New York University
Tisch School of the Arts
New York, NY

CROWDSOURCING METADATA IN CULTURAL INSTITUTIONS:

The story so far and a look to the future

by

Pamela Vizner Oyarce

A thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Arts

Moving Image Archiving and Preservation Program

Department of Cinema Studies

New York University

May 2014

ACKNOWLEDGEMENTS

I would like to thank all the people who made this thesis possible. For their support and help, thank you: Howard Besser (NYU), Antonia Lant (NYU), Kara van Malssen (NYU, AVPS), Mary Flanagan (Dartmouth College), Sukdith Punjasthitkul (Dartmouth College), Lacy Schultz (Museum of the City of New York), Todd Carter (Tagasauris), Johan Oomen (Netherlands Institute for Sound and Vision). Special thanks to my family in Chile for their unconditional support and to all my friends in New York for being my family away from home.

TABLE OF CONTENTS

1. INTRODUCTION 1

1.1 BACKGROUND 1

1.2 STATEMENT OF PURPOSE 3

1.3 CHALLENGES FOR CULTURAL INSTITUTIONS IN THE DIGITAL WORLD 6

1.4 MAIN ISSUES WITH CROWDSOURCING AND LINKED DATA 10

2. **CROWDSOURCING: Some examples** 14

2.1 THE COMMONS: Library of Congress Flickr Project 15

2.2 TROVE: The Australian Project 16

2.3 WIKIMEDIA FOUNDATION 18

2.4 GALAXYZOO: Classifying Galaxies 20

2.5 STEVE.MUSEUM PROJECT 21

2.6 MECHANICAL TURK 23

2.7 TAGASAURIS 25

3. CROWDSOURCING AND LINKED DATA: Basics 27

3.1 CROWDSOURCING: History and Definitions 27

3.2 CROWDSOURCING: Looking for a Definition 28

3.3 SEMANTIC WEB: History and definitions 37

4. UNDERSTANDING THE SOCIAL DIGITAL CULTURE: Communities' Online Behavior 43

4.1 CLAY SHIRKY'S THEORY: Means, Motive and Opportunity 44

4.1.1 Means 46

4.1.2 Motive 47

4.1.3 Opportunity 48

4.2 MEMBERSHIP AND DIGITAL CULTURE 49

4.3 FEWER RULES, BETTER RESULTS? 49

4.4 COGNITIVE AUTHORITIES 51

4.5 MARKETS, THE POSSITION OF NON-PROFITS AND PROFESSIONALISM 53

4.6 GAMES WITH A PURPOSE 55

4.7 QUALITY AND LONG-TERM VALUE 56

5. CROWDSOURCING METADATA: Issues and challenges 58

5.1 TAGGING TIME-BASED MEDIA 59

5.2 CONTROLLED VOCABULARIES 60

5.3 TAGS AND THE SEMANTIC WEB 62

5.4 DIVERSITY, A BLESSING AND A CURSE 63

5.5 PROFESSIONAL METADATA VS. USER METADATA 65

5.6 IMPROVING AND VALIDATING TAGS 67

5.7 COPYRIGHT AND LEGAL ISSUES 69

5.8 PLATFORMS FOR CROWDSOURCING TAGS 70

6. CROWDSOURCING METADATA: Three Case Studies 72

6.1 METADATA GAMES 73

6.1.1 Zen Tag, Nextag, Guess What! and Pyramid Tag 74

6.1.2 One Up 75

6.2 MUSEUM OF THE CITY OF NEW YORK AND TAGASAURIS 80

6.3 WAISDA? 86

7. CONCLUSIONS 96

7.1 BUILDING ONLINE COMMUNITIES 97

7.2 IN SEARCH OF THE PERFECT PLATFORM 99

7.3 INTEGRATING TAGS 102

7.4 VALIDATION OF TAGS AND LINKED DATA 103

7.5 LEAVING THE PROTOTYPE STAGE BEHIND 104

APPENDIX 1: More about Linked Open Data 106

APPENDIX 2: Glossary 109

BIBLIOGRAPHY 116

WEBOGRAPHY 120

1. INTRODUCTION

1.1 BACKGROUND

Since its creation, the Internet has progressively made its way into our lives, changing the way we relate to technology and people. First, before the introduction of the Web, Internet allowed people to connect, communicate and share, using tools such as FTP (to download and upload content) and usenet newsgroup systems (a peer-to-peer file exchange system). It was an interactive two-way communication tool.[1]

