible digital dark ages in this sector. Specifically, we believe that since language is the 'currency' both in indexing catalogues and in search and retrieval behaviour, it makes sense to accommodate potential changes or differences in usage which may provide a barrier to usage for a proportion of those accessing it. By automatically considering linguistic differences, we can identify cases in which performance of mitigating maintenance actions may reduce impact of change, adding further information to support the active curation processes implemented by cultural heritage professionals. In particular, we suggest that data drawn from social sensors and cultural media mining could usefully support such processes of revision.

Whilst the main body of this chapter deals with digital infrastructure, we remark that objects and their surrounding data are accessed in a variety of physical and online contexts. Many of the issues described in this chapter are relevant across online and physical contexts, and we acknowledge this in our literature review and analysis.

6.2 The usable catalogue

In the context of digital cultural heritage, interest in understanding the strengths and weaknesses of the online (electronic) catalogue grew with the evolution of the Web. It is notable that the online library catalogue significantly pre-dates broader digital cultural heritage efforts. Consequentially the following literature review considers evidence drawn from studies of the library catalogue as well as more recent studies focusing directly on cultural heritage. We begin with a discussion of the electronic catalogue, before moving onto physical and hybrid interactions with the system.

Two influential papers by Christine Borgman, published a decade apart, document the development of online catalogue usability between 1986 (Borgman, 1986) and 1996 (Borgman, 1996). Conceptual aspects of system design were a major focus in the earlier paper: misunderstanding of system features, lack of use of advanced search techniques and difficulty in identifying appropriate subject headings ('headline labels' for relevant categories). Two key types of knowledge were identified (Borgman, 1986): knowledge of search syntax, semantics, structure and system, and knowledge of the conceptual aspects of search methodologies. In the later paper (Borgman, 1996), Borgman refines the model, identifying

- conceptual knowledge—in which a person 'seeking knowledge or meaning [...] must

formulate a query in terms of the content of information entities' [or proxies],

- semantic knowledge of query implementation—in this case, the semantics of the catalogue system, and
- technical skills which allow the user to navigate the computer interface and query syntax (syntactic knowledge).

Borgman noted that the capabilities of information retrieval far exceeded those exhibited by catalogue interfaces. Through literature review, Borgman discussed the effects of various factors relating to the catalogue user, such as personality, age and experience, on user search behaviour.

In the 1990s, key questions about cultural heritage were asked about both physical and online visitors: wants, needs and strategies for information access (Cunliffe, Kritou, & Tudhope, 2001). In general, a greater focus was placed on developing well-informed *user models* (see section 6.2.1), seeking to understand patterns of use, visitor aims, information needs and search strategies.

'Next-generation' features (Hildreth, 1987) gained interest and currency, eventually entering the mainstream of catalogue design. Examples include faceted navigation, attempts at correcting user error via methods such as spell-checking and automated format validation, enriched search metadata, ranking of search results and greatly simplified interface design, lessons learned from the simple, sparse search interfaces offered by search engines such as Google (Breeding, 2007). Further innovations (Wilson, 2007) are often focused on active, ongoing user participation. Increased interest in the Semantic Web (Dokoohaki & Matskin, 2008) has facilitated further enrichment of catalogue records by providing technical scaffolds on which to build and datasets against which to work, prompting redesigns by major catalogue holders such as the Library of Congress (Lafrance, 2016).

6.2.1 User modeling

At the core of these developments is the placement of the user at the centre of design work, development and deployment. It is possible to work directly with users to establish user preferences and needs-a process of user centred design. However, it is not always possible at design time to fully anticipate uses that may arise during the operating life of a given software product, and for that reason, systems may be designed to *adapt* to changing circumstances (Van Velsen, Van Der Geest, Klaassen, & Steehouder, 2008). Such systems, to quote Van Velsen et al. (2008, p. 261), 'can alter aspects of their structure, functionality or interface on the basis of a user model generated from implicit and/or explicit user input, in order to accommodate the differing needs of individuals or groups of users and the changing needs of users over time'.

A user model is a structure that describes some aspect(s) of the user and their behaviour. For example, a structure of this kind may capture simple statistics about visitors: age, background, level of education, number of visits. A more complex model might aggregate information about visitor behaviour and context, working from a variety of sources—that is to say that, as well as storing direct evidence such as electronic search histories (explicitly coded interactions with the system), further information may be gleaned from observing a visitor's actions. How does the

visitor approach the collection? How does he/she act? What additional contextual information is available about him/her? (Ruotsalo et al., 2009)

User models typically consist of generalisations built from aggregated data. There are, consequently, ethical issues associated with their use. Trust and privacy are sometimes cited as a concern (Van Velsen et al., 2008). In particular, it is noted that such data is sometimes collected without user awareness of this fact (Kobsa & Schreck, 2003), a practice which would conflict with contemporary data protection legislation such as European data protection regulations. Various strategies have been proposed to mitigate this, such as anonymisation of contributors(Kobsa & Schreck, 2003). However, since subsequent research demonstrates that it is possible to deanonymise ('re-identify') participants in many cases given adequate information (Ohm, 2009), the developer working in this field must build consideration of the ethical and legal implications into each stage of her work.

As we see, the technical evolution of cultural heritage systems is bound closely with the developing story of research themes such as personalisation (Ardissono, Kuflik, & Petrelli, 2012), ubiquitous computing (Kuflik, Kay, & Kummerfeld, 2012) and augmented reality (Wojciechowski, Walczak, White, & Cellary, 2004).

In the following pages, we consider the evolution of metadata creation and use, as it moved from 'one person's view' (a manually created resource) to an automated indexing approach trained (taught) using an aggregate of individual human judgements, and thence towards interface design that explicitly acknowledges variation in user interests, preferences and goals. In section 6.3, we discuss the extension of systems into the physical domain, as experienced by visitors standing within the physical borders of the museum. We consider the role of the museum as expressed by the International Council of Museums, and demonstrate the centrality of the visitor to each goal provided in this definition. In section 6.4, we briefly introduce our research into the use of social sensors to capture contemporary texts relating to a museum and to related artefacts and activities, providing an aggregate model usable to support ongoing maintenance activities on textual material.

