Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges

Johan Oomen
Nederlands Instituut voor Beeld en Geluid
Sumatralaan 45
Hilversum
The Netherlands
+31 35 677 3434
joomen@beeldengeluid.nl

Lora Aroyo
Free University Amsterdam, FEW
De Boelelaan 1081a, 1081 HV
Amsterdam,
The Netherlands
+31 20 598 2868
l.m.aroyo@vu.nl

ABSTRACT
Galleries, Libraries, Archives and Museums (short: GLAMs) around the globe are beginning to explore the potential of crowdsourcing, i.e. outsourcing specific activities to a community though an open call. In this paper, we propose a typology of these activities, based on an empirical study of a substantial amount of projects initiated by relevant cultural heritage institutions. We use the Digital Content Life Cycle model to study the relation between the different types of crowdsourcing and the core activities of heritage organizations. Finally, we focus on two critical challenges that will define the success of these collaborations between amateurs and professionals: (1) finding sufficient knowledgeable, and loyal users; (2) maintaining a reasonable level of quality. We thus show the path towards a more open, connected and smart cultural heritage: open (the data is open, shared and accessible), connected (the use of linked data allows for interoperable infrastructures, with users and providers getting more and more connected), and smart (the use of knowledge and web technologies allows us to provide interesting data to the right users, in the right context, anytime, anywhere – both with involved users/consumers and providers). It leads to a future cultural heritage that is open, has intelligent infrastructures and has involved users, consumers and providers.

Keywords
Crowdsourcing, heritage, metadata, tagging, lifecycle model

1. INTRODUCTION

In his recent book *Cognitive Surplus: Creativity and Generosity in a Connected Age* Clay Shirky observes how the Internet changes the way we spend our spare time [39]. The so-called "cognitive surplus" that used to be spent on passive activities (notably watching television) can now be used in a profoundly different way, for new kinds of creativity and problem-solving. He writes, "the wiring of humanity lets us treat free time as a shared global resource, and lets us design new kinds of **participation and sharing** that can take advantage of that resource." Shirky offers Wikipedia as a compelling example. After calculating that creating Wikipedia as it stands today has taken one hundred million hours of cumulative thought, he juxtaposes this to the astounding 200 billion hours people watch TV in the US alone. 200 billion hours would amount to two thousand Wikipedia projects-worth of free time, annually. The statistics provided by Lasar [24] confirm once again this ever-growing reality, e.g. 35 hours of videos are uploaded to YouTube every minute, and 38,400 photos are uploaded on Flickr every hour, and in total 35% of Internet users have contributed a piece of user-generated content at least once.

The very design of the Internet makes these interactions possible. The core design principle underlying the Web's usefulness and growth is openness and universality. In his recent contribution to the debate on net-neutrality, Tim Berners-Lee notes how social-networking sites are creating silos of information that are only accessible under the conditions set by the entity that manages these sites [5]. According to him, locking up information will eventually hinder innovation. He observes, "when you make a link, you can link to anything. That means people must be able to put anything on the Web, no matter what computer they have, software they use, or human language they speak and regardless of whether they have a wired or wireless connection." All interactions and conversations on

the Web rely on the principle of universality. As a major implication this leads to a "**democratization" of innovation**. Williamson observes how this will empower "millions of people who hitherto had no means of connecting, networking and sharing their unique insights and knowledge" [47].

With the mass uptake of blogging and media sharing sites in the early 2000s, the social dynamic of the Web manifested itself more prominently. Shirky notes how the concept of cyberspace, where computers and networks are regarded as somewhat alien, is now disappearing [26]: "Our social media tools aren't an alternative to real life, they are part of it", he notes, adding that these tools are increasingly the coordinating tools for events in the physical world. Futurist Mark Pesce observes how the **human instinct of sharing** is amplified as the concept of distance evaporates. "The instinctual sharing behavior of humans remains as strong as ever before, but has extended to encompass communities beyond those within range of our voices" [36]. The egalitarian principles that form the foundation of the Internet, combined with our social instinct and explosion of access points to the network has resulted in an age of hyperconnectivity [11, 36].

Contributing masses tirelessly fill the Internet space with their content, e.g. blogs, comments, reviews, tags and multimedia. The agglomeration of individual contributions through online collaboration is having an important social and economic impact [44]. The term *outsourcing* - finding labor elsewhere - gets redefined on the Web as the *crowdsourcing phenomenon*: "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call" [16]. Various organizations are currently exploring ways of engaging the wisdom of the crowd for *creating and editing of content*, *solving problems* or the *organization of knowledge* structures.

In the heritage domain, Galleries, Libraries, Archives and Museums (abbreviated hereafter to 'GLAMs') around the globe are beginning to explore the potential of crowdsourcing. The mass digitisation of analogue holdings is key to heritage organizations becoming an integral part of the Web. In the case of fragile carriers (magnetic tapes and chemical film for instance) digitisation is a means to ensure long-term **preservation** of the information. Digitisation is also a precondition for creating **new access routes** to collections. Once digital and once part of an open network, cultural artifacts can be shared, recommended, remixed, mashed, embedded and cited. In this way attention can be brought to even the most obscure artifacts.

GLAMs and their users are now beginning to inhabit the same, shared information space. New services are being launched that explore this fundamentally new paradigm of participation in the GLAM domain. Participation can have a thorough impact on the workflows of heritage institutions, for instance, by inviting users to assist in the selection, cataloguing, contextualisation, and curation of collections [23]. These activities can be carried out by end-users remotely and can reduce operational costs. These new forms of usage of collections (beyond access) can also lead to a deeper level of involvement with the collections [15].

As funding of many heritage organizations is based on their societal impact, these initiatives will also be of growing importance from a managerial/PR perspective. In this paper, we focus on the potential impact of crowdsourcing, as one of the models in which participation manifests itself.

To study the potential impact, it is important to:

1. Classify the different types of crowdsourcing in the GLAM domain (Sections 2 and 3).
2. See where crowdsourcing can have an impact in the key areas of the so-called Digital Content Life Cycle around which many GLAMs are organized (Section 2).
3. Identify the mutual benefits for all stakeholders (opportunities) and identify potential challenges as a starting point for future work (Section 4).

It is important to undertake these actions, as this new paradigm of participation will have a lasting impact on institutional practice, on the visibility, and hence on sustaining the long-term relevance of GLAMs. Studying the opportunities and challenges will help build a more open, connected and smart cultural heritage.

2. CROWDSOURCING IMPACT ON GLAMs WORKFLOW

In this section, we classify the different types of crowdsourcing initiatives that are currently undertaken in the cultural heritage domain. Specifically, we look at initiatives that are coordinated by GLAMs. We map the different types against a model that describes the different stages of managing digital content projects.

2.1 Classifying the Domain

Several authors have been working on a classification of crowdsourcing projects. For instance, the "Crowdsourcing landscape" by Dawson [16] organizes the different crowdsourcing sites in 15 main categories. It illustrates the breadth of crowdsourcing initiatives, including activities related to clothing design, journalism, stock-picking, translation and fact checking.