However, the Web transformed what once was an interactive system into a one-way channel, where little information was actually introduced by users, who became plain consumers.[2] Access to information and communication were easier than ever, but users were very much limited to consumption since mass media corporations had the resources to control high-speed services. This of course did not stop its growth: by the beginning of the 21st century the number of computers in people's homes started increasing to reach the point where having a computer was no longer a luxury. However, it would take almost a decade for social interactions to flourish, allowing people to not only consume information but also to produce it, slowly going back to the two-way model, now called Web 2.0. It was precisely its massiveness, its progressively lower costs of access and our need for social interaction that would allow this phenomenon.

[1] "What life was before the Web", Matt Welsh, Feb 24, 2011, http://matt-welsh.blogspot.com/2011/02/what-life-was-like-before-web.html

[2] "Use of Non-Broadcast Channels to Communicate Information In Social Change Situations:Berkeley Anti-Apartheid and Solidarity Poland", Howard Besser, Jan. 21, 1986, http://besser.tsoa.nyu.edu/howard/Papers/Poland-berkeley/

The Web was no longer about consumption only; people could interact with each other through forums, participative and social websites and platforms that flooded the Web. The Web 2.0 or the so-called Social Web was here to stay and it became the perfect scenario for the creation of new models and innovative projects including the participation of the *crowd* bringing with them the *participatory culture* we live in online[3]. Open source and crowdsourcing initiatives were the perfect fit in this interactive digital universe. As Daren C. Brabham mentions in his book *Crowdsourcing,* these projects are possible because of the qualities of the Internet: speed, reach, temporal flexibility, anonymity, interactivity, low barriers to entry and the ability to carry all types of content.[4]

More recently, we have been witnesses of the advent of the Semantic Web, an improved and much more efficient Web, capable of connecting content in a way we never experienced before, which is possible thanks to its openness and the ability to allow machines to link data to establish complex relationships between objects and concepts. Certainly, Web 2.0 was part of the foundation for this idea, where free and openly available data can be easily shared and exchanged.

The Social Web has made a huge impact on our lives, and as such, this phenomenon must be studied from every angle. The Semantic Web, while still very new, has taken huge steps in the past two years, giving us a clue of what is about to happen. The application of crowdsourcing technologies, in particular, requires input from many different fields of study: from economics to

[3] We are used to call this new online culture as *participatory*. Clay Shirky, however, argues that the 20th century was an exception to our natural participatory or community-based behavior, therefore we had the need to use the work *participatory* to describe the new way we interact with each other online.
Clay Shirky, “Cognitive Surplus” (New York : Penguin Press, 2010)

[4] Daren C. Brabham, “Crowdsourcing” (Cambridge, Massachusetts ; London, England : The MIT Press, 2013)

behavioral sciences, to computer technologies. In the information sciences, not much research has been done regarding the effectiveness of crowdsourcing models for cultural institutions, but many organizations have applied the model in different ways.[5] Furthermore, considering the impact that the Semantic Web is having in the commercial world online, it is expectable to have these technologies as a must in the pool of resources of every cultural organization. With that in mind, it is necessary to look back to what we have done in the past years, to step into the future with clear goals and realistic expectations.

1.2 STATEMENT OF PURPOSE

Despite the growth in the use of crowdsourcing and Semantic Web applications among cultural institutions as a solution for obtaining basic descriptions for digital collections, there is still a lack of general assessment of the projects involving both models. I think that having an environmental scan of their effectiveness and sharing those results can help improve future initiatives. These types of studies are key to provide a wide view of what the community has done so far and to redirect efforts to solve issues collectively. It is true that many of those projects are in a prototype stage, and many details, observations and improvements coming from the institutions leading them are still a work in progress, but I believe that taking a time to slow down and look back will definitely give a more clear idea of what our roles and future challenges are, and more generally, how the community can help solve them.

[5] The most remarkable examples of research in this area are the work done by The Institute for Sound and Vision in The Netherlands and the steve.museum Project. Both will be further explored in this document.

This study intends to provide an outside view to the use of crowdsourcing and the Semantic Web to gather metadata for audiovisual collections in the form of tags - tags being annotations “attached” to media that serve as basic descriptive information. For the sake of clarity, throughout this document I will refer to “audiovisual materials” as photographs and time-based media. This thesis intends to picture the state of the art, to give future projects a comprehensive view of the models and to deliver general information with the aim of improving future implementations.