6.2.2 Populating a catalogue

6.2.2.1 Manually contributed metadata

Traditionally, catalogue information is manually generated by expert cataloguers chosen from subject experts in the field. It typically contains elements drawn from a controlled vocabulary (a taxonomy or list of subject headings) as well as free-text elements that may contain any choice of string input by the user.

As this process is manual and involves expert input, cataloguing is an expensive process. In some of its more expansive forms, cataloguing work is broadly viewed as prohibitively expensive. From time to time, efforts have been made to reduce cost by involving non-expert contributors: in the Great Depression, for example, non-experts were hired via the Welfare To Work platform to contribute to extensive cataloguing of historical records (Baltimore City Archives, 2014).

In more recent years, the annotation platforms offered on the Social Web provided further support for non-expert annotations (Trant, 2009), although opinions differ on the utility of the

outcome, which are sometimes viewed as privileging serendipity over structured search (Chan, 2007; Van Laere, Bordino, Mejova, & Lalmas, 2014).

6.2.2.2 Automated metadata

A further development is that of automated metadata generation, also known as automated indexing. The field of automated indexing has developed since the 1970s (Stevens, 1970) to encompass areas such as image and multimedia resource indexing. A wide variety of methods are used to extract specific types of information, ranging from extraction of textual features and video captions to extraction of image features. Because the features extracted are seldom directly comparable to the types of metadata created in a traditional catalogue record, there is often a semi-supervised training process necessary to 'translate' findings to compatible catalogue terms.

Although, prima facie, automated metadata extraction systems may be viewed as free from the variable human biases that characterise manually contributed metadata, it is important to recognise that many such systems are trained against exemplars of human performance in a given task, and hence are designed to approximate human performance as closely as possible. Consequentially, such systems replicate the biases common to the training set (Islam, Bryson, & Narayanan, 2016). As the audience changes, and as the material ingested by the museum itself continues to evolve, the assumptions made during system calibration about dataset features and distribution are likely to require revision, a task that is likely to be either expensive or technically challenging (Pan & Yang, 2010).

6.2.3 The search and retrieval process

In the most basic search process, the user simply formulates a search query by providing a number of textual key terms. By interrogating the index of objects, making use of catalogue data, the service is able to identify and return matches. This process draws heavily on the user's ability to generate appropriate search key terms, ordinarily with little prompting from the interface. S/he is likely to be supported by second-generation site features such as search query processing via a thesaurus built into the interface and automated error identification/correction: such thesauri are not automatically updated and require ongoing work to maintain.

6.3 Interactive cultural heritage: Collaborative performance and (re)construction

The increasing focus of the 2000s on supporting user activity beyond catalogue search and discovery heralded further research projects and practical developments focusing on a holistic understanding of the role of a cultural heritage platform. Broader platform functionality is intended to support individual participant needs. Roles that participants may hold include that of student, self-directed learner, an individual planning a physical visit to the museum, subject expert and teacher (Fantoni, 2006). Supporting the data and workflow requirements for each of these roles involves extensive information about the participant, the material held within collections, the physical context and contexts of creation (provenance) and curation of

the object. A broad understanding of these goals implies access to a considerable breadth of extrinsic information, giving, in broad terms, an understanding of the object's place in the world. In pursuit of such goals, data about navigation of a museum, particularly shared navigation, may be collected and used to enrich an existing user model.

6.3.1 Personalisation

Personalised access to museums, libraries and archives was explored during the 2000s by a large number of high-profile cultural heritage institutions (Borgman, 2003). The Rijksmuseum, for example, created a service called Cultural Heritage Information Presentation (CHIP) (Wang, Aroyo, Stash, & Rutledge, 2007). This made use of 'likes' and 'dislikes' expressed by the user; this 'non-obtrusive collection of user data' was designed to support the generation of personalised tours.

Under the name 'personal digital collections systems', functionality allowing users to select items from a museum's catalogue for a personalised collection was implemented by many museums, such as the Museum of Fine Arts in Boston, the National Museum of Australia and Tate Online (Marty, 2011). Use of these systems was found to be popular with a subset of users (Fantoni, 2006), although often collections were abandoned shortly after creation (Fantoni & Bowen, 2007), causing suggestions that the outcome of such systems is 'a landscape of "lost" personal museums' (Marty, 2011).

It is not clear whether the ephemeral nature of an individual's interaction with a cultural heritage system—creating and abandoning, or discarding, a list—should be seen to imply that the interaction is incomplete or superficial, as is often suggested in the literature. It is partially as a consequence of the short-lived nature of many visitor interactions, however, that many institutions have chosen to work with external services, often commercial in nature. These allow the institution to indirectly provide functionality such as social bookmarking or personal digital collection rather than providing the service themselves.

Introduction of a dependency on social websites introduces further risk of a 'digital dark age' (Jeffrey, 2012). Whilst it is not clear that either a self-hosted service model or an external service is sustainable, it is clear that the attrition rate for social websites is high (Tonkin, 2015). Where external services are used, there is a high probability that information (such as course notes, expert or public comment and further annotation) may not be captured by the institution itself. Consequentially, such interactions may be both ephemeral and unobserved by the museum itself, unless action is taken to observe and document patterns of usage. As an aside, even this manner of observation is increasingly challenging, as web technologies continue to develop, impacting on the technical accessibility of web services for purposes such as archiving of information (Kelly, Brunelle, Weigle, & Nelson, 2013).