More closely linked to the topic of this paper, the recent study by Bonney in *Public Participation in Scientific Research* (PPCR) lists three major models of participation within the domain of research [8]. It is also suitable as a categorisation of projects in the cultural heritage domain:

1. *Contributory projects* - designed by professionals, where members of the public contribute data;
2. *Collaborative projects* - designed by professionals, where members of the public contribute and analyze data, help in refining project design, or disseminate findings;
3. *Co-created projects* - designed by professionals, where members of the public are working together, and some of those public participants are actively involved in (all) steps of a process.

The level of the required skillset of users, and the interaction with the 'host' organisation increases; co-creation will require more engagement than adding a 'tag' as part of a contributory project. In effect, contributory projects will be able to attract a broader community, as the skillset is less specific.

Nina Simon, museum consultant and author of *The Participatory Museum* bases her discussion of "Models for Participation" in the museum domain on the work by PPCR [40]. She added a fourth category to the three listed above, namely *hosted projects*, "in which the institution turns over a portion of its facilities and/or resources to present programs developed and implemented by public groups or casual visitors". This additional category is a conceptual departure from the PPCR model, as it relates to the level of institutional involvement (cf. Section 2.2) rather than the required skillset.

Table 1. Classification of Crowdsourcing Initiatives

Crowdsourcing type	Short definition
Correction and Transcription Tasks	Inviting users to correct and/or transcribe outputs of digitisation processes.
Contextualisation	Adding contextual knowledge to objects, e.g. by telling stories or writing articles/wiki pages with contextual data.
Complementing Collection	Active pursuit of additional objects to be included in a (Web)exhibit or collection.
Classification	Gathering descriptive metadata related to objects in a collection. Social tagging is a well-known example.
Co-curation	Using inspiration/expertise of non-professional curators to create (Web)exhibits.
Crowdfunding	Collective cooperation of people who pool their money and other resources together to support efforts initiated by others.

Classifications like these are helpful in studying the differences between project types from the level of involvement. In our research we take a slightly different vantage point. As we would like to study the impact of crowdsourcing on workflows, we aim to classify the different types according to their **tangible outcomes.** Also, we wanted to create a classification that encompasses working practices at all GLAM domains. This classification is important in order to study the potential impact of crowdsourcing on current working practices in a systematic way. It will help identifying key challenges in operationalizing the concept of participatory GLAMs. It will help to define a research agenda to address these challenges.

We have been gathering examples of crowdsourcing initiatives across the globe. For instance, by studying the proceedings of leading conferences in this area (e.g. Museums and the Web, FIAT/IFTA, AMIA, Museumnext), interviews with practitioners (contacted through the GLAM-WIKI and Europeana Communities, and the Museum-L mailing list), and tracking a number of significant blogs. In studying the motivations and design of various projects and initiatives, the following classification of six main types of crowdsourcing initiatives emerges (Table 1).

As we will examine in Section 4, each type will present different challenges that will impact the success of the crowdsourcing initiative.

2.2 Grassroots Initiatives

Although it falls outside of the scope of this paper, it needs to be acknowledged that crowdsourcing initiatives in the GLAM domain can also be executed without institutions being in the lead. A number of coordinated initiatives by groups of users autonomously are currently being undertaken. To name just three examples as illustrations:

1. Open Plaques[1] - an initiative that aims to gather data about all the commemorative "plaques" across the globe.
2. The Johnny Cash Project[2] - a collective music video, fashioned from drawings done by users from across the web.
3. The International Amateur Scanning League - an experiment in crowd-sourced digitisation to help government and other institutions make their archives more widely available [28].

As the presence of GLAMs on the social web matures, we will begin to see crossovers between community- and organization-driven projects. To some degree, the collaboration between the Wikipedia community and the British Museum (cf. Section 3.2) is moving into this uncharted territory.

We would like to refer the reader to the research of Melissa Terras [45] that provides an in-depth discussion of grassroots initiatives in the heritage domain, zooming in on amateur digitisation. The seminal works of Yochai Benkler [4] and James Boyle [9] provide thorough analyses of the 'bigger picture' behind collaborative projects, investigating how the Internet is reshaping the current economic, social and political structures.

2.3 Crowdsourcing and the Digital Content Life Cycle

In the literature, we found models that define core activities of heritage organizations. These models are used in practice to plan curation and preservation activities for organizations to different levels of granularity. They are useful for heritage organizations as planning tools to organize their resources, and to support management decisions. Notably, the Athena Research Centre [17] examined the model created by the UK-based Digital Curation Centre (DCC). The DCC lifecycle model represents the complex processes found in digital curation in a comprehensive and generic model that can be applied to any discipline. It is widely adopted across the heritage domain. An alternative model proposed by [31] looks at activities undertaken by scholars throughout the research process. It identifies four core activities ("Discover," "Gather," "Create," and "Share") that can also be applied to other domains.

For this paper, we use the Digital Content Life Cycle model from the National Library of New Zealand [27]. It is

[1] http://openplaques.org

[2] http://www.thejohnnycashproject.com

a simplified model, which includes the main activities present in the more extensive and detailed DCC model. It encapsulates the main activities carried out by heritage organizations, through selecting to creating, managing, discovering, using and reusing (including licensing) as well as preservation. The model is cyclical, but it needs to be noted that in daily practice, the order can often differ. For instance, creating descriptions can be done in several stages.

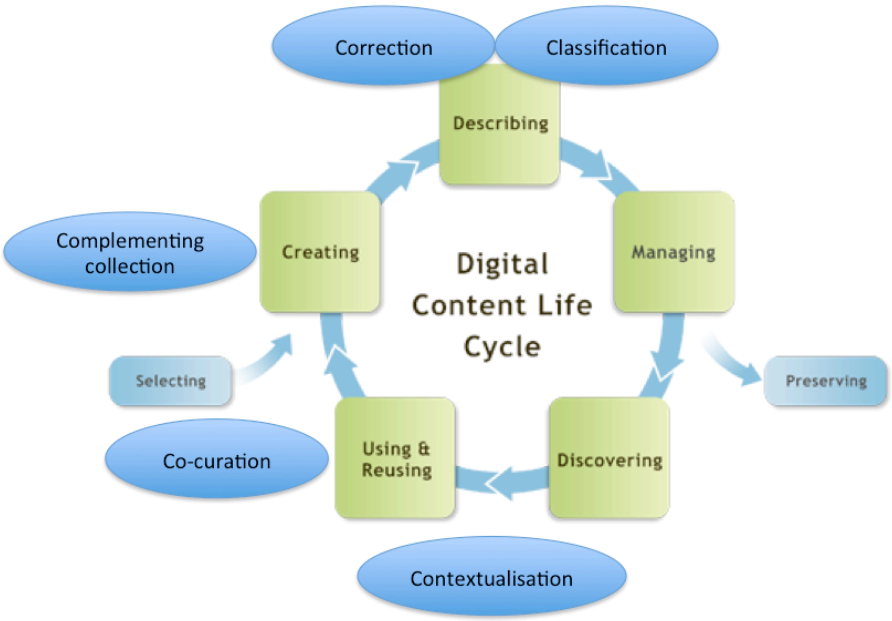

Figure 1. Digital Content Life Cycle and Crowdsourcing

When we look at the relationship between the stages in the Digital Content Life Cycle model and the types of crowdsourcing, the picture emerges that is shown in Figure 1. Crowdsourcing can play a role in all stages of the model: from selection and creation of content, to describing, discovery and use. This clearly underlines the enormous potential of these efforts. The management of the collection itself (including the backup and maintenance strategies of storage facilities) is, at least currently[3], the primary responsibility of professionals; in all other stages of the model, the added value of starting a dialogue between amateurs and professionals is currently being explored.

