Crowdsourcing in Web Archiving

A report on the crowdsourcing workshop on 4 May 2012
IIPC General Assembly April 30 - May 4, 2012I

1. Introduction

Crowdsourcing has been increasingly used by cultural heritage institutions to involve users in the development and curation of their collections. There are currently not many known or documented cases of applying crowdsourcing in Web Archiving. IIPC member institutions however have realised the potential and relevance of crowdsourcing for a number of reasons:

- Institutions operate large web archiving initiatives and programmes with minimal resource. As web archive collections grow institutions are in need for more manpower.
- Crowdsourcing is particularly attractive in the current economic climate.
- Crowdsourcing seems an effective way to reach out to stakeholders and increase the awareness of web archives.

The workshop was funded by the IIPC to explore the role of crowdsourcing in web archiving. While taking advantage of the annual General Assembly where representatives of member institutions meet physically, it was intended to stimulate discussions and to identify key web archiving workflows in which the "crowd" can contribute. The workshop aimed to deliver a number of crowdsourcing use cases for web archiving. In addition it aimed to provide a starting point from which institutions can initiate crowdsourcing projects and exchange best practices.

This report brings together the relevant materials used for the workshop, summarises the key elements and describes the identified uses cases. It will be circulated to the IIPC membership for comments and reference.

2. Format of the workshop

The workshop was advertised as a part of the IIPC General Assembly programme and participants were requested to register for the workshop. In order to ensure effective discussion, the number of participants was capped at 22. A discussion paper was circulated to the participants prior to the workshop, describing the concept of crowdsourcing. The paper considers examples of successful projects in the cultural heritage sector and how they relate to key stages of the web archiving workflow. It also suggests a list of further reading on the topic (see Appendix 1).

The day of the workshop started with a brief introduction of the participants, on their interest and experience with crowdsourcing and what they expected to get out of the workshop. This was followed by a thought-provoking presentation by Trevor Owens of the Library of Congress, who also chaired the all-day workshop. The workshop was further formed around two break-out sessions, looking at key features of existing crowdsourcing projects and developing use cases for web archiving. The 16 participants were divided in 4 break-out groups, each working on an example or a use case. At the end of each break-out session, the groups reported back on their findings.

The detailed workshop schedule is enclosed as Appendix 2.

3. Understanding crowdsourcing

Trevor Owen's presentation was entitled *The Crowd and the Library*, in which he set the scene by examining the term "crowdsourcing" and introduced a framework which was used throughout the workshop to analyse existing crowdsourcing initiatives and develop potential use cases. The presentation is summarised below.

Crowd sourcing is not a new concept to libraries, archives and museums as it connects with their long history of participation and engagement with members of the public. However crowdsourcing is a vague term which can be easily interpreted as exploitative relating to free labour. Most crowdsourcing projects in heritage institutions have not involved large crowds and have had very little to do with outsourcing labour. They are a continuation of the long standing volunteerism and involvement of citizens in the creation and development of public goods.

There are four key components that need to be considered for heritage institutions to be successful in inviting members of the public to participate as amateurs in the production, development and refinement of public goods:

1. Human computations

Human beings are capable of processing information and making judgements in ways that computers cannot. When designing crowdsourcing projects it is important that we do not waste the public's time by asking them to complete tasks that a computer could already complete. The key is to integrate the unique capabilities of the people into systems.

2. The wisdom of crowds or "why wasn't I consulted"

The wisdom of the crowds emerges from discussions and interactive platforms, such as the wiki, enabling individuals to edit and add to each other's work. However, the heart of the interactions seems to come from the human desire to respond. The web is particularly well suited to answer the question "Why wasn't I consulted"? So the key is to create the sense involvement by allowing people to provide their opinion.

3. Tools and software as scaffolding

The right tools are like scaffolding, putting people in position to do their job. When expertise can be embedded in the design of the tools, it will magnify users effort make it simpler and quicker to accomplish a task.

4. Sense of motivation

People support causes and projects that provide them with a sense of purpose. They feel motivated by doing things that matter to them and get a sense of belonging by being part of something bigger than themselves. As stewards of cultural memory, this is where the libraries, archives and museums have the most to offer.

The full paper by Trevor Owens, entitled *The Crowd & the Library – the Agony and the Ecstasy of "Crowdsourcing our Cultural Heritage"*, is enclosed as Appendix 3.

4. What works?

The first breakout session of the workshop focused on the analysis of a number of crowdsourcing projects, looking at their goals and audiences, as well as how the projects work and how they invite and encourage participation. The workshop participants were also asked to examine in detail the four components described above, applying to each of the 4 projects listed below which had been discussed.

1. Citizen Archivist Dashboard:
http://www.archives.gov/citizen-archivist/
Where citizen archivists can tag, transcribe, edit articles, upload scans, and participate in contests all related to the records of the US National Archives.

2. Old Weather
http://www.oldweather.org/
Old Weather invites you to help reconstruct the climate by transcribing old weather records from ships logs.

3. Galaxy Zoo
http://www.galaxyzoo.org/
Interactive project that allows the user to participate in a large-scale project of research: classifying millions of images of galaxies found in the Sloan Digital Sky.

4. What's on the menu
http://menus.nypl.org/
Help the New York Public Library improve a unique collection. We're transcribing our historical restaurant menus, dish by dish, so that they can be searched by what people were eating back in the day. It's a big job so we need your help!

Each of the projects has done things well but also has areas where improvements can be made. The New York Public Library's menu transcribing project for example did a good job making the website appealing to historians, foodies, and chefs. The project's homepage offers a fairly straightforward explanation of what the users are expected to do and it is initially fun to transcribe the menu items. However, the website does not always give users a clear sense of progress and it is not obvious which menu items have already been transcribed and which ones still need to be done. Another example is the Citizen Archivist Dashboard project by the US National Archives, which was considered a good starting point for engaging users of the Archives and offering a window into the Archives' collections. However the project seems to lack the focus of a crowdsourcing project in that it tries to do too many things and does not seem to have clearly defined audiences. Those who looked at the registration process on the website also found it lengthy and difficult to go through. The Galaxy Zoo and the Old Weather project both involved more specialist scientific tasks, which focus on the message "help the scientists" to invite participation. Both websites require registration, perhaps because of the nature of the tasks involved, and provide statistics on the number of people participating in the projects, creating a sense of community and belonging.

A numbers of key observations can be made and extracted from the overall discussion:

- Trade-offs quite often emerge between richer functionalities on a crowdsourcing website and forming barriers to participation by users. Requesting users to login for example has the advantage of being able to store information to enable personalised services but being able to start immediately without login is appealing
- It is important to provide feedback to users on how they are doing and how their contribution is furthering the overall progress of the project. This helps to keep users engaged.
- Advanced users and regular users should be given different tasks, fully utilising the wisdom of the crowd.
- Gamification can be used to motivate users but is tricky to manage. Projects need to make sure that they are not undoing the intrinsic motivation of the work through these techniques. This can be particularly problematic when when payment is involved.
- Crowdsourcing should be engaging, especially when users are asked to carry out repetitive tasks. It is easy to attract people to something new but more difficult to keep them interested.
- There may be sensitivity around areas there is already professional expertise within the organisation (eg cataloguing). It is important to design the project in such a way that the crowd and the expert each do what they are best at.

5. Crowdsourcing in web archiving

The web is a highly interactive and participatory platform. Its nature and scale lend itself well to crowdsourcing some of the work in archiving the web. Although the workflows of web archiving organisations differ, there are a number of key tasks which are common to web archiving and are good candidates for developing crowdsourcing initiatives.

Nomination: the process of suggesting candidate websites for long term preservation.

Selection: the decision-making process which determines what websites to archive and to include as part of a web archive collection.

Quality Assurance: the process of examining the characteristics of the websites captured by web crawling software, which is largely manual in practice, before making a decision as to whether a website has been successfully captured to become a valid archival copy.

Obtaining Permission: the process of contacting IPR owners and obtaining their agreement to archive selected websites.

Cataloguing / Describing websites: the process of adding descriptive metadata to archived websites.

Harvesting (or crawling): the automated process of downloading copies of selected websites, commonly using web crawling software.

A number of crowdsourcing use cases for web archiving have been identified by the workshop participants, which are described below:

1. Nomination / selection of at-risk websites

This potential project connects to the heroic sense of "saving the world" or "saving the web" and asks for people to nominate websites for preservation which are at risk of disappearing from the live web. The project will have a broad audience but should be particularly appealing to websites creators, government organisations and media experts. For the project to be successful it is important to maximise scaffolding. Instead of asking users to go to a website to nominate URLs, browser plugins, mobile applications etc. should be used to make the nomination process part of the users' day to day workflow of using the web. In addition, social networks are well placed to promote the project and collect nominations.

The Twittervane project, funded by the IIPC and carried out by the British Library, was discussed in this context and regarded as a good example of "scaffolding". A hash tag could be used and regularly communicated to users for the purpose of the project. The Twittervane application can then automatically harvest the URLs submitted using the specific hash tag.

2. Quality Assurance by web archive users

This potential project outsources the task of quality checking captured websites to the users of the web archive and will focus on the message “help the archivists” to invite participation. Contribution can be in a number of different ways, ranging from simple rating of quality, acceptance or rejection of archived websites as archival copy to more complex checking and reporting of missing content and broken links.

3. Quality Assurance and curation by website owners

This potential project is similar to the project above in that it outsources the quality assurance tasks. However it targets specifically the website owners who have a much bigger stake in good quality capture of their own sites. The project connects to websites owners by sharing the stewardship of content they care about. This could be combined with curation related tasks such as asking website owners to describe the websites or comment on descriptions of websites by curators.