6.3.2 Co-visiting and shared spaces

In the catalogue-focused examples shown in the previous section the system has been designed with the individual user in mind: one visitor's journey, be it physical, virtual or a combination, is considered as a complete interaction. Others, such as ARCHIE (Van Loon et al., 2006), were designed to support 'co-visiting'—interactive learning within the physical borders of the

museum. ARCHIE followed the contextual model of learning, which points to the contributions and influence of sociocultural, physical and personal context in interactions with objects and subsequent learning (Van Loon et al., 2006; Falk & Dierking, 2000). The Sotto Voce project explored co-visiting and the creation of shared audio media spaces, focusing in part on shared listening to promote interaction (Aoki et al., 2002). Further projects consider location-aware mobile gaming as elements in cooperative learning (Dini, Paternò, & Santoro, 2007).

6.3.3 The mission of the museum

These activities reflect what Lin and Gregor (2006), citing the International Council of Museums (ICOM), describe in a vision statement as the mission of the museum:

> A museum is a non-profit, permanent institution in the service of society and its development, open to the public, which acquires, conserves, researches, communicates and exhibits the tangible and intangible heritage of humanity and its environment for the purposes of education, study and enjoyment.

This definition contains an omission, likely intentional: in stating that a museum is 'open to the public', what would colloquially be described as the target audience is defined as broadly as possible. This statement provides clearer evidence as to the intended purposes of the interaction: communication for education, study and enjoyment. Perhaps this simply reflects the span of ICOM's membership—given over 20,000 museums, it is to be expected that a variety of answers might be given to the question, 'who is this museum primarily intended to serve?'

Yet the importance of specifying the intended audience is made clear by content preservation standards, such as the Open Archival Information System (OAIS) reference model (Lavoie et al., 2002), which identifies what it refers to as a 'designated community'—a construct possibly including several different groups—as a prerequisite for digital preservation activities. This follows when one considers any test of the success or failure of the museum in achieving its stated goals: *communicating* – with whom? Supporting study and education *by which students?* A wholly open mandate is almost impossible to evaluate. The goals must be made concrete and achievable.

In any case, this definition provides us with a series of items that a museum must support: the requirements of education, of study, and of enjoyment. The evidence suggests that provision of a service capable of providing all three of these is an ongoing process that is deeply dependent on a strong and current understanding of the visitor. In particular, we suggest that ongoing service provision in the museum context in particular requires careful monitoring and proactive response to changes in user behaviour, profile and context.

It is not yet clear in what time span issues associated with the types of change process monitored by 'long term' digital preservation become significant to each of these activities. Metrics for establishing the quality of engagement with the visitor, on the other hand, are well-established in general. For the virtual visitor, a body of literature exists on website design and associated metrics for evaluating enjoyment (Lin & Gregor, 2006; Lin, Gregor, & Ewing, 2008). A broader literature on factors associated with enjoyment in human-computer interaction, such as flow (Ghani & Deshpande, 1994), suggests a link between enjoyment and level of

challenge (a factor associated with individual level of certain forms of pre-existing knowledge). Education and study may be viewed as relating to pre-existing formal knowledge, to lived experience, to exposure to contexts, peers and experts with shared areas of interests and to learning opportunities (Ito et al., 2013).

We suggest that the effect of ageing on each key aspect of the museum's mission is a useful area of study, in that the frequency of intervention and hence the sustainability of any institution's services is greatly influenced by this factor. In the remainder of this chapter, however, we will not consider these services further. Rather, we will, by reference to relevant theory, consider the effect of ageing on one relatively straightforward element of the system: the museum catalogue and its accompanying index.

6.4 The Ageing of Cultural Collections

We are familiar with the marks of time. On stone, we expect to see erosion where water once flowed or generations of visitors have walked; on skin, liver-spots and crows' feet. Some of these marks are informationally rich physical clues to the experiences lived by other visitors in past years, such as a well worn passage in a book, or the scuffed floor where furniture once stood. A great deal of recent research on digital preservation has focused on forms of erosion that affect digital objects themselves, such as lack of compatibility with modern software (Factor et al., 2009). The structural conveniences that accompany these objects—the texts and interlinks that provide context and render the material searchable and accessible—are generally engineered for a shorter timescale. What are the effects of age on the tertiary indexes and metadata structures that accompany those objects and make them actionable resources, such as catalogues of digital heritage material?

There are relatively few studies of the effects of age on catalogues over a significant period of time. There are various reasons for this, notably the fact that online catalogues are of themselves a fairly recent phenomenon. The opportunity to observe issues that occur over the long term, as in long-term data preservation scenarios, has therefore been limited. Confounding factors may also be identified. As we see above, catalogue interfaces have typically undergone significant amounts of re-engineering as new technologies and standards are developed and gain in popularity. Where funding exists, cultural heritage organisations will often look to state-of-the-art research, implementation guidelines and even upcoming trends in interface design to overhaul aspects of their online and public presence, meaning that the presentation of information is likely to change rapidly and often, even if the information itself is edited relatively infrequently. This adds a confounding aspect to any longitudinal study. Waterfield's review (Waterfield, 2000) of the development of art cataloguing demonstrates the significance that pragmatic, idealistic and nationalistic concerns took in this nascent field, discussing the complex interplay between the factors mentioned above.

We must therefore look for evidence from indirect sources, such as theoretical and experimental research, which may inform our hypotheses about digital preservation issues on the humble museum catalogue and the information it contains.

6.4.1 The problem of ageing indexes and potential mitigation

6.4.1.1 Studies in recognition and term generation

A contemporary text is written with an audience in mind, according to the author's perception of the strategies that will most effectively transmit his or her point. The same is true of catalogue entries, particularly elements that serve an interpretive or contextual purpose, such as descriptive text.

A subtle effect of passing time is an ongoing change in the style of written and spoken speech. Terminology, in particular, undergoes ongoing processes of change, with terms falling out of favour and being replaced by alternatives. A similar effect occurs between speech communities, which may prefer different terminology to others. Despite this change in preference, low-frequency lexical items are typically still recognised long after falling out of regular use. However, low-frequency words are not well recalled (Lohnas & Kahana, 2013).