Five of the six crowdsourcing types we identified can be linked to the stages of the model Figure 1. The sixth one, related to funding, can play a role in each of the stages. Most Crowdfunding initiatives today (cf. Section 3.6) are primarily focusing on projects dealing with the stages 'Using and Reusing' and 'Creating'.

2.4 Knowledge Transfer and Organizational Change

In Section 4. we will explore some of the major challenges related to the successful uptake of crowdsourcing. Here, we will show how current advances in research on scientific areas such as knowledge engineering, human-computer interaction and communication sciences can help to address these. The importance of knowledge exchange between the research domain and the operational services cannot be underestimated [35]. Both stakeholders should invest ample resources in learning each other's vocabulary, working methods and embrace opportunities for joint, multidisciplinary projects. This forms the basis for the establishment of an ecosystem for applied research and ongoing innovation. This activity of knowledge transfer can be regarded as an additional dimension to the Digital Content Life Cycle model.

[3] Note we decided not take peer to peer hosting of files into account. We did not find examples of GLAMs that choose to host their content on peer to peer networks.

3. TYPES OF CROWDSOURCING

In this section, we study the different types of crowdsourcing listed in Table 1 in more detail.

3.1 Correction and Transcription

A typical example of crowdsourcing corrections is the Australian Newspaper initiative from the National Library of Australia. The Library is overseeing the mass digitisation of 830,000 newspaper pages dating from 1803. The newspaper pages are converted into electronically translated, searchable text through the use of Optical Character Recognition (OCR). Using this technology for historical newspapers delivers poor and inaccurate results. The library launched the first service in the world that allows users to correct the OCR'ed text (Figure 2). Without too much active promotion, a subsequent call for user participation in 2008 was greeted with great enthusiasm by end-users. The administrators note that by "October 2009 over 6000 members of the public had already enhanced the data significantly by correcting over 7 million lines of text in 320,000 articles, and adding 200,000 tags and 4,600 comments to articles. One exceptional user has corrected over 285,000 lines of text in over 7,000 articles." [20].

Figure 2. Australian Newsletters

Similarly, the project Transcribe Bentham[4] based at University College London (UCL) is working with a range of end-users to complete the transcription of 12,400 of manuscripts of the philosopher and jurist Jeremy Bentham [30]. In October 2010, a consortium including the National Maritime Museum and the Citizen Science Alliance launched its initiative "Old Weather"[5], that aims to collect data on temperatures from historical ship logs. These detailed logs were kept by ships of the British Royal Navy, that sailed around the world from 1905 to 1929. Sailors wrote down temperature, wind and other climate data every four hours. Users perform a task comparable to the Australian Newspapers project (Figure 3). The speed in which this work is carried out is just stunning. By December 2010, 202,904 pages have been transcribed, 25% of the total amount. With this data, scientists will be able to study how oceans transport heat and water around the globe and try to determine how this affects temperature. The Old Weather project is the latest citizen-based science project by the Citizen Science Alliance community, which has enrolled 349,000 volunteers in to process images of stars, galaxies and other astronomical formations [32].

[4] http://www.ucl.ac.uk/transcribe-bentham

[5] http://www.oldweather.org

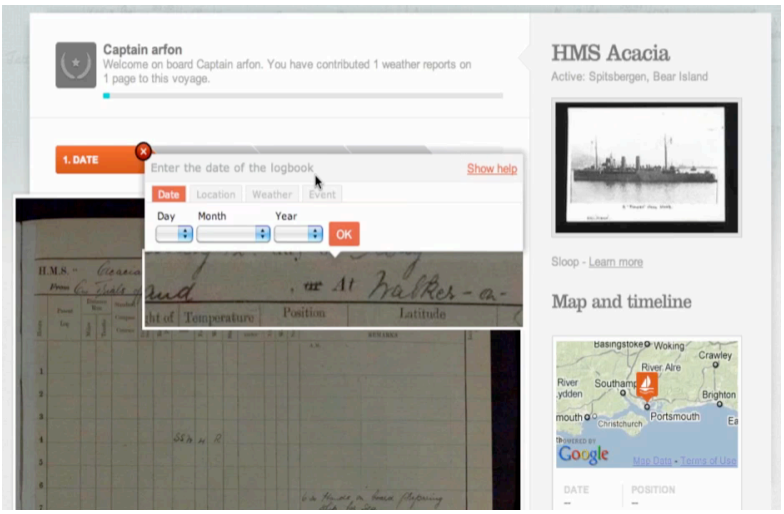

Figure 3. Old Weather: transcribing ship logs

A fourth example of this type of crowdsourcing is New York Public Library's Map Rectifier Project (Figure 4). This is an online environment in which the public aligns ("rectifies") historical maps from the NYPL's collections to match today's precise maps [19].

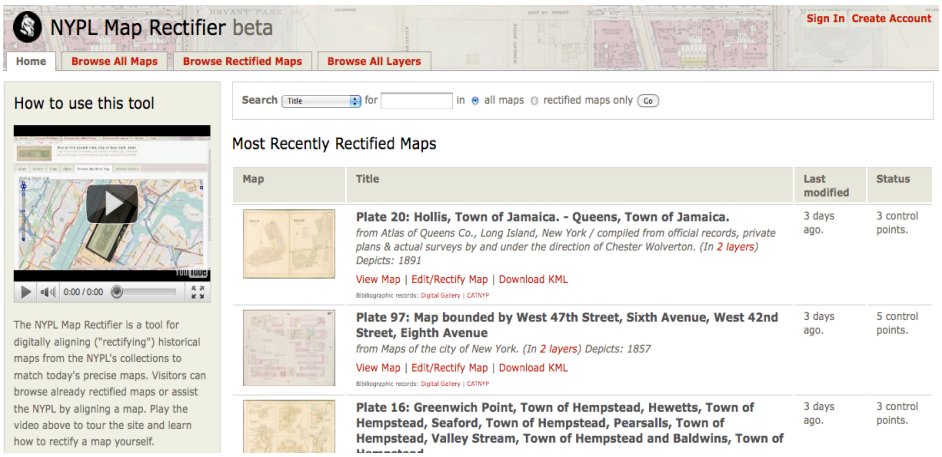

Figure 4. Aligning historical maps

The outcome of this activity will make it possible to create visualizations showing changes in maps over time.

3.2 Contextualisation

The term 'contextualisation' has many connotations in the heritage domain. Here, we propose to frame contextualisation in the cultural heritage domain as activities that aim to place or study objects in a meaningful context. Contextualisation has always been part of the 'mission statement' of cultural heritage organizations. Or, in the words of Bruce Sterling "The grand plan here is to protect the legacy of the past while also ensuring one's relevance to the present and future" [42].