4. Describing / Tagging archived websites

Due to the scale of web archives, especially those based on broad domain crawls, there is general lack description of web archive content. This potential project connects to amateur curators to help describe or tag web archive collections. This could be done in a number of different ways. We could ask people to describe / tag embedded object such as images, or to tag names of people, places and organisations. Users could also build their own collections, describe these and share them with others. The descriptions and tagged content can be used to improve search and access.

6. Recommendations

The workshop confirmed the potential of crowdsourcing in web archiving and provided a forum to understand and discuss the various aspects of its application in detail. While encouraging IIPC member institutions to take advantage of this form of collaborative working, the workshop participants made a number of recommendations:

1. Crowdsourcing is a great way to collect ideas from your crowd and interact with them to keep the conversation going.

2. There is no free lunch and costs of crowdsourcing projects should not be overlooked. Crowdsourcing requires community management.

3. Crowdsourcing projects should have a clear sense of purpose and scope, focusing on distinctive tasks which connect to the target audiences.

4. Users need to be supported by tools which act as scaffolds to make the most of their effort.

5. Make crowdsourcing engaging and interesting for the participants.

6. Quality of crowdsourced work should be taken into account. Some sort of vetting or assurance may be required of the work done by the crowd.

7. Crowdsourcing could be disruptive to workflows and expertise which are already in place. Adjustment to existing workflow / resources should be part of the considerations when developing crowdsourcing initiatives.

Helen Hockx-Yu
British Library
May 2012

Maureen Pennock, British Library 18 April 2012

DISCUSSION PAPER
Can Crowdsourcing play a role in archiving the web?

Cultural heritage organisations are increasingly inviting users to contribute to the growth and curation of their collections through so-called 'crowdsourcing' initiatives. Most of this work has focused on digitised resources. Very little has taken place to explore whether crowdsourcing can be similarly employed for born digital content, particularly in the web archiving domain. This short paper is an introduction to the role that crowdsourcing may play in archiving the web and is intended to stimulate thinking prior to attendance at the IIPC workshop on May 4th.

Crowdsourcing offers web archives an opportunity to increase the amount of people power available to a project or initiative, at little or no extra cost. But what exactly is crowdsourcing, and what are the real benefits of getting an often unskilled workforce to perform specialised tasks?

Wikipedia, a crowdsourced encyclopaedia, describes crowdsourcing as *'the act of outsourcing tasks traditionally performed by an employee or contractor to an undefined, large group of people or community (a 'crowd') through an open call'*.[1]

The tasks undertaken can vary wildly. A recent paper by Oomen & Aroyo explored opportunities and challenges for crowd sourcing in the cultural heritage domain and suggested the following classification of crowd sourcing tasks:

- **'Correction & transcription** - 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.
- **Complimenting 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.'[2]

Before we explore how this may relate to web archives, let us first consider some examples of successful crowdsourced projects in the cultural heritage sector.

[1] http://en.wikipedia.org/wiki/Crowdsourcing
[2] http://www.cs.vu.nl/~marieke/OomenAroyoCT2011.pdf

1. The British Library's Georeferencing project

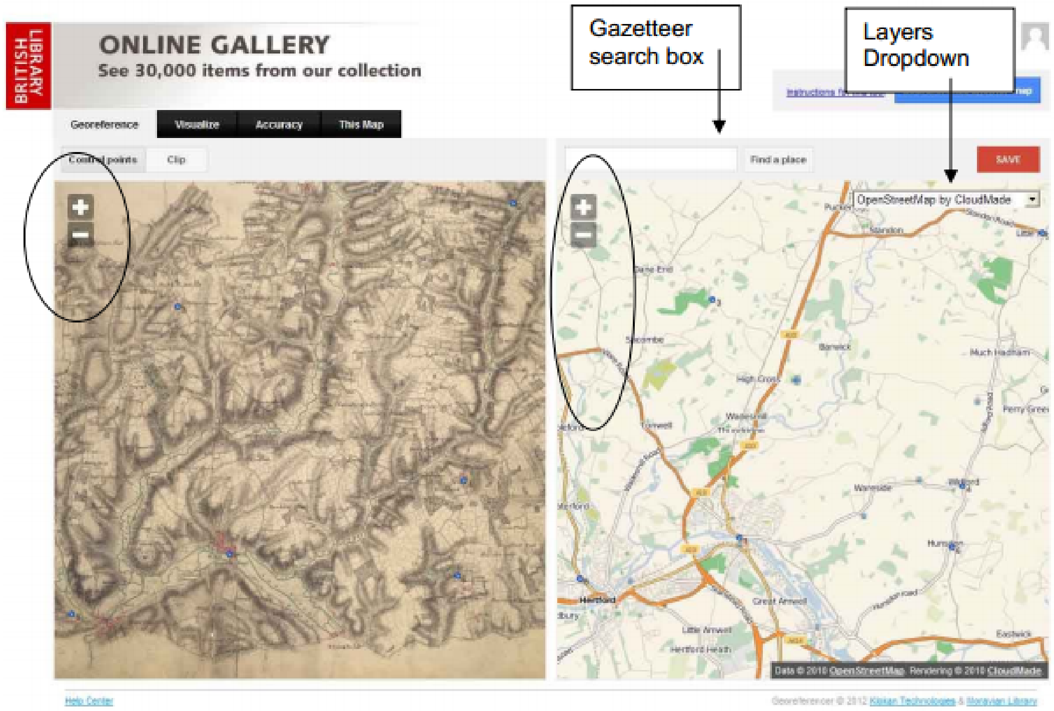


Figure Screenshot of the British Library's Georeferencer, with zoom-pan controls of the two maps circled, and labelled arrows pointing to the Gazetteer search box and the Layers dropdown. OSD 141 of Ware is pictured.

From the website:

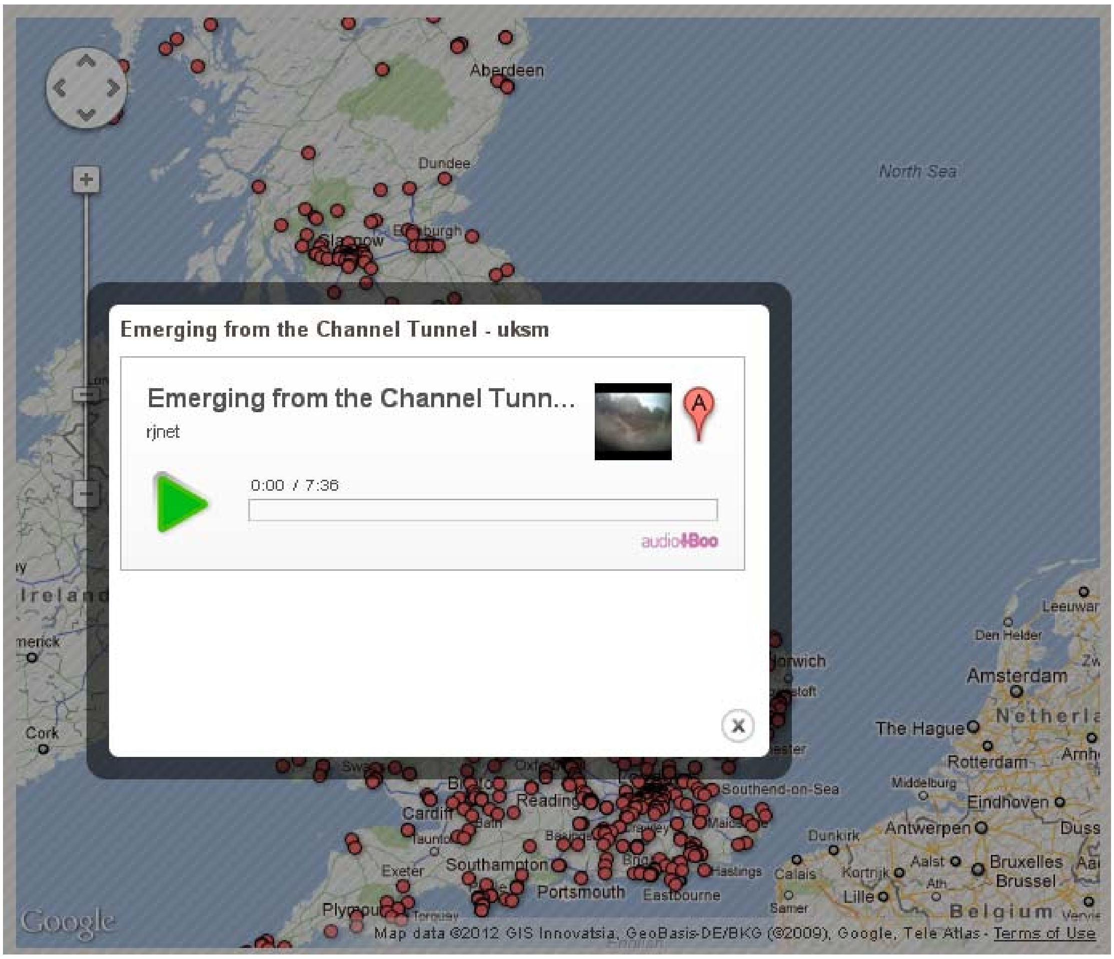
'The British Library's georeferencing project crowdsourced location data on Britain's historic landscape to make a selection of maps fully searchable and viewable using popular online geotechnologies.