Thus, as an unmodified index continues to age, the search process is compromised by increased user difficulty in generating the necessary search terminology—a subtle but measurable process. The search process becomes slower and more error-laden, and the accessibility of the collection consequentially reduces despite the fact that the index itself has not changed at all.

In practice, catalogue and index are not immutable: they are frequently updated for particular purposes, ranging from addition of new items to correction of existing entries for various purposes. A high-profile recent review of Rijksmuseum metadata, for example, saw the alteration of over a hundred items containing racially-charged terminology (Siegal, 2015). User complaints have been identified as a factor in the decision to review the metadata.

6.4.1.2 Mitigation

Taking these things together, we use them as our motivation and basis for addressing digital preservation within cultural heritage institutions.

Words come and go from fashion. In some cases historical or outmoded terms may cause offence. In others, they may cause a breakdown in understanding or ability to retrieve certain information. A strongly data-driven and frequently updated set of user profiles facilitates detection of both conditions. Yet the museum as an operating context introduces particular complexities. The 'user' of a museum is not clearly defined. Consider for example the reuse of museum resources, which may be considered by many individuals or groups for a wide variety of purposes, from the sheer joy of collection[1] to academic research.

6.4.1.3 Scalable, sustainable information sourcing

To inform itself on this and other subjects, an institution may reach out to the public, perhaps through directly working with museum visitors or, as proposed by Wrigglesworth and Watts in Chapter 8, volunteers. Crowdsourcing, as explain van der Lans et al. in Chapter 7, is an alternative approach that, though it may come at a cost, offers the possibility of receiving input from participants across the world. The downside of both approaches is simply the paucity of

[1] Analysis of the social collection website Pinterest (Mull & Lee, 2014) shows that users of the site self-report as finding the process both enjoyable and inspiring

available agents; a large collection cannot hope to regularly revisit every element of a large and growing collection. It is therefore interesting to consider the coexistence of participant-driven approaches with observations generated elsewhere: that is to say, we assert that a machine-driven 'shallow reading' (Etzioni, Banko, & Cafarella, 2006) of a broad variety of sources can usefully support cultural heritage in various tasks related to catalogue maintenance.

As an institution, it is preferable to schedule preventative maintenance in good time than to discover issues that have significantly inconvenienced large numbers of users. Thus, it is proposed that ongoing and non-invasive methods of data collection are used to identify cases in which catalogue (or, more frequently, thesaural) maintenance may become necessary, such that the proposed changes may be reviewed and implemented as part of the routine of data preservation.

6.4.1.4 Recording change, purpose and provenance

In addition, metadata versioning can be put in place to help reflect changes in our understanding of the object (Zavalina, Shakeri, & Kizhakkethil, 2015). The addition of extra information within the metadata, or accompanying the metadata, allows us to record explanations for the changes made. being able to add extra information within the metadata record to provide explanation surrounding the changes made. The metadata surrounding an object, be it digital or physical, is itself a living record, with a chain of provenance that tells its life story.

The provision of service infrastructures to support maintenance tasks in preservation is a fairly common pattern: various digital preservation frameworks have been proposed to this end, including OAIS. These often focus primarily on concrete tasks such as ensuring the ongoing technical accessibility of the materials held and on ensuring that adequate metadata records are stored. For this purpose, necessary technical infrastructure, such as registries of document formats, format validators, etc., are provided. Whilst these standards do not exclude ongoing user modelling, tracking of audience engagement and so forth, they also do not explicitly mandate such activities. Yet it is only through interaction and observation of those engaging with materials that a cultural heritage organisation can establish whether they are achieving their goals.

At times, explanatory information is explicitly provided by archivists; however, the contexts in which such decisions are taken is often not explicitly encoded. We suggest that efforts should be made to capture this information and, perhaps more importantly, to ensure that the information captured is informationally rich. This implies, from the metadata design perspective, providing rich provenance and annotation capabilities.

6.4.1.5 Aggregate user data collection through social sensors

Where created user data is viewed as of interest, an appropriate social sensor may be used to evaluate specific dimensions of social annotations or interactions, such as information shared, emotional responses and geographical localisation (Resch, Summa, Sagl, Zeile, & Exner, 2015). A social sensor is a source of information—which could be anything from a mobile or pervasive computing device to information sequences gleaned from Web services such as Web tools—Flickr, Twitter, search engine queries and activities, and so forth (Rosi et al., 2011)—which tells

us something interesting. Social sensors are used for an increasingly significant set of purposes. For example, using a sensor of this kind it is possible to rapidly detect earthquakes (Sakaki, Okazaki, & Matsuo, 2010) through automated review of Twitter posts, and to provide sufficient information about everyday life in cities to support decision-making for city councils and utilities (Domingo, Bellalta, Palacin, Oliver, & Almirall, 2013) and characterise social interactions within a group (Bell, McDiarmid, & Irvine, 2011) using information from mobile devices.

6.4.1.6 The use of social sensors for the museum observatory

We propose the use of a set of social sensors (Manovich, 2011) and corresponding sensors operating on cultural and political data (e.g., Zeng & Greenfield, 2015) that, in combination, represent a Museum Observatory. Such an observatory allows the museum, in an automated manner, to look out at the constellations of human activity and discourse that surround it and to identify and react to ongoing processes of change. For example, such an observatory might process data relating to news reports referencing the cultural institution, in order to get a sense of the context in which it operates and the discussions surrounding it at a formal level. In addition, social media may provide indications of the concerns of the general public (or sections of the general public). Inclusion of more structured data such as the catalogue search terms (or those in relation to the website), will also allow a more focused perspective on users of the cultural institution's resources.