There is a long tradition of contextualisation of content in collections by a wide range of users, including scholars, amateur historians and other enthusiasts. They have done so by writing scientific publications, compiling magazines that document the history of the city they live in, studying their family histories, using archival footage as illustrations for monographs and so on. The involvement of curators, librarians and archivists in these private/scholarly endeavors range considerably, from looking up information to pre-processing data. These interactions between professionals and 'amateurs' are now also taking place online, using an impressive variety of tools and platforms.

The project 1001 Stories Denmark[6] for instance, based at the Danish Heritage Board offers an impressive insight in the history of Denmark by linking objects from contributing heritage institutions to times, places and (perhaps most interestingly) personal stories contributed by end-users that provide context. End-user contributions are a key feature of the interactive design of the portal (Figure 5), giving users explicit attribution for their additions, and prominently inviting users to contribute their own stories about an object.

Figure 5. 1001 Stories

Wiki-style platforms are embraced by many heritage organizations as a means to 'harvest' contextual knowledge from their user base, as they are a way both to facilitate collaborative contributions and to track the history of successive contributions from multiple users. The Netherlands Institute for Sound and Vision, for example, uses a wiki platform to gather contextual information on television programmes, broadcasters, presenters and so on[7]. The service was launched in 2008 and by 2010 over 38,000 pages have been added by almost 2,000 users. The information on the wiki pages will be linked to the catalogue of the Sound and Vision, which is managed by professional cataloguers. Sound and Vision actively solicits contributions by the community, for instance by collaborating with several media studies departments, where contributing content to the wiki is part of a research Masters course.

Collaborations between heritage organizations and the Wikipedia community will have a great impact on the way contextual knowledge will be added to cultural heritage content. A first collaboration was initiated by a consortium of US/UK based museums in 2008, within an initiative called "Wiki Loves Art", that aimed to increase the amount of images from museum objects on Wikimedia Commons (the media repository of open content hosted by the Wikimedia Foundation)[8]. This has been repeated a number of times since. For a limited time period, participating institutions open their doors for users to take photographs that are subsequently uploaded to Wikimedia Commons. Contextual information is added as these pictures are attached to Wikipedia pages. In the summer of 2010, a so-called "Wikipedian" (i.e. an individual contributing to Wikipedia) joined the British Museum as part of the museum's first "Wikipedian in residence" programme. Activities[9] consisted of the in-depth curation of Wikipedia pages on masterpieces in the collections of the British Museum. This included one-on-one collaborations, where individual Wikipedians worked with curators on a particular topic [13]. In many cases, the Wikipedia pages now offer more detail than the information available in the museum space itself. The museums decided to use mobile devices (co-called Wikireaders) to offer visitors access to this resource. Another effect of the collaboration has been the increase of traffic to the Website of the British Museum, as the Wikipedia pages include deep links to the institute's website [49].

[6] http://www.kulturarv.dk/1001fortaellinger/en_GB

[7] http://beeldengeluidwiki.nl/index.php/Hoofdpagina

[8] http://commons.wikimedia.org/wiki/Category:Commons_partnerships

[9] https://secure.wikimedia.org/wikipedia/en/wiki/Wikipedia:GLAM/BM

3.3 Complementing Collections

Crowdsourcing can be employed in order to fill gaps in collections. A good example is the UK_Soundmap project[10], launched by the British Library in July 2010. The British Library's Sound Archive wanted to facilitate research into the changing "soundscape" of the UK by providing a rich corpus of sounds. From the UK_Soundmap website: "By capturing sounds of today and contributing to the British Library's digital collections you can help build a permanent researchable resource."[11] The British Library decided to invite users to provide these sounds, using the mobile application Audioboo as one of the main instruments. Users install this application on their smartphone, make recordings and subsequently upload them, together with some contextual metadata including a geo-coordinate. After six months, at the project's mid-way point, the British Library had managed to gather 1,200 sounds though this project. The sound clips are placed on an interactive map (Figure 6) and are also searchable through a set of metadata.

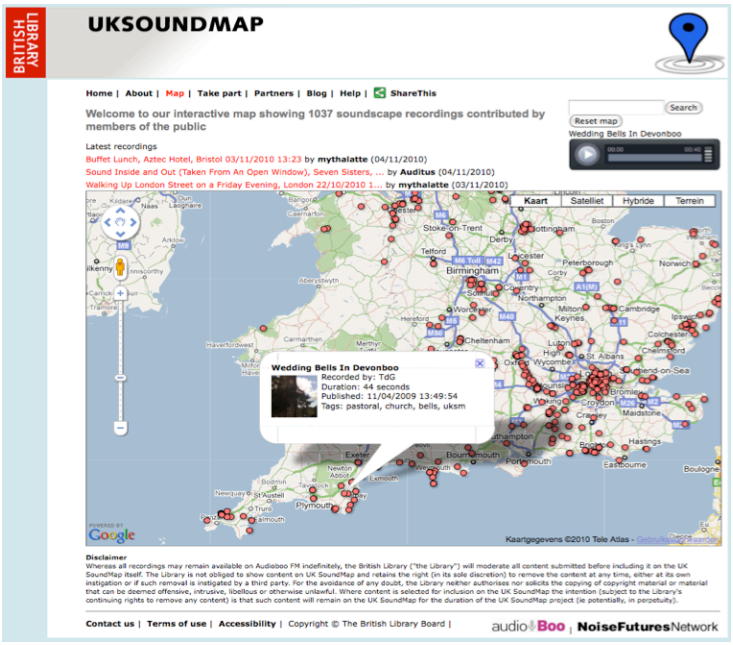

Figure 6. Interface showing the sounds on a map

Wir Waren So Frei is a cooperative project by the Deutsche Kinemathek and the Bundeszentrale für politische Bildung has gathered an impressive collection of images related to the fall of the Berlin Wall by issuing an open call for contributions of content and the stories behind them. Here, the Kinemathek took the responsibility for digitising the photographs and small-gauge films that were contributed by the public, resulting in a unique resource including almost 7,000 items that are available online (Figure 7).

In the UK, the RunCoCo[12] initiative delivers training and support to groups wishing to run community collections. The project is developing training materials, and organizing and running workshops to support projects wishing to follow the community collection model. The project is a direct continuation of the work done for the University of Oxford's Great War Archive, which created an online resource of 6,500 items contributed by the general public[13]. This project is also developing the open-source CoCoCo (community contributed content) software to make it available for any other project to collect user-generated content via the Web.