Online geographic tools allow historic maps to be overlaid on modern mapping, enhancing the ability to view and compare the past with the present, and improving findability. Georeferencing, i.e. assigning points on a map image to corresponding geographical co-ordinates, links the map to its spatial location on the ground using universal geographic standards (latitude / longitude).'[3]

725 maps were assigned spatial metadata. The project was planned to span a year yet all maps were georeferenced within a week of the project going live. Formal publicity was minimal and word was spread mostly through social media. The project had around 90 participants in the week it was live, half of the work was completed by just five of those. The data quality was very good, with less than 3% of the total maps requiring correction.

[3] http://www.bl.uk/maps/georefabout.html

2. The British Library's UK SoundMap project

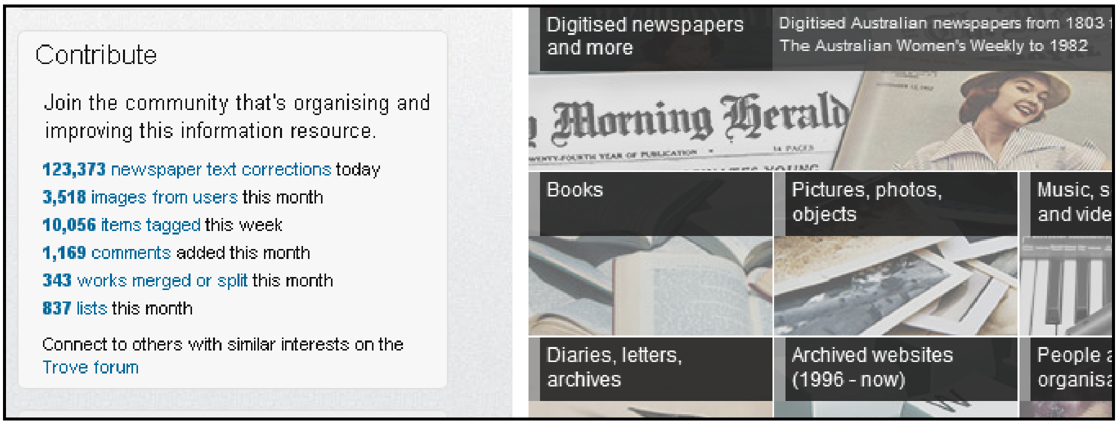


From the website:

'The UK Soundmap, the first nationwide sound map, invited people to record the sounds of their environment, be it at home, work or play. Over 2,000 recordings were uploaded by some 350 contributors during the period July 2010 to July 2011.'

Around 80% of uploads were made from mobile devices. Roughly 7% of uploads were rejected for reasons of quality, copyright, or inappropriate sounds. The project had an active Twitter presence and required users to upload content from the AudioBoo service and to tag contributions with metadata including the #uksm tag and location data if location data was not supplied by the upload device. After moderation, contributions were mapped using Googlemaps and made available from the Sound Archive website. The project won the 2010 SomeComms award for best public sector use of social media.

1. TROVE at the National Library of Australia.



Contribute

Join the community that's organising and improving this information resource.

123,373 newspaper text corrections today
3,518 images from users this month
10,056 items tagged this week
1,169 comments added this month
343 works merged or split this month
837 lists this month

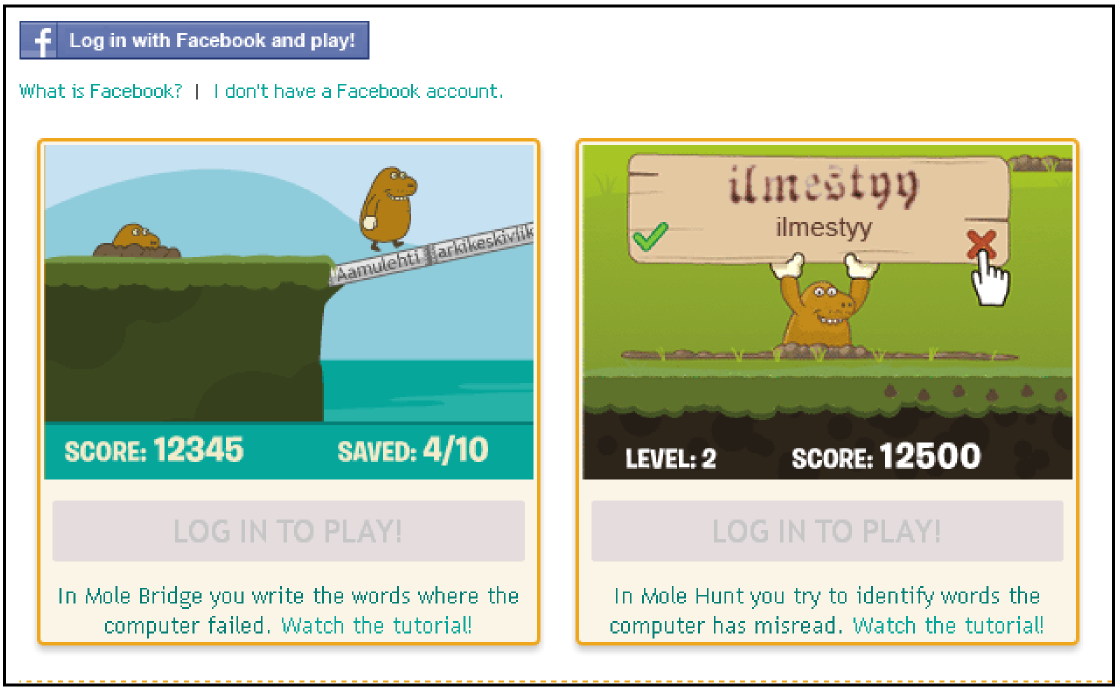
Connect to others with similar interests on the Trove forum

Digitised newspapers and more
Digitised Australian newspapers from 1803 t
The Australian Women's Weekly to 1982
Morning Herald
Books
Pictures, photos, objects
Music, s and vide
Diaries, letters, archives
Archived websites (1996 - now)
People a organisa

Trove invites users to contribute in a variety of ways with different sorts of content. Items can be tagged with keywords, comments added, content uploaded to Flickr and digitised text corrected. You can tag anything in Trove except for archived websites.

The text digitisation correction programme runs on the NLA collection of digitised newspapers. Around 40,000 volunteers have corrected over 5o million lines of OCR text in historic newspapers. The programme was nominated for the 2011 Digital Heritage Award (organised by the Dutch Institute for Cultural Heritage). It is a very successful crowdsourcing venture though it was not devised as such: the goal was to digitise and make the content available online; error correction was ‘a bit on the side’. Read Rose Holley’s blog for more details.

4. DigitalKoot from the National Library of Finland



In Mole Bridge you write the words where the computer failed. Watch the tutorial!

In Mole Hunt you try to identify words the computer has misread. Watch the tutorial!

DigitalKoot is a text correction game for OCR digitised newspapers. From the website:

‘The current project consists of the National Library's archive of the issues of the newspaper Aamulehti from the end of the 19th century.

So far 103,212 people have visited the Digitalkoot web site. Volunteers have contributed a total of 352,486 minutes (6,791,628 microtasks) of their time.’

The gamification element is widely considered to be one of the reasons why the project is so successful.

5. See also…

- WAISDA crowdsourcing description project from the Netherlands’ institute for Sound & Vision – described here
- Library of Congress on Flickr – social description & tagging
- Galaxy Zoo – not a GLAM project but a fascinating example of an early simple and hugely successful classification project.

What are the Benefits of crowd sourcing?

The benefits of crowdsourcing range from increased awareness and community engagement, to low cost labour, knowledge exchange, broad input, and faster progress. A D-lib article by Rose Holley in 2010 identified the following benefits of crowdsourcing for Libraries:

- Achieving goals the library would never have the time, financial or staff resource to achieve on its own.
- Achieving goals in a much faster timeframe than the library may be able to achieve if it worked on its own.
- Building new virtual communities and user groups.
- Actively involving and engaging the community with the library and its other users and collections.
- Utilising the knowledge, expertise and interest of the community.
- Improving the quality of data/resource (e.g. by text, or catalogue corrections), resulting in more accurate searching
- Adding value to data (e.g. by addition of comments, tags, ratings, reviews).
- Making data discoverable in different ways for a more diverse audience (e.g. by tagging).
- Gaining first-hand insight on user desires and the answers to difficult questions by asking and then listening to the crowd.
- Demonstrating the value and relevance of the library in the community by the high level of public involvement
- Strengthening and building trust and loyalty of the users to the library. Users do not feel taken advantage of because libraries are non-profit making.[4]

Many of these are just as applicable to web archiving initiatives as they are to libraries generally. But there are, of course, also risks involved. The quality of work provided can vary wildly, the initiative may fail to attract the desired attention and input, the amount of effort expended on organising the initiative may outweigh the returned effort from the crowd, and it may even result in negative exposure if received badly. Several sites have discussed possible pitfalls - interestingly, more than you might expect given that most of the crowdsourcing projects we hear about are successful:

- Added costs to bring a project to an acceptable conclusion
- Increased likelihood that a crowdsourced project will fail due to lack of monetary motivation, too few participants, lower quality of work, lack of personal interest in the project, global language barriers, or difficulty managing a large-scale, crowdsourced project
- Projects can fail to attract participants and sometimes require the agency involved to create and seed their own films
- No written contracts, non-disclosure agreements, or employee agreements

[4] http://www.dlib.org/dlib/march10/holley/03holley.html

- Difficulties maintaining a working relationship with crowdsourced workers
- Susceptibility to faulty results caused by targeted, malicious work efforts
- Criticism for using free labour

So on one hand it may be a case of ‘many hands make light work; on the other, that ‘too many cooks spoil the broth’. There is therefore an element of risk management in any crowdsourcing initiative.