The incorporation of a wide range of relevant resources which form the immediate ecosystem of the museum, enable a broad and varied view of its context. This observatory could then be integrated into the museum catalogue, for example, by generating relevant terms for automated (or semi-automated) annotation of objects, or for identifying and generating new relationships within the catalogue (e.g., between objects or between different indexing terms, or between objects and indexing terms). Although itself an actor in the ecosystem, the museum is one among many. We argue that to remain accessible to the visitor, the museum must retain an active connection with the general public and along with its wider context. Such an observatory may act as a virtual mirror on an institutional level (Gloor, Oster, Raz, Pentland, & Schoder, 2010). It would, as the poet Robert Burns once put it, 'the giftie gie us/To see oursels as ithers see us!'

Such infrastructure, although cheap by comparison with the human-led processes which it can inform, is far from free: we remark, however, that such observatories do not need to exist in large numbers. In the case of larger institutions, they will most likely have the expertise and resources to create, and curate, such and observatory. However smaller institutions may need to form consortia or create links with the larger institutions; as with significant manual infrastructural efforts such as the development of openly accessible thesauri or taxonomies, the work is better shared than individually replicated.

6.5 Social sensors in corpus review: a preliminary exploration

We developed an initial prototype for the purpose of investigating the feasibility of this approach. A large number of museums, both in the United Kingdom and internationally, have published metadata of varying levels of comprehensiveness describing their collections. In some cases,

providers also offer contextual material such as information about the artist, prior provenance and history of artefacts. In our initial study, we selected a network of UK-based museums as the subject for our work, for several reasons. Firstly, extensive contemporary material about the institution's holdings had been made available online under an appropriate licence[2]. Secondly, we were aware that further relevant material, such as historical guides from the earliest days of this institution, has now been digitised by the Getty Research Institute. Additionally, we had already developed a corpus of statistical information characterising online discussion about several institutions, including Tate (Kontopoulos et al., 2016).

In an initial study, quantitative methods have been explored to support direct comparison with modern corpora drawn from the same topic area and corresponding repair processes. Material drawn from social sensors is considered as a source for data relating to impact and affect, as well as an entry point into relevant online discourse. From this pilot study, we provide a few manually derived examples.

6.5.1 What people say: methods for sensing change through observation of contemporary data sources

Contemporary social media platforms frequently provide an application programming interface (API) that makes material provided by users of these platforms, such as text or multimedia content, available to software. Programmers can straightforwardly work with these interfaces by making use of libraries available in many contemporary programming languages such as R and Python; in our case, we primarily made use of Python in our research.

In our initial work, we began by focusing on two such platforms: the microblogging and social networking websites Tumblr and Twitter. Each of these platforms can be searched for content relating to certain keywords, although Twitter restricts search matches to contemporary material (normally within the last few hours or days), whilst Tumblr allows searching of material from the entire history of the site, with the exception of material posted by since-deleted accounts. We began by exploring a confined space: published material directly relating to a particular museum. This was used as a starting-point from which to identify cliques, in the social network sense (which is to say, clusters of people—or, sometimes, agencies such as news agencies or, indeed, museum social media accounts—between whom there exist many links, such as cross-references, mentions or 'likes').

Through this means, we built up a corpus of material relating to the institution in its many facets—a building; a destination; a sequence of events, of exhibitions, of performances; individuals associated to it directly (for example, the subject of specific exhibitions). Caught, also, in this web, we see people, places and things that are more tangentially connected, such as people producing comparable or relevant work, collaborating institutions, relevant news, and so forth. So, too, do we find references to activities directed by educational institutions—reading material for a course, for example. In itself, such material can be accessed serendipitously—by happy chance, perhaps, something interesting for a reader may be found. It is valuable in part, even through direct inspection, for its exuberant presentation of the multiplicity of faces presented

[2] `https://github.com/tategallery/collection`, available under the Creative Commons Public Domain CC0 licence.

by the institution; indeed, many institutions already seek to characterise electronic participation and interaction.

Less directly, this material also provides us with a reference point: speaking informally, it tells us a little more about the way people write, what they say—and what they do not. There are, of course, limits to the use of this resource, and obvious objections; any corpus is limited in scope, and the contributors to the corpus may or may not have contemporary relevance to a given institution. A shallow reading is only the beginning of the process. To work with a corpus such as the one described above, we need good, relevant questions, complementary datasets and supplementary information.

Whilst we do not propose to go into technical detail in this chapter, we note that methods to achieve various relevant tasks have been reported elsewhere. Consider for example the problem of identifying new uses of existing terminology within a corpus (Dorow & Widdows, 2003). Solutions to this problem allow us to identify existing words that now hold different meanings or connotations, especially problematic connotations. Methods from recent research (Hamilton, Clark, Leskovec, & Jurafsky, 2016) show promise in robustly identifying term sentiment in a new corpus of text. Such methods allow us to identify words that hold positive or negative connotations, allowing systems to track change in the way that terms are perceived and the connotations that they have. Using graph-theoretical methods, it is also possible to identify orthographic changes (that is, changes in the way that words are written) over long periods of time.

By comparing large amounts of text taken from similar subject areas written over very distinct periods of time, we find that using sample corpora we are able to demonstrate differences, the extent of the differences, and, in general, to identify pathways that can be used to bring material up to date.

6.5.2 Examples of use

Consider as a first case the comparison of historical materials with contemporary material: we take several examples of artwork titles from an open catalogue, and evaluate them against our corpus. A proportion of terms simply do not appear in randomly sample modern corpora, or if so, appear very seldom; others have gained or lost connotations over the years.

One example of an uncommon term is 'Newsmongers', the title of an artwork by Sir David Wilkie[3]. The term is defined by the Oxford English Dictionary as referencing a gossip. From the artwork and its accompanying taxonomic information (which includes 'reading, writing, printed matter' and 'newspaper'), it is not clear that the reader would interpret this term accordingly.