[10] http://sounds.bl.uk/uksoundmap/index.aspx

[11] http://sounds.bl.uk/uksoundmap/index

[12] http://runcoco.oucs.ox.ac.uk

[13] http://www.oucs.ox.ac.uk/ww1lit/gwa

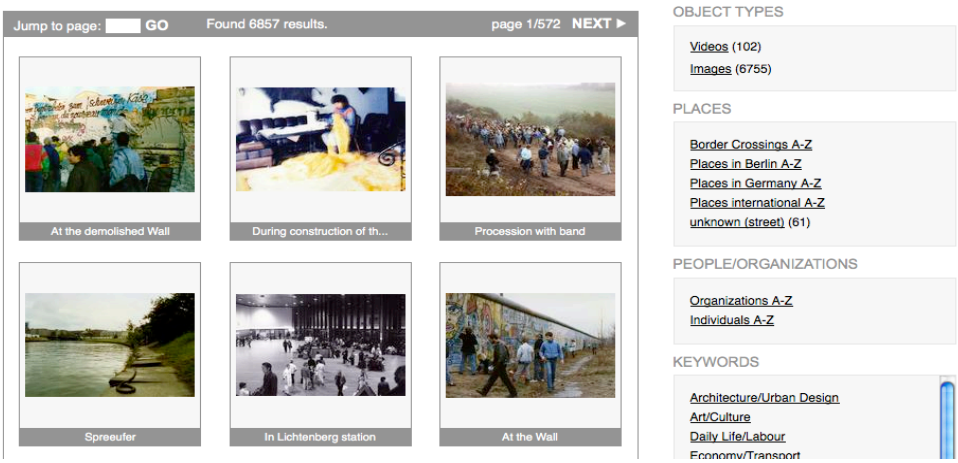

Figure 7. Wir Waren So Frei

A final example of organizations soliciting objects to be added to their collection is the Wedding Fashion[14] initiative from the V&A museum in the UK, that creates a database of photographs of clothes worn for weddings between 1840 and the present. In order to ensure the creation of a useful historical record all entries are accompanied by the year of the event and the names of the bride and groom or partners [17].

3.4 Classification

Social tagging has grown to be a popular way for institutions to explore the potentially positive implications of presenting their collections on-line. steve.museum [46], the first large-scale project to explore the concept of tagging by "the crowd" in the heritage domain, was launched in 2005 [3]. It brings together a number of US and UK based museums that collaboratively "explore the role user-contributed descriptions play in improving on-line access to works of art" [46]. An online environment[15] (Figure 8) was created that allows registered users to add tags to a selection of works from participating museums.

In little over two years, steve.museum managed to gather 36,981 terms, comprising 11,944 terms in 31,031 term/work pairs. At the end of 2010, this number has risen to a stunning 468,120. Tagging is shown to provide a significantly different vocabulary to museum documentation: 86% of the tags contributed through the steve.musem tagging environment were not found in museum documentation [46].

Dozens of GLAMs have embarked on similar projects to steve.museum [46]. For instance, the Powerhouse Museum in Sydney launched their social tagging project in 2006. When a user submits a tag, it is incorporated in the online catalogue. Tags can also be corrected and removed by other users if they are deemed incorrect. Within six months, almost 4,000 tags were added to over 2,200 objects in the online catalogue. Over 500 of these tags "were deleted, edited for spelling, or removed by other users and the system administrator" [12].

[14] http://www.vam.ac.uk/things-to-do/wedding-fashion

[15] http://tagger.steve.museum

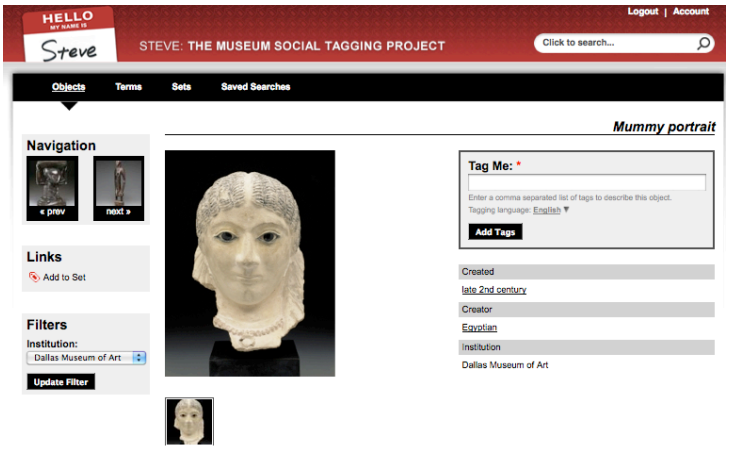

Figure 8. steve.museum tagging interface

In January 2008, the Library of Congress (LoC) published a set of about 3,000 pictures on Flickr, with the goal to reach out "to unknown as well as known audiences" and to collect information about these photos through the audiences' comments and tags [41]. The launch was heavily advertised in the blogosphere and, within a day of the launch of the project, the collection of images had been viewed over a million times. The LoC photo set on Flickr has been expanded gradually since, and still receives about 500,000 views monthly [41]. During the first ten months of the project, the Flickr community placed over 7,000 comments on more than 2,800 pictures. People often commented on the aesthetic qualities of the pictures, but a lot of additional factual information was added as well. Within this timeframe, a total of 2,518 people left over 67,000 tags. Of these tags, 1,000 (21%) were unique. On average, 14 tags were added to each photo [41]. The LoC and the photo sharing website Flickr later teamed up to develop a communal page for other cultural heritage institutions with photography collections: Flickr: The Commons[16]. To date, over forty organizations have joined The Commons, and it has established itself as one of the most prominent examples of social tagging.

In 2009, a consortium including the Netherlands Institute for Sound and Vision, KRO broadcasting, and VU University Amsterdam launched the video labeling game called Waisda? (Figure 9). It uses gaming as method to annotate television heritage. Similar to the 'Games With A Purpose' serious gaming concepts developed by Von Ahn, players receive points if their tag matches a tag that their opponent has also typed in within a given time-frame [1]. The game-play of Waisda? focuses on reaction and precision. From the point of the archive, the underlying assumption is that tags are probably valid if there is mutual agreement between players.

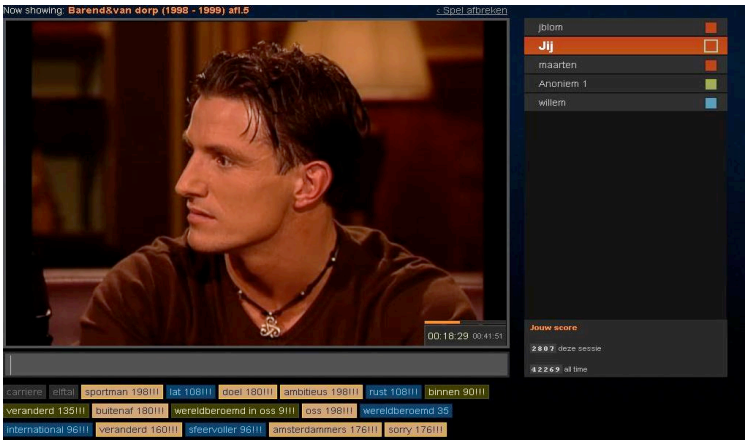

Figure 9. Waisda? Tagging interface

16 http://www.flickr.com/commons/institutions/

Within a period of 7 months, 340,000 tags were added though Waisda?, of which 40.3% are of matching tags, i.e. tags added by two more players within a time frame of 10 seconds [33].

3.5 Co-curation

Projects belonging to the category co-curation focus on the interaction between users and institutions regarding selection activities for (online) publication. "Click! A Crowd-Curated Exhibition" from the Brooklyn Museum is a good example. The museum invited artists to submit electronically a work of photography that responded to the exhibition's theme, "The Changing Faces of Brooklyn", plus a 100-word artist statement. 389 photographs were submitted and subsequently judged by the public, using an custom built evaluation tool. 3,344 evaluators cast 410,089 evaluations and the top 78 images were put on display in the museum [43]. Remarkably, there was a lot of agreement between the crowd's judgment and the judgment of the experts.