How does all this relate to web archiving?

The web is a social platform, built by people and organizations for people and organizations. The sheer scale of the web archiving challenge suggests that web archives could potentially benefit greatly from crowdsourcing some of their work. But how? Let us first consider the main stages of a generic selective web archiving workflow:

1. Identify sites
2. Enter details into web archiving system
3. Obtain permission
4. Crawl
5. QA Crawl results
6. Catalogue
7. Provide access

Appreciating that the order of these stages may vary, crowdsourcing contributions could be made at least at stages 1, 2, 5 and 6, whilst still enabling the web archiving institution to retain control over the infrastructure.

In other scenarios and outside of a business as usual setting, the entire process could potentially be crowdsourced.

But setting a non-standard scenario aside, how might a web archive maximise the potential of the crowd to enhance their capacity for web archiving?

1. Identify sites.

Selective archives that accept nominations from the public have an element of crowdsourcing, particularly for specific collections that focus on a given event or theme. Nominations may be sought in a variety of means, such as social media, an applet, an online form or a broader engagement programme such as K12.

The Twittervane application developed by the British Library is similar to crowdsourcing as it 'outsources' parts of the selection/nomination process. However, as it does not ask the crowd to do anything different from what they already do, it is not strictly speaking a crowdsourcing application.

2. Enter details into web archiving system

Whilst unlikely that institutions would completely open up access to their system, the nominations stage could request sufficient details that the system can then be populated automatically.

5. QA crawl

Quality assurance has already been successfully crowdsourced by a number of institutions for digitisation projects. The QA process followed by most web archives is time consuming and potentially complicated, depending on the size of the site, the type of content hosted, and the technical structure. however, it is conceivable that crowdsourcing could supported targeted elements of the QA process. The comparative aspect of QA (does the archived site look/behave the same as the live one?) lends itself well to 'quick wins' for participants.

6. Cataloguing

Cataloguing and describing sites is a valuable activity that is often best carried out by people already familiar with the subect. Many people are already familiar with the concept of annotation or tagging content online. Crowdsourcing this process could work on a number of counts.

Conclusion

Crowd sourcing could support the growth and curation of web archives in several different ways. There are thus many different crowd sourcing scenarios that could be drafted. Given that there are as yet *no* formal crowd sourcing initiatives underway for web archives, it may be most appropriate to start small and grow through experience, rather than devise a large scale project based on relatively little practical crowdsourcing experience.

A key challenge for web archives is to not only to devise an appropriate project but also to attract a sizable audience to participate in the challenge. The number and enthusiasm of participants is key to all projects. Web archives often struggle to achieve a high profile in the cultural heritage and scholarly communities, suggesting that a strong engagement and promotional programme would be required. It may be that a gamification & reward element would help achieve this.

Another challenge is the longevity of scale of the challenge. Successful projects are typically broken down into achievable projects, against which progress can be measured and the goal 'smashed'. This is a potential issue for web archives that collect on an ongoing basis and suggests that crowdsourcing solutions that target specific 'special collections' would have a greater chance of success than open-ended challenges.

Further reading

1. Oomen & Aroyo (2011) - Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges
2. Holley (2010) - Crowdsourcing: How and Why Should Libraries Do It?

Crowd Sourcing Web Archives – IIPC Workshop -- 2012

Workshop Schedule

9:00-9:15: Brief Introductions (Go around the room and make sure everyone knows everyone and have everyone explain what their interests and experiences are with crowdsourcing, ask them what they would like to get out of the workshop.)

9.15 - 10.00 The Crowd and The Library: Trevor Owens (Presentation)

10.00 - 10:20 Time for questions and discussion of the presentation.

10.20 – 10:40 Break

10:40 – 11:30 Components of Crowdsourcing initiatives (group activity) We will break into groups to explore individual crowdsourcing projects. Each group will consider goals of the project, its structure, its intended audience, and its communications plan.

11:30-12:00 Group report outs and discussion of components (report outs) Each group will give a five minute report out to the rest of the workshop. The goal in this case is to start synthesizing and explicating key features of crowdsourcing projects.

12.00 - 13.00 Lunch

13.00 - 13.20 Key features of crowdsourcing projects (Full Group Discussion) An open discussion focused on synthesizing and abstracting key design features and principles from the morning presentation with the morning break out session.

13.20 - 14.20 Developing Crowdsourcing Use Cases for Web Archives (Breakout Groups) Each group will focus on developing a proposal for a specific use case for using crowdsourcing for web archiving projects.

14.20 - 14.30 Break (though groups can keep working if they wish)

14.30 - 15.00 Each group reports back to the group as a whole, walking through their use cases and responding to questions from the other groups.

15.15 - 15.30 Summary and closing.

Crowdsourcing Sites for Consideration

Citizen Archivist Dashboard:
http://www.archives.gov/citizen-archivist/
Where citizen archivists can tag, transcribe, edit articles, upload scans, and participating in contests all related to the records of the US National Archives.

Trove
http://trove.nla.gov.au/
User’s correct ocr’ed newspaper, upload images, tagged items, post comments and add lists.

GLAM Wiki
http://outreach.wikimedia.org/wiki/GLAM/Model_projects
The GLAM-WIKI project supports GLAMs and other institutions who want to work with Wikimedia to produce open-access, freely-reusable content for the public.

Old Weather
http://www.oldweather.org/
Old Weather invites you to help reconstruct the climate by transcribing old weather records from ships logs.

Galaxy Zoo
http://www.galaxyzoo.org/
Interactive project that allows the user to participate in a large-scale project of research: classifying millions of images of galaxies found in the Sloan Digital Sky.

UK Sound Map
http://sounds.bl.uk/Sound-Maps/UK-Soundmap
http://britishlibrary.typepad.co.uk/archival_sounds/uk-soundmap/
The UK Soundmap, the first nationwide sound map, invited people to record the sounds of their environment, be it at home, work or play. Over 2,000 recordings were uploaded by some 350 contributors during the period July 2010 to July 2011.

What's on the menu
http://menus.nypl.org/
Help The New York Public Library improve a unique collectionWe're transcribing our historical restaurant menus, dish by dish, so that they can be searched by what people were eating back in the day. It's a big job so we need your help!

STEVE
http://tagger.steve.museum/
A place where you can help museums describe their collections by applying keywords, or tags, to objects.

Crowd Sourcing Web Archives – IIPC Workshop -- 2012

Questions to Ask of Crowdsourcing Sites:

Goals: What is the apparent goal of the project?

How does it work? Briefly describe how it works, what users do and what the system does.

Audience: Who is the audience it is targeted at? How does it try to invite that audience? Do you think it is effective?

Invitation: How does it invite participation? Does it try to keep users engaged and interested? Does it give us a sense of progress?

Human Computation: How could we use human judgment to augment computer processable information?

Wisdom of Crowds: How could we empower and consult with a community of users?

Scaffolding: How can our tools act as scaffolds to help make the most of users efforts?

Motivation: Whose sense of purpose does this project connect to? What identities are involved?

Crowd Sourcing Web Archives – IIPC Workshop -- 2012

Crowdsorciong Use Case Templates

Project Name:

Goals: What is the goal of the project? What are you starting with what do you want to end with and why is the endstate valuable?

How does it work? Work up a brief gloss of a workflow, what users do and what the system does.

Audience: Who is the audience it is targeted at? How would you get them to know about it?

Invitation: How would the system invite participation? How would it try to keep users engaged and interested?

Human Computation: How would we use human judgment to augment computer processable information?

Wisdom of Crowds: How could we empower and consult with a community of users?

Scaffolding: How can our tools act as scaffolds to help make the most of users efforts?

Motivation: Whose sense of purpose does this project connect to? What identities are involved?

References, Further Reading & Watching

Ahn, L. von. (2006). *Human Computation*. Google TechTalks. Retrieved from http://video.google.com/videoplay?docid=-8246463980976635143

Brumfield, B. W. (2012, March 17). Collaborative Manuscript Transcription: Crowdsourcing at IMLS WebWise 2012. *Collaborative Manuscript Transcription*. Retrieved April 25, 2012, from http://manuscripttranscription.blogspot.com/2012/03/crowdsourcing-at-imls-webwise-2012.html

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, USA.

Crowdsourcing Cultural Heritage: The Objectives Are Upside Down | Trevor Owens. (n.d.). Retrieved April 25, 2012, from http://www.trevorowens.org/2012/03/crowdsourcing-cultural-heritage-the-objectives-are-upside-down/

deterding, sebastian. (2011, February 19). *Meaningful Play: Getting Gamification Right*. Retrieved from http://www.youtube.com/watch?v=7ZGCPap7GkY&feature=youtube_gdata_player

Ford, P. (2011, January 6). The Web Is a Customer Service Medium (Ftrain.com). Retrieved May 3, 2012, from http://www.ftrain.com/wwic.html

Gee, J. P. (2000). Identity as an analytic lens for research in education. *Review of research in education*, *25*(1), 99.

Gee, James Paul. (2003). *What Video Games Have to Teach Us About Learning and Literacy* (New Ed.). Palgrave Macmillan. Retrieved from http://www.amazon.com/dp/1403965382

Holley, R. (2010). Crowdsourcing: How and Why Should Libraries Do It? *D-Lib Magazine*, *16*(3/4). doi:10.1045/march2010-holley

Hutchins, E. (1995). How a Cockpit Remembers Its Speed. *Cognitive Science*, *19*, 288, 265.