An example of the latter—a term used in a manner which is now considered archaic—is in the title of a piece by Gilbert Stuart Newton, 'Yorick and the Grisette'. The Oxford English Dictionary reminds us that a grisette is a common edible woodland mushroom, and, historically, referred to a young working-class Frenchwoman[4]. Neither definition is common on social media; in fact, a grisette is also a type of beer, and this is overwhelmingly the most commonplace use of the term today. There exist two parts to this problem: detection of the fact that the connotations of the token 'grisette' have changed over time, achievable given sufficient information through

[3] `http://www.tate.org.uk/art/artworks/wilkie-newsmongers-n00331`

[4] `https://en.oxforddictionaries.com/definition/grisette`

the use of frequency and co-occurrence information, and selection, by evaluation of the context of use of the term, of appropriate equivalent terms.

Identification of candidate terms is not always straightforward and sometimes, as here, there are no straightforward and commonplace equivalent terms available to us. In other cases, as with the Caravaggio artwork, 'The Decollation of St John', there is a more widely accepted alternative term (i.e., 'beheading') readily usable in its place. Such simple cases are straightforwardly and reliably automated, and can be achieved in the context of 'just-in-time' service provision.

The identification of problematic terminology is useful in that it allows for the helpful provision of (for example) an adaptive thesaurus, accompanying glossary or an explanatory note. The automated nature of this approach enables it to reach into the 'long tail' of a museum's collections, where serendipitous encounters with confused visitors are somewhat less likely to spark spontaneous review of catalogue data.

Beyond description of individual holdings, we have not yet spoken about encounters with a museum—with temporary exhibitions, for example, which attract visitors and comment during their lifetime. Our published findings (Maronidis et al., 2016) show that exhibitions, tours and events are prominent in social media datasets. If we aim to compare perspectives on cultural heritage items, we must also consider the contexts in which they encounter the work, and perhaps also (as a matter of future work), how they react to the work within this context.

6.5.3 Discussion

This study has given us confidence that text analytic methods based on the distributional hypothesis, twinned with topic-comparable corpora from distinct temporal or social contexts, can be used both to describe the gulf that lies between the two corpora, and to identify strategies that enable the construction of knowledge structures that facilitate the bridging of the gap. We expect to report more extensively on this work in an upcoming publication.

Maintaining relevant and accessible descriptors for information objects (i.e., a usable catalogue) is a time-consuming task. That said, data-driven approaches are themselves time-consuming. Data collection is a lengthy task, although it can be automated to a large extent; maintenance of software is time-consuming and potentially expensive. For a single institution, data-driven approaches may not be an economical solution.

It is therefore recommended that efforts continue to be cast as collaborative between multiple cultural heritage institutions and groups likely to make use of the data. Siloised models of operation are commonplace, but, as in this instance, there is much to be gained from collaboration between institutions.

It is also worth noting that the sampling methodology has a significant impact on the relevance and accuracy of the results. An observer receives a partial and biased view of user activities and responses, so there is a risk that an institution may overfit their solution, hence further reducing the accessibility of the result.

6.6 Conclusion

Digital preservation of cultural heritage often focuses on digital objects or digital proxies of physical objects, however museums require significant surrounding infrastructure to fulfil their

mission which includes both the core function of search and discovery of information, as well as supporting visitors in education, learning and enjoyment.

In this chapter, we have begun to explore the impact of language change in accessing catalogue information. Given that catalogue function is a key element of cultural heritage sites, the cumulative result of this process may cause a significant decline in accessibility of information. One proposal to potentially mitigate this might be the use of social sensors and cultural media mining. In this paper, we have proposed that the collection of this kind of information from open data sources could be used to provide a landscape from which to understand and interpret catalogue information. By automatically and periodically collecting this information, in a manner that takes into consideration ethical concerns, it becomes possible for the catalogue/index to model—and therefore take account of—changes in common understanding or usage of language, with a view to cultural context, including specific search terms. This therefore may go some way towards supporting pro-active maintenance of museum infrastructure as it is currently supported for digital object formats. We hope that by doing so, this may begin to address the problems of catalogue accessibility with relation to averting a digital dark age. Such an approach may also contribute to supporting the increasingly extensive objectives of cultural organisations in supporting education, study and enjoyment.

We hope to explore the use of similar linguistically inspired approaches across heterogeneous groups for practical purposes. One such approach is the support of system development by identifying problematic or specialised jargon. A second is to provide support for detailed analysis of specialised cataloguing approaches and practices, which is of use in the normalisation of catalogue metadata, and in the sociological study of practices in a specialised field.

References

Abu-Shumays, M. & Leinhardt, G. (2002). Two docents in three museums: central and peripheral participation. *Learning conversations in museums*, 45–80.

Aoki, P. M., Grinter, R. E., Hurst, A., Szymanski, M. H., Thornton, J. D., & Woodruff, A. (2002). Sotto voce: exploring the interplay of conversation and mobile audio spaces. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 431–438). ACM.

Ardissono, L., Kuflik, T., & Petrelli, D. (2012). Personalization in cultural heritage: the road travelled and the one ahead. *User Modeling and User-Adapted Interaction*, *22*(1), 73–99. doi:10.1007/s11257-011-9104-x

Baltimore City Archives. (2014). Transcribing and inventorying the records of baltimore city, 1905–1940. Retrieved from http://baltimorecityhistory.net/research-at-the-baltimore-city-archives/transcribing-and-inventorying-the-records-of-baltimore-city-1905-1940/

Bell, S., McDiarmid, A., & Irvine, J. (2011, May). Nodobo: mobile phone as a software sensor for social network research. In *2011 ieee 73rd vehicular technology conference (vtc spring)* (pp. 1–5). doi:10.1109/VETECS.2011.5956319

Borgman, C. L. (1986). Why are online catalogs hard to use? lessons learned from information-retrieval studies. *Journal of the American society for information science*, *37*(6), 387–400.

Borgman, C. L. (1996). Why are online catalogs still hard to use? *JASIS*, *47*(7), 493–503.

Borgman, C. L. (2003). Personal digital libraries: creating individual spaces for innovation. In *Nsf post-digital library futures workshop*.

Breeding, M. (2007, July). Next-gen library catalogs. *Library Technology Reports*, 10–13.