Figure 10. Expose: my favorite landscape

A second example is a campaign initiated by the Dutch modern art museum Kröller-Müller Museum earlier this year. The museum invited children to select their favorite landscape from the museum's collection (Figure 10). The 20 works of art with the highest number of votes were put on display in the Winter 2010 exhibition.

The Danish broadcaster DR used a similar methodology in their Bonanza project, which invited the audience to vote for their favorite show from the archive collections to be digitised and made available on-demand first. Bonanza was a great success, as noted by DR in [7]: "The site had more than 12,000,000 video streams during the voting which is remarkable for a country with a total population of about 5,500,000".

3.6 Crowdfunding

The final type in our classification is Crowdfunding. This activity refers to the collective cooperation of people who pool their money and other resources together to support efforts initiated by others. The Louvre, for instance, recently managed to raise one million Euro's from online donors to buy a Renaissance painting by Lucas Cranach the Elder. Within a few weeks after the appeal was announced, 5,000 donors responded, donating an average 150 Euros [11].

Initiatives like Kickstarter [26], IndieGoGo, RocketHun, and Voor de Kunst[17] can serve as funding platforms for both artists and cultural heritage organizations. The funding mechanisms on these platforms are quite similar. Creators indicate the target amount, and duration of the campaign. Visitors begin pledging contributions, committing to donate the promised amount if the project reaches or exceeds its funding goal before time expires. Recent research, based on data from 12 leading crowdfunding sites, found that approximately $80 million has been pledged until early 2011. An estimated one million people were responsible for these pledges [2]. However,

[17] http://www.kickstarter.com, http://www.indiegogo.com, http://www.rockethub.com, http://www.voordekunst.nl/

the study indicated that not all projects get funded, and returns for the crowdsourcing platforms themselves appear modest.

4. CHALLENGES

Most of the activities in the section above existed in some form before the Web became what it is today. For instance, quite a lot of heritage institutions have been working with volunteers coming to the institute to help to assist on cataloging of curatorial tasks. Thus crowdsourcing can be seen as a remediation: the effect of new media on old media, causing old media to 'refashion' themselves [6]. The essential difference are those of scale, connectedness between 'e-volunteers', and ease of use, which are fundamentally different in a global online environment thanks to the universality principle of the Internet.

New technology can be used to overcome major challenges in *turning the potential of cognitive surplus into a key asset*. As Holley [20] notes "Libraries and Archives will never have the resources to fully do what they or the users want, so crowdsourcing is an opportunity that should be seriously considered". Technology can ensure a maximum impact of these initiatives by combining the strengths of the machines and the knowledge, common sense, and potential of the human crowd. Important here is to seek the optimal combination for a user-friendly and shared-initiative interaction.

4.1 Identifying Critical Challenges

As mentioned in Section 2.4 an important aspect in achieving open, connected, and smart cultural heritage where consumers and providers are closely involved, is to provide an innovative environment (ecosystem) where the "inhabitants explore the adjacent possible, because they expose a wide and diverse sample of spare parts - mechanical or conceptual - and they encourage novel ways of recombining those parts" (Stephen Johnson, "Where good Ideas come from"). In this section we focus on some of the technological challenges that require multidisciplinary teams working in all the phases of the Digital Content Life Cycle to realize a functional and successful deployment of crowdsourcing in the life cycle.

Currently Semantic Web techniques and methods appear to gain quite some momentum in their deployment in Social Web applications and other mainstream tools. For example, Facebook's use of Open Graph for connecting people and content items across applications, as well as Google's semantics-based search and auto-completion. In this work we explore further the challenges related to the application of such techniques for crowdsourcing in cultural heritage.

Currently, Semantic Web techniques are aiming to (1) improve the understanding of machines of different knowledge domains, (2) aid their reasoning, and (3) discover serendipitous links between items in the collections. In addition, using linguistic, image and video analysis techniques, builds the basis for a new generation of collections with support for quickly growing amount of objects and where annotations capture diverse set of dynamics and perspectives. For example, (1) integrating both professional and amateur perspectives, (2) combining depicted and contextual annotations, (3) allowing for diversity in the type, level of specificity and granularity of the annotations, (4) allowing for multimodal, mixed-initiative, interactive exploration, (5) interlinking with objects from other collections and additional information from external sources.

Challenges that are **related to Semantic Web techniques:**

- Dealing with complex underlying knowledge is challenging in terms of providing explanations to the users.
- Having simple interaction interfaces with a multitude of complex, analytical, summary and interlinked information.
- Providing scalable and robust solutions.
- Stimulating users to contribute specific types of knowledge through engaging them via semantic-based tags and suggestions.

Challenges that are **related to Linguistic techniques:**

- Offering proper exploitation and presentation of multilingual information.
- Providing efficient and effective quick learning mechanisms.

Challenges that are **related to Quality of the data:**

- Maintaining/resolving conflicting information.
- Maintaining and presenting extensive (ever growing) provenance information.
- Creating open and clear reviewing procedures.
- Evenly distributing the contributions of the users over the entire collection.
- Indicating when an annotation is 'good' or 'finished'.

Although all the challenges listed above are at some level important for successfully maintaining the new generation of cultural heritage collection, in this paper we will focus on two critical ones: to (1) bootstrap the process with sufficient knowledgeable and loyal-over-time users, and to (2) maintain a reasonable level of quality, in order to sustain the existing levels of reputation, or to expand it.

4.2 Motivations – Gathering Loyal Users

Users engage in crowdsourcing for either extrinsic motivations or intrinsic motivations. Amazon's Mechanical Turk is probably the most famous example of a platform that is built around interactions based on extrinsic motivations [34]. Mechanical Turk employs individuals (so-called Workers in Amazon's jargon) to perform simple tasks in return for monetary payment. The service has (January 2011 data) over a half a million Workers worldwide from over 190 different countries, performing a wide variety of tasks, ranging from creating subtitles, categorising websites, counting instances of words in audio files and so on.

The examples in Section 3 are all focused on intrinsic motivations. Here, both GLAMs and their users benefit from mutual recognition. It will foster a more profound way of engagement [25]. On the merits of tapping in to intrinsic motivation, Clay Shirky notes "Amateurs are sometimes separated from professionals by skill, but always by motivation; the term itself derives from the Latin amare—to love. The essence of amateurism is intrinsic motivation: to be an amateur is to do something for the love of it." [39]

Table 2. Motivational factors: two clusters

Motivation	Examples
Connectedness and membership	**Old Weather** builds upon the existing Citizen Science Alliance community that brings together a community of many thousands of volunteers. The enthusiasm of this enormous volunteer army can be leveraged for many projects.
	The Brooklyn Museum tagging project created an online environment where the ***posse members*** (taggers) can meet[18]. This clearly taps into the feeling of belonging to a group.
	Contributors to the **UK_Soundmap** receive a personal thank you note for each recording they upload, e.g. through Twitter messages.
	The **Great War Archive** organized a series of 'real live' events supporting the collection of items that are placed in the online archive.
Sharing and Generosity	The altruistic nature of playing the ***Waisda?*** **Video Labeling Game** is made explicit: "Help to improve access to audiovisual archives"
	Flickr the Commons has goals to show its users (1) "the hidden treasures in the world's public photography archives", and (2) how their "input and knowledge can help make these collections even richer."
	British Museum curators worked together with Wikipedians in the "Wikipedian in residence" programme for sharing knowledge between amateurs & professionals.
	Initiatives aiming at complementing the collections with content contributed by users, e.g. **Wir Waren So Frei** and **Flickr the Commons,** use open licenses that allow the reuse of content on other platforms. Here, sharing is an integral part of the design of the service.