Itō, M., Baumer, S., Bittanti, M., boyd, danah, Cody, R., Herr-Stephenson, B., Horst, H. A., et al. (2010). *Hanging Out, Messing Around, and Geeking Out: Kids Living and Learning with New Media*. Cambridge, Mass: MIT Press.

Juul, J. (2011, April 2). Gamification Backlash Roundup. *The Ludologist*. Retrieved April 25, 2012, from http://www.jesperjuul.net/ludologist/gamification-backlash-roundup

Karen Smith-Yoshimura. (2012). *Social Metadata for Libraries, Archives, and Museums: Executive Summary*. Dublin, Ohio:: OCLC Research. Retrieved from http://www.oclc.org/research/publications/library/2012/2012-02.pdf

Oomen, J., & Aroyo, L. (2011). Crowdsourcing in the cultural heritage domain: Opportunities and challenges. *Proceedings of the 5th International Conference on Communities and Technologies* (pp. 138–149).

Pea, R. (1997). Practices of distributed intelligence and designs for education. In G. Salomon (Ed.), *Distributed cognitions: psychological and educational considerations,* Learning in doing. Cambridge, UK: Cambridge University Press.

Pink, D. H. (2009). *Drive: The Surprising Truth About What Motivates Us* (1st ed.). Riverhead Hardcover.

Sternberg, R. J. (2005). Intelligence, Competence, and Expertise. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of Competence and Motivation* (pp. 15-30). New York: NY: Guilford Press.

The Crowd & the Library

The Agony and the Ecstasy of “Crowdsourcing” our Cultural Heritage

Trevor Owens, Digital Archivist, The Library of Congress

Libraries, archives and museums have a long history of participation and engagement with members of the public. This essay connects these traditions with current discussions of crowdsourcing. Crowdsourcing is a bit of a vague term, one that comes with potentially exploitative ideas related to uncompensated or undercompensated labor. This essay focuses on how a set of related concepts, human computation, the wisdom of crowds, thinking of tools and software as scaffolding, and understanding and respecting end users motivation can both help clarify what crowdsourcing can do for cultural heritage organizations while also clarifying a clearly ethical approach to inviting the public to help in the collection, description, presentation, and use of the cultural record.

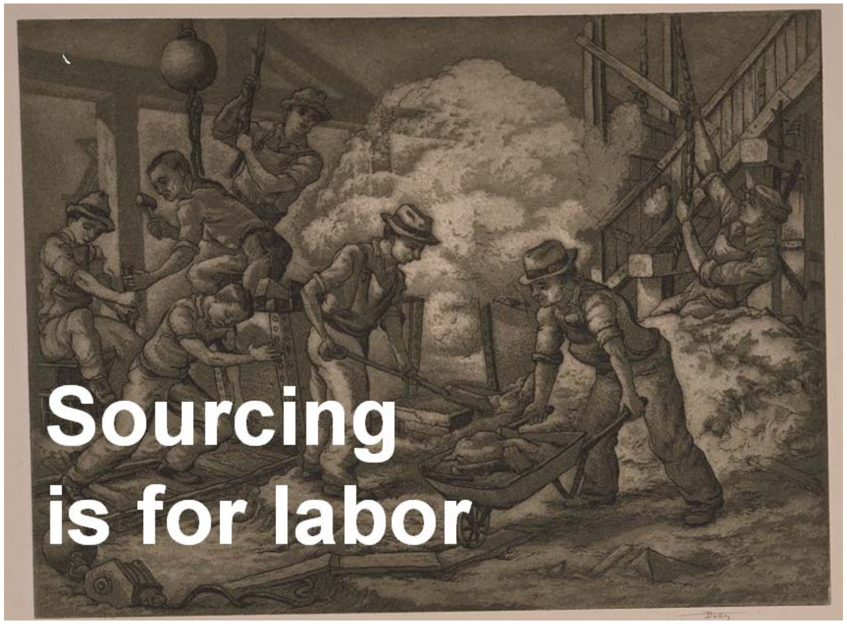
The Two Problems with Crowdsourcing: Crowd and Sourcing

There are two primary problems with bringing the idea of crowdsourcing into cultural heritage organizations. Both the idea of the crowd and the notion of sourcing are problematic terms. The most successful projects crowdsourcing projects in libraries, archives and museums have not involved large and massive crowds and they have very little to do with outsourcing labor.

Most successful crowdsourcing projects are not about large anonymous masses of people. They are not about crowds. They are about inviting participation from interested and engaged members of the public. These projects can continue a long standing tradition of volunteerism and involvement of citizens in the creation and continued development of public goods.

For example, the New York Public Library’s menu transcription project, *What’s on the Menu?*, invites members of the public to help transcribe the names and costs of menu items from digitized copies of menus from New York restaurants. Anyone who wants to can visit the project website and start transcribing the menus. However, in practice it is a dedicated community of foodies, New York history buffs, chefs, and otherwise self-motivated individuals who are excited about offering their time and energy to help contribute, as volunteers, to improving the public library’s resource for others to use[1]. Far from a break with the past, this is actually a clear continuation of a longstanding tradition of inviting members of the public in to help refine, enhance, and support resources like this collection as public goods. In the case of the menus, years ago, it was actually volunteers who sat at a desk in the reading room who had cataloged the original collection[2]. In short, crowdsourcing the transcription of the menus project is not about crowds at all, it is about using digital tools to invite members of the public to volunteer in much the same way members of the public have volunteered to help organize and add value to the collection in the past.

The problem with the term sourcing is its association with labor. Wikipedia’s definition of crowdsourcing helps further clarify this relationship, “Crowdsourcing is a process that involves outsourcing tasks to a distributed group of people.” The keyword in that definition is outsourcing. Crowdsourcing is a concept that was invented and defined in the business world and it is important that we recast it and think through what changes when we bring it into cultural heritage.



At this point, we need to think for a moment about what we mean by terms like work and labor. While it might be ok for commercial entities

[1] Ben Vershbow, Bringing in the Crowd: Effects, Affects and a Few (minor) Defects, Presented at Make It Work: Improvisations on the Stewardship of Digital Information, July 19-21, 2011 Washington, DC Slides online at http://www.digitalpreservation.gov/meetings/documents/ndiipp11/vershbow.pdf

[2] Barbara Taranto, Crowdsourcing Metadata, Presented at the Coalition for Networked Information Fall 2011 Membership Meeting, December 12-13-2011, Video available online at http://vimeo.com/38196574

to coax or trick individuals to provide free labor the ethical implications of such trickery should give pause to cultural heritage organizations. It is critical to pause here and unpack some of the different meanings we ascribe to the terms work. When we use the term “a day’s work” we are directly referring to labor, to the kinds of work that one engages in as a financial transaction for pay. In contrast, when we use the term work to refer to someone’s “life’s work” we are referring to something that is significantly different. The former is about acquiring the resources one needs to survive. The latter is about the activities that we engage in that give our lives meaning. In cultural heritage we have clear values and missions and we are in an opportune position to invite the public to participate. However, when we do so we should not treat them as a crowd, and we should not attempt to source labor from them. When we invite the public we should do so under a different set of terms.

Citizen Scientists, Archivists and the Meaning of Amateur

Some of the projects that fit under the heading of crowdsourcing have chosen very different kinds of terms to describe themselves. For example, the Galaxy Zoo project, which invited anyone interested in Astronomy to help catalog a million images of stellar objects, refers to its users as citizen scientists[3]. Similarly, the United States National Archives and Records Administration recently launched crowdsourcing project, the Citizen Archivists Dashboard, invites citizens, not members of some anonymous crowd, to participate[4]. The names of these projects highlight the extent to which they invite participation from members of the public who identify with and the characteristics and ways of thinking of particular professional occupations. While these citizen archivists and scientists are not professional, in the sense that they are unpaid, they connect with something a bit different than volunteerism. They are amateurs in the best possible sense of the term.

Amateurs have a long and vibrant history as contributors to the public good. Coming to English from French, the term Amateur, means a “lover of.” The primarily negative connotations we place on the term are a relatively recent development. In other eras, the term Amateur simply meant that someone was not a professional, that is, they were not paid for these particular labors of love. Charles Darwin, Gregor Mendal, and many others who made significant contributions to the sciences did so as Amateurs. As a continuation of this line of thinking, the various Galaxy Zoo projects see the amateurs who participate as peers, in many cases listing them as co-authors of academic papers published as a result of their work. I suggest that we think of crowdsourcing not as extracting labor from a crowd, but of a way for us to invite the participation of amateurs (in the non-derogatory sense of the word) in the creation, development and further refinement of public goods.

Toward a better, more nuanced, notion of Crowdsourcing

With all this said, fighting against a word is rarely a successful project, from here out I will continue to use and refine a definition for crowdsourcing that I think works for the cultural heritage sector. In the remainder of this essay I will explain what I think of as the four key components of this ethical crowdsourcing, this crowdsourcing that invites members of the public to participate as amateurs in the production, development and refinement of public goods. For me these fall into the following four considerations, each of which suggests a series of questions to ask of any cultural heritage crowdsourcing project. The four concepts are;

1. Human Computation
2. The Wisdom of Crowds
3. Thinking of Tools and Software as Scaffolding
4. Understanding A Deep Sense of Motivation

Together, I believe these four concepts provide us with the descriptive language to understand what it is about the web that makes crowdsourcing such a powerful tool. Not only for improving and enhancing data related to cultural heritage collections, but also as a way for deep engagement with the public.