Brunsmann, J. (2011). Product lifecycle metadata harmonization with the future in oais archives. In *International conference on dublin core and metadata applications* (pp. 126–136).

Burrows, A., Gooberman-Hill, R., & Coyle, D. (2015, December). Shared language and the design of home healthcare technology. In *Proceedings of the acm conference on human factors in computing systems*.

Canter, D., Rivers, R., & Storrs, G. (1985). Characterizing user navigation through complex data structures. *Behaviour & Information Technology*, *4*(2), 93–102.

Carlo Bertot, J., Snead, J. T., Jaeger, P. T., & McClure, C. R. (2006). Functionality, usability, and accessibility: iterative user-centered evaluation strategies for digital libraries. *Performance Measurement and Metrics*, *7*(1), 17–28.

Chan, S. (2007). Tagging and searching: serendipity and museum collection databases.

Cunliffe, D., Kritou, E., & Tudhope, D. (2001). Usability evaluation for museum web sites. *Museum Management and Curatorship*, *19*(3), 229–252.

Dini, R., Paternò, F., & Santoro, C. (2007). An environment to support multi-user interaction and cooperation for improving museum visits through games. In *Proceedings of the 9th international conference on human computer interaction with mobile devices and services* (pp. 515–521). ACM.

Dokoohaki, N. & Matskin, M. (2008). Personalizing human interaction through hybrid ontological profiling: cultural heritage case study. In M. Ronchetti (Ed.), *1st workshop on semantic web applications and human aspects,(swaha08)* (pp. 133–140). In conjunction with Asian Semantic Web Conference.

Domingo, A., Bellalta, B., Palacin, M., Oliver, M., & Almirall, E. (2013, winter). Public open sensor data: revolutionizing smart cities. *IEEE Technology and Society Magazine*, *32*(4), 50–56. doi:10.1109/MTS.2013.2286421

Dorow, B. & Widdows, D. (2003). Discovering corpus-specific word senses. In *Proceedings of the tenth conference on european chapter of the association for computational linguistics-volume 2* (pp. 79–82). Association for Computational Linguistics.

Etzioni, O., Banko, M., & Cafarella, M. J. (2006). Machine reading. In *Aaai* (Vol. 6, pp. 1517–1519).

Factor, M., Henis, E., Naor, D., Rabinovici-Cohen, S., Reshef, P., Ronen, S., … Guercio, M. (2009). Authenticity and provenance in long term digital preservation: modeling and implementation in preservation aware storage. In *First workshop on on theory and practice of provenance* (6:1–6:10). TAPP'09. San Francisco, CA: USENIX Association.

Falk, J. H. & Dierking, L. D. (2000). *Learning from museums: visitor experiences and the making of meaning*. Altamira Press.

Fantoni, S. F. (2006). Web-based solutions: save it for later. Retrieved from http://www.artsprofessional.co.uk/magazine/article/web-based-solutions-save-it-later

Fantoni, S. F. & Bowen, J. P. (2007). Bookmarking in museums: extending the museum experience beyond the visit. In J. Trant & D. Bearman (Eds.), *Museums and the web 2007*. Archives and Museum Informatics. Toronto.

Furnas, G. W. (1985). *Experience with an adaptive indexing scheme*. ACM.

Ghani, J. A. & Deshpande, S. P. (1994). Task characteristics and the experience of optimal flow in human-computer interaction. *The Journal of psychology*, *128*(4), 381–391.

Gloor, P. A., Oster, D., Raz, O., Pentland, A., & Schoder, D. (2010). The virtual mirror: reflecting on the social and psychological self to increase organizational creativity. *International Studies of Management & Organization*, *40*(2), 74–94.

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. *arXiv preprint arXiv:1606.02820*.

Hedstrom, M. (1997). Digital preservation: a time bomb for digital libraries. *Computers and the Humanities*, *31*(3), 189–202.

Hildreth, C. (1987, Spring). Beyond boolean; designing the next generation of online catalogues. *Library Trends*, 647–67.

Islam, A. C., Bryson, J. J., & Narayanan, A. (2016). Semantics derived automatically from language corpora necessarily contain human biases. *CoRR, abs/1608.07187*. Retrieved from http://arxiv.org/abs/1608.07187

Ito, M., Gutierrez, K., Livingstone, S., Penuel, B., Rhodes, J., Salen, K., … Watkins, S. C. (2013). *Connected learning: an agenda for research and design*. Digital Media and Learning Research Hub.

Jeffrey, S. (2012). A new digital dark age? collaborative web tools, social media and long-term preservation. *World Archaeology*, *44*(4), 553–570.

Kelly, M., Brunelle, J. F., Weigle, M. C., & Nelson, M. L. (2013). On the change in archivability of websites over time. In T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, & C. J. Farrugia (Eds.), *Research and advanced technology for digital libraries: international conference on theory and practice of digital libraries, tpdl 2013, valletta, malta, september 22-26, 2013. proceedings* (pp. 35–47). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-40501-3_5

Kobsa, A. & Schreck, J. (2003). Privacy through pseudonymity in user-adaptive systems. *ACM Transactions on Internet Technology (TOIT)*, *3*(2), 149–183.

Kontopoulos, E., Riga, M., Mitzias, P., Andreadis, S., Stavropoulos, T., Konstantinidis, K., … Tonkin, E. L. (2016). Pericles deliverable 4.4: modelling contextualised semantics. PERICLES project.

Kuflik, T., Kay, J., & Kummerfeld, B. (2012). Challenges and solutions of ubiquitous user modeling. In A. Krüger & T. Kuflik (Eds.), *Ubiquitous display environments* (pp. 7–30). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-27663-7_2

Kuny, T. (1998). The digital dark ages? challenges in the preservation of electronic information. *International preservation news*, (17), 8–13.