18 http://www.brooklynmuseum.org/community/posse/

Discussing social motivations, Clay Shirky points to the work of Yochai Benkler and Helen Nisembaum: “They divide social motivations into two broad clusters - one around connectedness or membership and the other around sharing and generosity” [39]. Both types of social motivations have been taken into consideration in analyzing the successful examples listed above. In Table 2 we organize the projects according to the functionality they provide to motivate the crowd to contribute.

Next to these social motivations listed above, altruism, fun and competition are also regarded as important incentives for users to participate [33].

4.3 Quality assurance

Despite the fact that we are witnessing an explosion of user-generated content on the Web, only a small portion of people contribute most of it. About 90% of the online users only consume content and from the 10% left only 1% actively and consistently contribute the majority of the user-generated content [19]. Another issue here is the quality of this content. As noted by Foresman [18], 95% of this content is either spam or malware. Thus, motivating not only for participation but also supporting quality contributions appears to be a major challenge.

GLAMs earned their reputation over the years by preserving the quality and truthfulness of the information they offered by having full control over the acquisition, organization and the annotation of the collection items. As is often noted, for example by the Europeana initiative [50], one of the distinguishing qualities of heritage organizations is their authority: provide context and trusted factual information. Nowadays, online search engines and “the people formally known as the audience” [37] can easily perform the same activities. This could be seen as a threat to the position of heritage institutions. Allowing the end-users to actively participate, for instance by adding descriptive metadata to catalogues, could corrode this (perceived) qualitative distinction between users and organization staff [10]. Thus, a fundamental change is required of the *old in-situ culture* based on controlled authority and the *new in-vivo reality* based on the wisdom of the crowd and crossing various geographical, age and competency boundaries. Even if the institutions embrace the new style of building reputation (not based on the distinction between amateur/professional, but on the merits of the contributor’s knowledge on the subject), they are still facing a quality assurance challenge in an open, decentralized space.

It needs to be acknowledged that the social web is not all "Blue skies and sunshine”. The last quote is from Andrew Keen from his book “The Cult of the Amateur” [17], a major critique of peer production, user generated content and phenomena related to the social web. He points at the merits of the expert-based filtering process as beneficial to the quality of information and criticizes how the advent of participatory culture undermines this process. Other authors, for instance Mirko Tobias Schäfer [38] and Jonathan Zittrain [51] also point to potential risks that are unquestionably linked to the concept of participatory culture. But their critiques take a more constructive vantage point.

After analyzing the initiatives, we can conclude most of the GLAMs are very much aware of potential pitfalls in working with the public. For instance, the institutions evaluated the quality of the contributions by end-users. In most cases, the benefits outweigh the caveats. As noted by *Wikipedian in residence* Liam Wyatt: “Unknown risks are accounted for, overestimated, unknown rewards are discounted, underestimated” [49]. As referred to in [23] knowledge is created though conversation. This includes controlling quality in these crowdsourcing initiatives. A combination of technological and interaction aids, psychology principles and community building rules can help to (1) establish behavioral norms, (2) build an image of the desired quality of content, and (3) filter or correct erroneous information. For example, in the *Waisda?* Video Labeling Game, the community itself acts as a filter, as only those terms for which there is mutual agreement between players are considered for inclusion in the archive. Next to this, interactive user feedback is used in order to support users in learning the aspects of good quality contributions. Finally, the creation of a strong sense of belonging to an altruistic community, and making explicit the mutual benefits of the contributed tags, attracts users with diligence and ethical behaviour.

5. CONCLUSION

We have shown how GLAMs are currently leveraging the ‘cognitive surplus’ of their user base. By classifying ongoing projects, and mapping these against current work processes (following the Digital Content Life Cycle

model) we can conclude that there is an enormous potential for GLAMs to explore making crowdsourcing into an integral part of their workflow. GLAMs need to be aware of motivational factors, as participation of users is key to the success of these projects. Also, we've shown how technology can aid institutions to improve the quality of contributions, for example by applying filters or linking with external resources.

Crowdsourcing has the potential to help build a more open, connected, and smart cultural heritage with involved consumers and providers: **open** (the data is open, shared and accessible), **connected** (the use of linked data allows for interoperable infrastructures, with users and providers getting more and more connected), and **smart** (the use of knowledge technologies and Web technologies allows us to provide interesting data to the right users, in the right context, anytime, anywhere). It is of crucial importance for all stakeholders to invest in knowledge transfer from the research to operational services.

We envision the future cultural heritage to be open, built on intelligent infrastructures and on the concept of participation between the various stakeholders. This will allow GLAMs to excel in terms of knowledge, applications and technologies for the wide range of end users they cater for.

ACKNOWLEDGMENTS

We would like to express our gratitude to Lotte Belice Baltussen, Mia Ridge, Mike Ellis and Liam Wyatt for their helpful contributions. A special thank you to Alexandra Eveleigh for reviewing the final version of the paper. This research has been supported by the NWO project Agora and the EU FP7 project PrestoPRIME.

REFERENCES

1. Ahn, Luis von. Games with a Purpose. IEEE Computer 39(6): 92-94, 2006.
2. Andrews, Robert. Is Crowdfunding Working? Here's What We Know. February 8, 2011. Available at: http://paidcontent.org/article/419-is-crowdfunding-working-heres-what-we-know/
3. Bearman, David and Jennifer Trant. "Social Terminology Enhancement through Vernacular Engagement: Exploring Collaborative Annotation to Encourage Interaction with Museum Collections" *D-Lib Magazin*e. September 2005, Volume 11, Number 9
4. Benkler, Yochai. *The Wealth of Networks: How Social Production Transforms Markets and Freedom.* New Haven. Conn: Yale University Press, 2006.
5. Berners-Lee, Tim. Long Live the Web: A Call for Continued Open Standards and Neutrality. *Scientific American, November 22, 2010*
6. Bolter, J.D. and Grusin, R. *Remediation: Understanding New Media*. Cambridge, MIT Press, 1999
7. Bonanza Story: Launching the Cultural Heritage Project in the Danish Broadcasting Company. Available at: http://www.dr.dk/Kulturarv/Bonanza.htm
8. Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., and Wilderman, C. *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), 2009.
9. Boyle, James. *The Public Domain: Enclosing the Commons of the Mind.* New Haven, Conn: Yale University Press, 2008. Print.
10. Brothman, Brien. *Declining Derrida: Integrity, Tensegrity, and the Preservation of Archives from. Deconstruction.* Archivaria, 1999.
11. Carvajal, Doreen. 5,000 Donors Help Louvre Buy a Painting. Available at: http://www.nytimes.com/2010/12/18/world/europe/18iht-louvre18.html?_r=3.
12. Chan, Sebastian. Tagging and Searching – Serendipity and Museum Collection Database. (2007*) In Museums and the Web 2007: Proceedings*. Toronto: Archives & Museum Informatics, ed. Jennifer Trant and David Bearman. http://www.archimuse.com/mw2007/papers/chan/chan.html.