Human Computation

Human Computation is grounded in the fact that human beings are able to process particular kinds of information and make judgments in ways that computers can’t. To this end, there are a range of projects that are described as crowdsourcing that are anchored in the idea of treating people as processors. The best way to explain the concept is through a few examples of the role human computation plays in crowdsourcing.

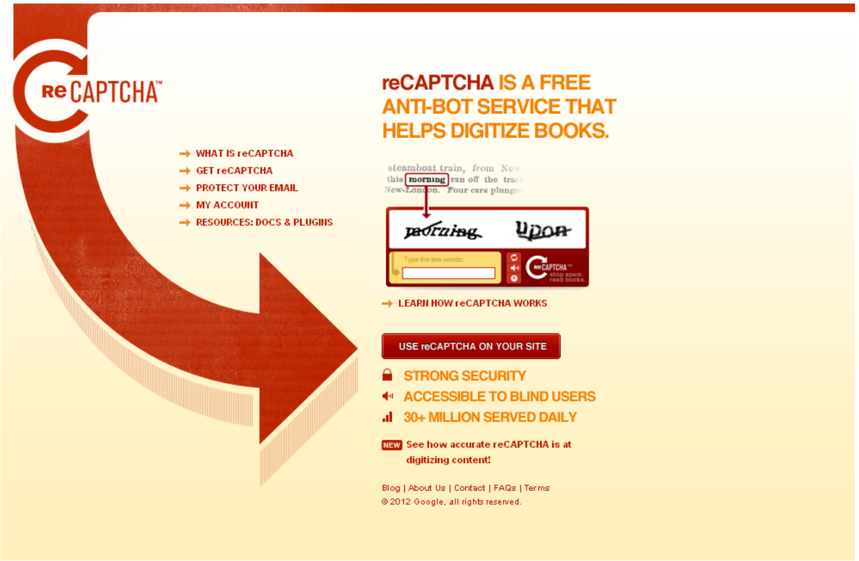
ReCaptcha is a great example of how the processing power of humans can be harnessed to improve cultural heritage collection data[5]. Most readers will be familiar with the little ReCaptcha boxes we fill out when we need to prove that we are in fact a person and not an automated system attempting to login to some site. Our ability to read

[3] Galaxy Zoo can be found online at http://www.galaxyzoo.org/

[4] Citizen Archivist Dashboard can be found online at http://www.archives.gov/citizen-archivist/

[5] Luis von Ahn, Ben Maurer, Colin McMillen, David Abraham and Manuel Blum (2008). "reCAPTCHA: Human-Based Character Recognition via Web Security Measures" *Science* 321 (5895): 1465–1468. doi:10.1126/science.1160379

the strange and messed up text in those little boxes proves that we are people, but in the case of ReCaptcha it also helps us correct the OCR'ed text of digitized New York Times and Google Books transcripts. The same capability that allows people to be differentiated from machines is what allows us to help improve the full text search of the digitized New York Times and Google Books collections.

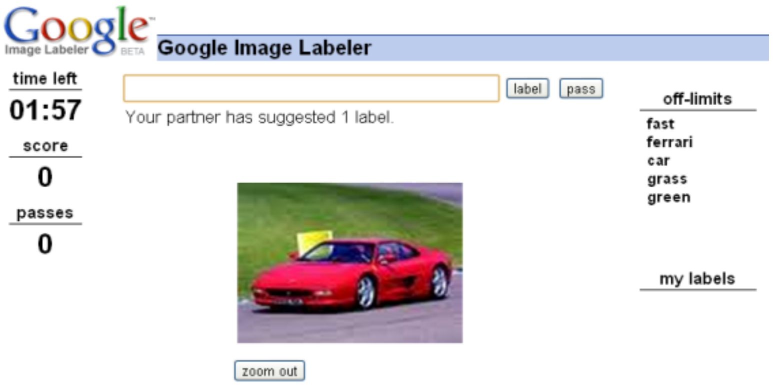

reCAPTCHA IS A FREE ANTI-BOT SERVICE THAT HELPS DIGITIZE BOOKS.

steamboat train, from New this morning ran off the track New-London. Four cars plunge

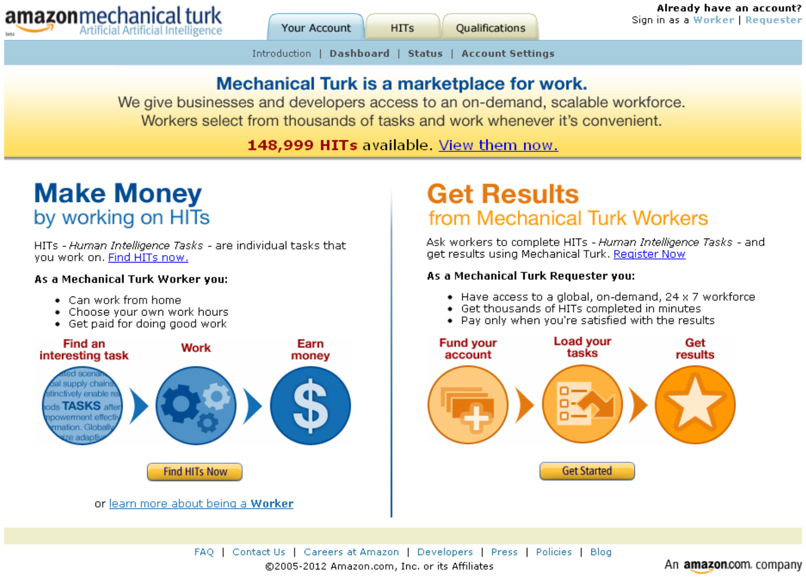
See how accurate reCAPTCHA is at digitizing content!

© 2012 Google, all rights reserved.
reCAPTCHA™
WHAT IS reCAPTCHA
GET reCAPTCHA
PROTECT YOUR EMAIL
MY ACCOUNT
RESOURCES: DOCS & PLUGINS
morning
upon
Type the two words:
reCAPTCHA™ stop spam. read books.
LEARN HOW reCAPTCHA WORKS
USE reCAPTCHA ON YOUR SITE
STRONG SECURITY
ACCESSIBLE TO BLIND USERS
30+ MILLION SERVED DAILY
NEW
Blog | About Us | Contact | FAQs | Terms

The principles of human computation are similarly on display in the Google Image Labeler. From 2006-2011 the Google image labeler game invited members of the public to describe and classify images.[6] For example, in the image below a player is viewing an image of a red car. Somewhere else in the world another player is also viewing that image. Each player is invited to key in labels for the image, with a series of "off-limits" words which have already been associated with the image. Each label I can enter which matches a label entered by the other player results in points in the game. The game has inspired an open source version specifically designed for use at cultural heritage organizations[7]. The design of this interaction is such that, in most cases, it results in generating high quality description of images.


Your partner has suggested 1 label.

Both the image labeler and ReCaptcha are fundamentally about tapping into the capabilities of people to process information. Where I had earlier suggested that the kind of crowdsourcing I want us to be thinking about is not about labor, these kinds of human computation projects are often fundamentally about labor. This is most clearly visible in Amazon's Mechanical Turk project.


Already have an account?
Sign in as a Worker | Requester

Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

148,999 HITs available. View them now.

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

or learn more about being a Worker

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Register Now

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

©2005-2012 Amazon.com, Inc. or its Affiliates
amazonmechanical turk beta
Artificial Artificial Intelligence
Your Account
HITs
Qualifications
Introduction | Dashboard | Status | Account Settings
Find an interesting task
Work
Earn money
TASKS
Find HITs Now
Fund your account
Load your tasks
Get results
Get Started
FAQ | Contact Us | Careers at Amazon | Developers | Press | Policies | Blog
An amazon.com. company

The tagline for Mechanical Turk is that it "gives businesses and developers access to an on-demand, scalable workforce" where "workers select from thousands of tasks and work whenever it's convenient." The labor focus of this site should give pause to those in the cultural heritage sector, particularly those working for public

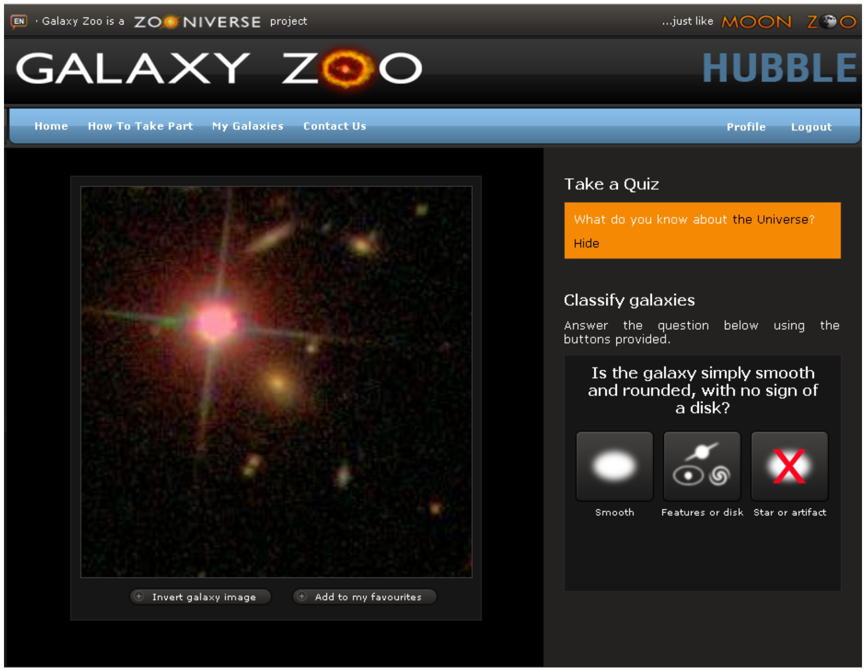
[6] For some further context on both ReCaptcha and Google Image Labler see Luis von Ahn. *Human Computation*. Google TechTalks, 2006. http://video.google.com/videoplay?docid=-8246463980976635143

[7] See Metadata Games, http://www.tiltfactor.org/metadata-games

institutions. There are very legitimate concerns about this kind of labor as serving as a kind of “digital sweatshop.[8]

While there are legitimate concerns about the potentially exploitive properties of projects like Mechanical Turk, it is important to realize that many of the same human computation activities which one could run through Mechanical Turk are not really the same kind of labor when they are situated as projects of citizen science.