Lafrance, A. (2016). Archaeology's information revolution - the atlantic. Retrieved February 3, 2017, from https://www.theatlantic.com/technology/archive/2016/03/digital-material-worlds/471858/

Lavoie, B., Alexander, M., Rieger, O., Bradley, K., Sergeant, D., Day, M., … Woodyard, D. (2002). *Preservation metadata and the oais information model. a metadata framework to support the preservation of digital objects*. OCLC Online Computer Library Center, Inc. Dublin, OH. Retrieved from http://www.oclc.org/content/dam/research/activities/pmwg/pm_framework.pdf

Lin, A. C. H. & Gregor, S. D. (2006). Designing websites for learning and enjoyment: a study of museum experiences. *The International Review of Research in Open and Distributed Learning*, *7*(3). Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/364/735

Lin, A. C. H., Gregor, S. D., & Ewing, M. (2008). Developing a scale to measure the enjoyment of web experiences. *Journal of Interactive Marketing*, *22*(4), 40–57. doi:10.1002/dir.20120

Lohnas, L. J. & Kahana, M. J. (2013). Parametric effects of word frequency in memory for mixed frequency lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1943–1946.

Manovich, L. (2011). Trending: the promises and the challenges of big social data. *Debates in the digital humanities*, 460–475.

Maronidis, A., Chatzilari, E., Kontopoulos, E., Nikopoulos, S., Riga, M., Mitzias, P., … other. (2016). *Pericles deliverable 4.3: content semantics and use context analysis techniques*. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-11750

Marty, P. F. (2011). My lost museum: user expectations and motivations for creating personal digital collections on museum websites. *Library & information science research*, *33*(3), 211–219.

Mull, I. R. & Lee, S.-E. (2014). "pin" pointing the motivational dimensions behind pinterest. *Computers in Human Behavior*, *33*, 192–200. doi:10.1016/j.chb.2014.01.011

Ohm, P. (2009). Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA Law Review*. U of Colorado Law Legal Studies Research Paper No. 9-12. *57*, 1701–1777. Retrieved from https://ssrn.com/abstract=1450006

Pan, S. J. & Yang, Q. (2010, October). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. doi:10.1109/TKDE.2009.191

Resch, B., Summa, A., Sagl, G., Zeile, P., & Exner, J.-P. (2015). Urban emotions — geo-semantic emotion extraction from technical sensors, human sensors and crowdsourced data. In *Progress in location-based services 2014* (pp. 199–212). Springer.

Rosi, A., Mamei, M., Zambonelli, F., Dobson, S., Stevenson, G., & Ye, J. (2011, March). Social sensors and pervasive services: approaches and perspectives. In *2011 ieee international conference on pervasive computing and communications workshops (percom workshops)* (pp. 525–530). doi:10.1109/PERCOMW.2011.5766946

Ruotsalo, T., Mäkelä, E., Kauppinen, T., Hyvönen, E., Haav, K., Rantala, V., … Matskin, M. (2009). Smartmuseum – personalized context-aware access to digital cultural heritage.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web* (pp. 851–860). WWW '10. Raleigh, North Carolina, USA: ACM. doi:10.1145/1772690.1772777

Siegal, N. (2015). Rijksmuseum removing racially charged terms from artworks, titles and descriptions. *New York Times*. Retrieved from http://nyti.ms/1SQeoEX

Stevens, M. E. (1970). Automatic indexing: a state-of-the-art report.

Tonkin, E. L. (2015, January). *Supporting unsupervised context identification using social and physical sensors* (Doctoral dissertation, Department of Computer Science, The University of Bristol). Retrieved from http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.686425

Trant, J. (2009). Studying social tagging and folksonomy: a review and framework. *Journal of Digital Information*, *10*(1).

Van Laere, O., Bordino, I., Mejova, Y., & Lalmas, M. (2014). Deesse: entity-driven exploratory and serendipitous search system. In *Proceedings of the 23rd acm international conference on conference on information and knowledge management* (pp. 2072–2074). CIKM '14. Shanghai, China: ACM. doi:10.1145/2661829.2661853

Van Loon, H., Gabriëls, K., Teunkens, D., Robert, K., Luyten, K., Coninx, K., & Manshoven, E. (2006). Designing for interaction: socially-aware museum handheld guides. *NODEM 06-Digital Interpretation in Cultural Heritage, Art and Science*.

Van Velsen, L., Van Der Geest, T., Klaassen, R., & Steehouder, M. (2008). User-centered evaluation of adaptive and adaptable systems: a literature review. *The Knowledge Engineering Review*, *23*(3), 261–281. doi:10.1017/S0269888908001379

Wang, Y., Aroyo, L. M., Stash, N., & Rutledge, L. (2007). Interactive user modeling for personalized access to museum collections: the rijksmuseum case study. In *User modeling 2007* (pp. 385–389). Springer.

Waterfield, G. (2000). The origins of the early picture gallery catalogue in europe, and its manifestation in victorian britain. *Art in Museums*, 42–73.

Weller, K. (2007). Folksonomies and ontologies: two new players in indexing and knowledge representation. *Applying web*, *2*, 108–115.

Wilson, K. (2007). Opac 2.0: next generation online library catalogues ride the web 2.0 wave! *Online Currents*, *21*(10), 406.

Wojciechowski, R., Walczak, K., White, M., & Cellary, W. (2004). Building virtual and augmented reality museum exhibitions. In *Proceedings of the ninth international conference on 3d web technology* (pp. 135–144). Web3D '04. Monterey, California: ACM. doi:10.1145/985040.985060

Zavalina, O. L., Shakeri, S., & Kizhakkethil, P. (2015). Metadata change in traditional library collections and digital repositories: exploratory comparative analysis. In *Proceedings of the 78th asis&t annual meeting: information science with impact: research in and for the community* (146:1–146:5). ASIST '15. St. Louis, Missouri: American Society for Information Science.

Zeng, R. & Greenfield, P. M. (2015). Cultural evolution over the last 40 years in china: using the google ngram viewer to study implications of social and political change for cultural values. *International Journal of Psychology*, *50*(1), 47–55.