13. Cohen, Noam. Venerable British Museum Enlists in the Wikipedia Revolution. Available at: https://www.nytimes.com/2010/06/05/arts/design/05wiki.html.
14. Constantopoulos, Panos, Costis Dallas, Ion Androutsopoulos, Stavros Angelis, Antonios Deligiannakis, Dimitris Gavrilis, Yannis Kotidis, and Christos Papatheodorou. *DCC&U: An Extended Digital Curation Lifecycle Model.*International Journal of Digital Curation 4, no. 1 2009.
15. Huvila, I. Participatory archive: towards decentralized curation, radical user orientation and broader contextualisation of records management. *Archival Science*, 2008, 8 (1), 15-36. (Springer).
16. Dawson, Ross. Crowdsourcing landscape Available at: http://crowdsourcingresults.com/competition-platforms/crowdsourcing-landscape-discussion.
17. Durbin, G., More than the Sum of its Parts: Pulling Together User Involvement in a Museum Web Site. *In J. Trant and D. Bearman (eds). Museums and the Web 2010: Proceedings. Toronto: Archives & Museum Informatics.*
18. Foresman, Chris. All that user-generated content?. Ars Technica. Available at: http://arstechnica.com/web/news/2010/02/all-that-use-all-that-user-generated-content-95-is-malware-spam-r-generated-content-95-of-is-malware-spam.ars.
19. Hinkle, Karyn. NYPL's Map Division gets more amazing by the moment. Available at: http://bgc-apps.rubensteintech.com/blogs/bgc-library-blog/archives/212.
20. Holley, Rose. Crowdsourcing: How and Why should Libraries do it? *DLIB Magazine, March/April2010*
21. Johnson, Steven. *Where Good Ideas Come from: The Natural History of Innovation*. New York: Riverhead Books, 2010.
22. Keen, A. *The cult of the amateur: How today's Internet is killing our culture*. New York: Doubleday/Currency, 2007.
23. Lankes RD, Silverstein J, Nicholson. *Participatory networks: The library as conversation.* Syracuse University, 2007
24. Lasar, Matthew. Most Internet time now spent with social networks, games. Available at: http://arstechnica.com/web/news/2010/08/nielsen-social-networking-and-gaming-up-email-uncertain.ars.
25. Leadbeater, Charles. *We-think: The Power of Mass Creativity.* London: Profile, 2007.
26. Levy, Shawn. Kickstarter raises money online for artistic endeavors, tapping into Portland ethos. Available at: "http://www.oregonlive.com/living/index.ssf/2010/05/kickstarter_raises_money_onlin.html.
27. Make It Digital Guides. National Library of New Zealand. Available at: http://makeit.digitalnz.org/guidelines.
28. Malamud, Carl. International Amateur Scanning League. Available at: http://radar.oreilly.com/2010/02/international-amateur-scanning.html
29. Mechanical Turk Office Hours. Available at: http://mechanicalturk.typepad.com/blog/2011/02/aws-office-hours-recap-.html
30. Moyle, M. and Tonra, J. and Wallace, V. Manuscript transcription by crowdsourcing: Transcribe Bentham. *LIBER Quarterly*. 20 (2011).
31. A Multi-Dimensional Framework for Academic Support. University of Minnesota, 2005.
32. Niiler, Eric. WWI-Era Ships Enlisted as Climate Guideposts: Citizen scientists can help researchers sift through weather data from these ships, providing critical info about Earth's climate history. Available at: http://news.discovery.com/earth/wwi-ships-climate-weather.html.
33. Oomen, Johan and Belice Baltussen, Lotte and Limonard, Sander and van Ees, Annelies and Brinkerink, Maarten and Aroyo, Lora and Vervaart, Just and Asaf, Kamil and Gligorov, Riste (2010) Emerging Practices in the Cultural Heritage Domain - Social Tagging of Audiovisual Heritage*. In: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, April 26-27th, 2010, Raleigh, NC: US.*

34. Panagiotis G. Ipeirotis. Analyzing the Amazon Mechanical Turk marketplace. *XRDS* 17, 2 (December 2010), 16-21.
35. Peacock, D., Digital ICTs: Driver or vehicle of organisational change? *, in International Cultural Heritage Informatics Meeting (ICHIM07): Proceedings, J. Trant and D. Bearman (eds). Toronto: Archives & Museum Informatics. 2007.*
36. Pesce, Mark, The New Tookit. February 20, 2011. Available at: http://blog.futurestreetconsulting.com/2011/02/20/the-new-toolkit.
37. Rosen, Jay. The People Formerly Known as the Audience. *Huffington Post June 30, 2006.* http://www.huffingtonpost.com/jay-rosen/the-people-formerly-known_1_b_24113.html.
38. Schäfer, Mirko Tobias. *Bastard culture!: how user participation transforms cultural production.* Amsterdam: Amsterdam University Press, 2011.
39. Shirky, C. *Cognitive surplus: Creativity and generosity in a connected age*. New York: Penguin Press, 2010.
40. Simon, Nina. *The Participatory Museum*. Santa Cruz, California: Museum 2.0, 2010.
41. Springer et al., For the Common Good*: The Library of Congress Flickr Pilot Project* Washington: Library of Congress, 2008.
42. Sterling, Bruce. Revisions of Digital Culture. In: Me You and Everyone *We Know Is A Curator.* Graphic Design Museum, March 2009.
43. Surowiecki, James. Reflections on Click! Available at: http://www.brooklynmuseum.org/community/blogosphere/2008/08/08/reflections-on-click-by-james-surowiecki/.
44. Tapscott, Don, and Anthony D. Williams. *MacroWikinomics rebooting business and the world.* Soundview Executive Book Summaries, 2011.
45. Terras, M. *Digital Curiosities: Resource Creation Via Amateur Digitisation*. University of Maryland, 2009.
46. Trant, Jennifer, Tagging, Folksonomy and Art Museums: Results of steve.museum's research. 2009. Available at: http://conference.archimuse.com/jtrants/stevemuseum_research_report_available
47. Williams, Anthony. Wikinomics and the Era of Openness: European Innovation at a Crossroads. *The Lisbon Council e-brief*. Issue 05/2010.
48. Wyatt, Liam. End of my residency. Available at: http://www.wittylama.com/2010/07/end-of-my-residency.
49. Wyatt, Liam. Wikipedia and GLAMS, November 2010. Presentation at the Europeana conference, Amsterdam. Available at: http://prezi.com/fdjjl0tfpghu/wikipedia-glams.
50. Zeinstra, Maarten and Paul Keller. *Open Linked Data and Europeana.* Kennisland, 2010.
51. Zittrain, Jonathan. Minds for Sale. Available at: http://cddrl.stanford.edu/news/jonathan_zittrain_on_minds_for_sale_20110106.