For example, the Galaxy Zoo invites individuals to identify galaxies. The activity is fundamentally similar to the Google image labeler game. Users are presented with an image of a galaxy and invited to classify it based on a simple set of taxonomic information. While the interaction is more or less the same the change in context is essential.



Take a Quiz

What do you know about the Universe?

Classify galaxies

Answer the question below using the buttons provided.

Is the galaxy simply smooth and rounded, with no sign of a disk?

Galaxy Zoo invites amateur astronomers to help classify images of galaxies. While the image identification task here is more or less the same as the image identification tasks previously discussed, at least in the early stages of the project, this site often gave these amateur astronomers the first opportunity to ever see these stellar objects[9]. As these images were all captured by a robotic telescope the first galaxy zoo participants who looked at these images were actually the first people ever to see each of these stellar objects. In this case, the amateurs who catalog these galaxies do so because they want to contribute to science[10]. Beyond engaging in this classification activity, the Galaxy Zoo project also invites members to discuss the galaxies in a discussion forum. This discussion forum ends up representing a very different kind of crowdsourcing, one based not so much on the idea of human computation but instead on a notion which I refer to here as the wisdom of crowds.

Key questions from human computation: How could we use human judgment to augment computer processable information? It would be a waste of the public’s time to invite them in to complete a task that a computer could already complete. The value human computation offers is the question of how the unique capabilities of people can be integrated into systems for the creation of public goods.

The Wisdom of Crowds, or Why Wasn’t I Consulted

The Wisdom of Crowds comes from James Surowiecki’s 2004 grandiosely titled book, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* In the book, Surowiecki talks about a range of examples of how crowds of people can create important and valuable kinds of knowledge. Unlike human computation, the wisdom of crowds is not about highly structured activities. In Surowiecki’s argument, the wisdom of crowds is an emergent phenomena resulting from how discussion and interaction platforms, like wikis, enable individuals to add and edit each other’s work.

The wisdom of crowds notion tends to come with a bit too much utopian baggage for my tastes, in contrast, I find Paul Ford’s reformulation of this notion particularly compelling. Ford suggests that the heart of this matter is that the web, unlike other mediums, is

[8] For some brief coverage of these discussions see Williams, George. “The Reliability, Efficiency, and Affordability of Amazon’s Mechanical Turk.” The Chronicle of Higher Education. *ProfHacker*, February 22, 2010. http://chronicle.com/blogs/profhacker/the-reliability-efficiencyaffordability-of-amazons-mechanical-turk/22994 and Williams, George. “The Ethics of Amazon’s Mechanical Turk.” The Chronicle of Higher Education. *ProfHacker*, March 1, 2010. http://chronicle.com/blogs/profhacker/the-ethics-of-amazons-mechanical-turk/23010

[9] For an account of the history of the Galaxy Zoo Project see *Chris Lintott on The Galaxy Zoo*, 2010. http://www.youtube.com/watch?v=j_zQIQRr1Bo

[10] Raddick, M. Jordan, Georgia Bracey, Pamela L. Gay, Chris J. Lintott, Phil Murray, Kevin Schawinski, Alexander S. Szalay, and Jan Vandenberg. “Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers.” *Astronomy Education Review* 9, no. 1 (2010): 010103.

particularly well suited to answer the question “Why wasn’t I consulted.[11]” It is worth quoting him here at length:

> Why wasn't I consulted,” which I abbreviate as WWIC, is the fundamental question of the web. It is the rule from which other rules are derived. Humans have a fundamental need to be consulted, engaged, to exercise their knowledge (and thus power), and no other medium that came before has been able to tap into that as effectively.

He goes on to explain a series of projects that succeed because of their ability to tap into this human desire to be consulted.

> If you tap into the human need to be consulted you can get some interesting reactions. Here are a few: Wikipedia, StackOverflow, Hunch, Reddit, MetaFilter, YouTube, Twitter, StumbleUpon, About, Quora, Ebay,Yelp, Flickr, IMDB, Amazon.com, Craigslist, GitHub, SourceForge, every messageboard or site with comments, 4Chan, Encyclopedia Dramatica. Plus the entire Open Source movement.

Each of these cases tap into our desire to respond. Unlike other media, the comments section on news articles, or our ability to sign-up for an account and start providing our thoughts and ideas on twitter or in a tumblr is fundamentally about this desire to be consulted.

Returning to the example from Galaxy Zoo, where the carefully designed human computation classification exercise provides one kind of input, the projects very active web forums capitalize on the opportunity to consult. Importantly, some of the most valuable discoveries in the Galaxy Zoo project, including an entirely new kind of green colored galaxy, were the result of users sharing and discussing some of the images from the classification exercise in the open discussion forums.

Key Wisdom of Crowds Questions: How could we empower and consult with a community of users? Unlike human computation, the goal here is not users ability to process information or make judgments but their desire to provide their opinion.

Tools as Scaffolding

Helping someone succeed is often about getting them the right tools. Consider the image of scaffolding below. The scaffolding these workers are using puts them in a position to do their job. By standing on the scaffolding they are able to do their work without thinking about the tool at all[12]. In the activity of the work the tool disappears and allows them to go about their tasks taking for granted that they are suspended six or seven feet in the air. It is fruitful to think about a wide range of tools as serving as scaffolds.[13]

Scaffolding puts one in position to do a job.

All tools can act as scaffolds to enable us to accomplish a particular task. At this point it is worth briefly considering an example of how this idea of scaffolding translates into a cognitive task. In this situation I will briefly describe some of the process that is part of a park rangers regular work, measuring the diameter of a tree[14].

[11] Ford, Paul. “The Web Is a Customer Service Medium (Ftrain.com)”, January 6, 2011. http://www.ftrain.com/wwic.html

[12] This line of thinking is tied to Harman’s take on Heidegger’s notion of tools as either “read-to-hand” or “broken-at-hand” see Harman, Graham. *The Quadruple Object*. Zero Books, 2011 p 51-56.

[13] Here I am drawing on the Vygotskyan tradition of talking about “scaffolding” see Wood, D., Bruner, J., & Ross, G. (1976). The role of tutoring in problem solving. Journal of child psychology and psychiatry, 17, 89-100 and more broadly on the idea of cultural mediation see Vygotsky, L. S. *Mind in Society: The Development of Higher Psychological Processes*. Edited by Michael Cole, , 1978.

[14] This example comes from Pea, Roy. “Practices of Distributed Intelligence and Designs for Education.” In *Distributed Cognitions: Psychological and Educational Considerations*, edited by Gavriel Salomon. Learning in Doing. Cambridge, UK: Cambridge University Press, 1997.

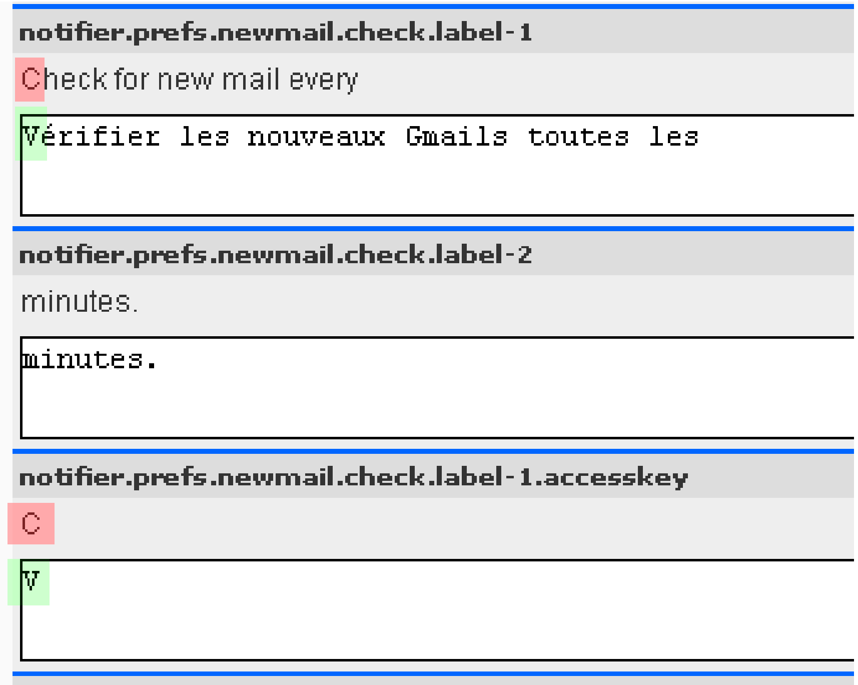
If you want to measure a tree you take a standard tape measure and do the following;

1. Measure the circumference of the tree
2. Remember that the diameter is related to the circumference of an object according to the formula circumference/diameter
3. Set up the formula, replacing the variable circumference with your value
4. Cross-multiply
5. Isolate the diameter by dividing
6. Reduce the fraction

Alternatively, you can just use a measuring tape that has the algorithm for diameter embedded inside it. In other words, you can just get a smarter tape measure. You can buy a tape-measure that was designed for this particular situation that can think for you (see the image below). Not only does this save you considerable time, but you end up with far more accurate measurements. There are far fewer moments for human error to enter into the equation.

The design of the tape measure has quite literally embedded the equations and cognitive actions required to measure the tree in its design.

This has a very direct translation into the design of online tools as well. For example, before joining the Library of Congress I worked on the Zotero project, a free and open source reference management tool. Zotero was translated into more than 30 languages by its users. The translation process was made significantly easier through BabelZilla. BabelZilla, an online community for developers and translators of extension for Firefox extensions, has a robust community of users that work to localize various extensions. One of the neatest features of this platform is that it stripes out the strings of text that need to be localized from the source code and then presents the potential translator with a simple web form where they just type in translations of the lines of text.



notifier.prefs.newmail.check.label-1

Check for new mail every

Vérifier les nouveaux Gmails toutes les

notifier.prefs.newmail.check.label-2

minutes.

minutes.

notifier.prefs.newmail.check.label-1.accesskey

C

V

This not only makes the process much simpler and quicker it also means that potential translators need absolutely no knowledge of the programming to contribute a localization. Without BabelZilla, a potential translator would need to know about how Firefox Extension locale files work, and be comfortable with editing XML files in a text editor. But BabelZilla scaffolds the user over that required knowledge and just lets them fill out translations in a web form.

If we again return to the Galaxy Zoo example, we can now think of the classification game as a scaffold which allows interested amateurs to participate at the cutting edge of scientific inquiry. In this scenario, the entire technical apparatus, all of the technical equipment used in the Sloan Digital Sky Survey, the design of the Galaxy Zoo site, and the work of all of the scientists and engineers that went into those systems are all part of one big scaffold that puts a user in the position to contribute to the frontiers of science through their actions on the website[15].

Scaffolding Users Key Question: How can our tools act as scaffolds to help make the most of users efforts? What expertise can we embed inside the design of our tools to magnify our users efforts? How can our tools put a potential user in exactly the right position with the right just in time knowledge to accomplish a given activity?

Understanding User Motivation

Asking why someone would want to participate is a critical question. Before going into explaining why I think people want to participate I will provide an example from a crowdsourcing transcription project.

Ben Brumfield runs a range of crowdsourcing transcription projects[16]. At one point in a transcription project he noticed that one of his power users was slowing down, cutting back significantly on the time they spent transcribing these manuscripts. The user explained that they had seen that there weren't that many manuscripts left to transcribe. For this user, the 2-3 hours a day they spent working on transcriptions was an important part of their day that they had decided to deny themselves some of that experience. For this users, participating in this project was so important to them, contributing to it was such an important part of who they see themselves as, that they needed to ration out those remaining pages. They wanted to make sure that the experience lasted as long as they could. When Ben found that out he quickly put up some more pages. This particular story illustrates several broader points about what motivates us.

After a person's basic needs are covered (food, water, shelter etc.) they tend to be primarily motivated by things that are not financial. People identify and support causes and projects that provide them with a sense of purpose. People define themselves and establish and sustain their identity and sense of self through their actions. People get a sense of meaning from doing things that matter to them. People find a sense of belonging by being a part of something bigger than themselves[17]. Projects that can tap into these identities, senses of purpose that can provide a little bit of a sense of meaning are projects that far from exploiting people can provide a way for people to connect with each other and make meaningful contributions to public goods.

This is one of the places where Libraries, Archives and Museums have the most to offer. As stewards of cultural memory these institutions have a strong sense of purpose and their explicit mission is to serve the public good. When we take seriously this call, and think about what the collections of culture heritage institutions represent, instead of crowdsourcing representing a kind of exploitation for labor it has the possibility to be a way in which cultural heritage institutions connect with and provide meaning full experiences with the past.

Motivating Users Key Questions: Whose sense of purpose does this project connect to? What identities are involved? What kinds of people does this matter to and how can we connect with and invite in the participation of those people.

Key Questions for Crowdsourcing Projects

To recap, here again are the four areas and their four sets of key questions. I think if a project has good answers to each of these four sets of questions it is well on its way toward success.

Key human computation questions: How could we use human judgment to augment computer processable information? It would be a waste of the public's time to invite them in to complete a task that a computer could already complete. The value human computation offers is the question of how the unique capabilities of people can be integrated into systems for the creation of public goods.

Key Wisdom of Crowds questions: How could we empower and consult with a community of users? Unlike human computation, the goal here is not users ability to process information or make judgments but their desire to provide their opinion.

Key Scaffolding Users Key Questions*:* How can our tools act as scaffolds to help make the most of users efforts? What expertise can we embed inside the design of our tools to magnify our users efforts? How can our tools put a potential user in exactly the right position with the right just in time knowledge to accomplish a given activity?

Key Motivating Users Key Question: Whose sense of purpose does this project connect to? What identities are involved? What kinds of people does this matter to and how can we connect with and invite in the participation of those people.

[15] This broader understanding of tools combines Andy Clark's notion of cognitive extension see Clark, Andy. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, USA, 2008.

[16] For background on Ben's work and projects see his blog http://manuscripttranscription.blogspot.com/

[17] For a popular account of much of the research behind these ideas see Pink, Daniel H. *Drive: The Surprising Truth About What Motivates Us*. Riverhead Hardcover, 2009 for some of the more substantive and academic research on the subject see essays in Elliot, Andrew J., and Carol S. Dweck, eds. *Handbook of Competence and Motivation*. The Guilford Press, 2005.

Example Cultural Heritage Crowdsourcing Projects

Citizen Archivist Dashboard
http://www.archives.gov/citizen-archivist/
Where citizen archivists can tag, transcribe, edit articles, upload scans, and participating in contests all related to the records of the US National Archives.

Trove
http://trove.nla.gov.au/
User's correct ocr'ed newspaper, upload images, tagged items, post comments and add lists.

GLAM Wiki
http://outreach.wikimedia.org/wiki/GLAM/Model_projects
The GLAM-WIKI project supports GLAMs and other institutions who want to work with Wikimedia to produce open-access, freely-reusable content for the public.

Old Weather
http://www.oldweather.org/
Old Weather invites you to help reconstruct the climate by transcribing old weather records from ships logs.

Galaxy Zoo
http://www.galaxyzoo.org/
Interactive project that allows the user to participate in a large-scale project of research: classifying millions of images of galaxies found in the Sloan Digital Sky.

UK Sound Map
http://sounds.bl.uk/Sound-Maps/UK-Soundmap
http://britishlibrary.typepad.co.uk/archival_sounds/uk-soundmap/
The UK Soundmap, invited people to record the sounds of their environment, be it at home, work or play.

What's on the menu
http://menus.nypl.org/
Help The New York Public Library improve a unique collectionWe're transcribing our historical restaurant menus, dish by dish, so that they can be searched by what people were eating back in the day. It's a big job so we need your help!

STEVE
http://tagger.steve.museum/
A place where you can help museums describe their collections by applying keywords, or tags, to objects.

Further Reading & Viewing

Ahn, L. von. (2006). *Human Computation*. Google TechTalks. Retrieved from http://video.google.com/videoplay?docid=-8246463980976635143

Brumfield, B. W. (2012, March 17). Collaborative Manuscript Transcription: Crowdsourcing at IMLS WebWise 2012. *Collaborative Manuscript Transcription*. Retrieved April 25, 2012, from http://manuscripttranscription.blogspot.com/2012/03/crowdsourcing-at-imls-webwise-2012.html

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, USA.

Crowdsourcing Cultural Heritage: The Objectives Are Upside Down http://www.trevorowens.org/2012/03/crowdsourcing-cultural-heritage-the-objectives-are-upside-down/

deterding, sebastian. (2011, February 19). *Meaningful Play: Getting Gamification Right*. Retrieved from http://www.youtube.com/watch?v=7ZGCPap7GkY&feature=youtube_gdata_player

Ford, P. (2011, January 6). The Web Is a Customer Service Medium (Ftrain.com). Retrieved May 3, 2012, from http://www.ftrain.com/wwic.html

Gee, J. P. (2000). Identity as an analytic lens for research in education. *Review of research in education*, *25*(1), 99.

Gee, James Paul. (2003). *What Video Games Have to Teach Us About Learning and Literacy* (New Ed.). Palgrave Macmillan. Retrieved from http://www.amazon.com/dp/1403965382

Holley, R. (2010). Crowdsourcing: How and Why Should Libraries Do It? *D-Lib Magazine*, *16*(3/4). doi:10.1045/march2010-holley

Hutchins, E. (1995). How a Cockpit Remembers Its Speed. *Cognitive Science*, *19*, 288, 265.

Juul, J. (2011, April 2). Gamification Backlash Roundup. *The Ludologist*. Retrieved April 25, 2012, from http://www.jesperjuul.net/ludologist/gamification-backlash-roundup

Karen Smith-Yoshimura. (2012). *Social Metadata for Libraries, Archives, and Museums: Executive Summary*. Dublin, Ohio:: OCLC Research. Retrieved from http://www.oclc.org/research/publications/library/2012/2012-02.pdf

Oomen, J., & Aroyo, L. (2011). Crowdsourcing in the cultural heritage domain: Opportunities and challenges. *Proceedings of the 5th International Conference on Communities and Technologies* (pp. 138–149).