This document will study the successful projects as well as those that, for different reasons, didn’t have optimum or desired results, which will allow us to understand what decisions, features and applications could make a successful crowdsourcing project. In order to do that, in the first three chapters I will provide background and definitions to built a common ground for the topics and discussions presented later on in this document

Chapter one presents the statement of purpose and some general background on the challenges that cultural institutions face in the advent of the digital era. It also provides some general concerns and observations about crowdsourcing, the Semantic Web and their applications; issues that will be addressed more in depth in later chapters.

In Chapter two I will first present an overview of some projects that were developed to improve collection descriptions in the recent years. Some of them will be reviewed due to their importance in the field, whether because they were pioneers or because their results and implementations give us an idea of what is crowdsourcing and what aspects are important to consider when embarking in crowdsourcing projects. They also represent a good background for the study of the implementation of crowdsourcing projects that include Semantic Web features. While this thesis will not deal with the Semantic Web directly, I will provide some basic information about it to contextualize the most recent crowdsourcing projects that include this model.

Chapter three will provide a detailed description of crowdsourcing, its history, definitions and fundamental features to give a common ground for the discussion in the following chapters. As a new model, some definitions, structures and descriptions are still in the process of development and there are still not many publications about them, hence the importance of providing a theoretical framework for further discussion. I will also provide a general description of the Semantic Web and Linked Data, which will suffice to understand how crowdsourcing can benefit from it.

Crowdsourcing and Linked Data, as models based on crowd and community participation should include basic discussions about social impact and mass behavior. Part of the success of these models lies in the way online communities interact with the Web and their perception of their own status inside the crowd. Chapter four will address these issues, primarily focusing on the description of such behaviors and their impact on community projects developed by cultural institutions. This chapter will also discuss institutions´ cognitive authority, the concept of Games with a Purpose and long-term value of crowdsourcing products.

Once the background and basis of both models are explained, I will tackle the implementation issues of both models. Chapter five will discuss difficulties in the development and application to discover their origins and how these could be solved. Discussions on crowdsourcing and copyright will also be briefly discussed.

Chapters six will serve to present and study three current initiatives using crowdsourcing and Linked Data to gather descriptive tags for audiovisual collections: *Metadata Games* by Tiltfactory at Dartmouth College, the project developed by the Museum of the City of New York and Tagasauris, and *Waisda?* by the Netherlands Institute for Sound and Vision and University of Amsterdam.

Finally, chapter seven will gather all the conclusions, observations and reflections of this study. Readers must understand that due to the highly exploratory characteristics of the projects using crowdsourcing and Linked Data, the conclusions drawn here could disappoint people looking for answers for implementation of future projects – although I do mention a couple of open source tools that are currently available. However, this section can be used as a guideline to assess your own status to evaluate the readiness of your institution in the implementation of crowdsourcing and social-media-based projects. As you will learn after reading this document, there's much more to it than just good and well-designed platforms.

This document is intended to be used by members of the archival community in general and people from cultural institutions looking to implement this type of projects, not only for audiovisual collections.

1.3 CHALLENGES FOR CULTURAL INSTITUTIONS IN THE DIGITAL WORLD

The beginning of the Social Web - and even the very beginning of the Internet – was already a huge benefit for cultural institutions. The new way of sharing information had an enormous impact on how collections are accessed by users. The social media also changed the way in which institutions and their target communities interact with each other, moving from unidirectional, where the museum talks at the public without feedback, to multi-directional. This more participative environment presented by technology offers new ways of interaction that can be beneficial both for institutions – who can provide better access – and users – who can enjoy better access.

But, at the same time, technology presents new challenges for institutions. Keeping up with the fast pace of technology developments is hard, it sometimes requires many resources - both economic and human – and they put institutions in a position where systems and workflows have to be rethought and redesigned in order to integrate them. Users and communities want modern institutions that fit their needs and provide easy access to their services.

That implies looking for solutions to provide access to all different types of collections, including paper documents and audiovisual materials. Preserving and providing access to some collections is a real challenge; to do so, especially for audiovisual collections, institutions have to deal with issues from natural chemical decay of physical items to technological obsolescence, i.e. the discontinuity in the manufacture of playback machines and their replacement parts. The latter will prevent us from extracting any content from them, even if the tapes are in good condition. Under this complicated scenario, the only realistic solution for preservation and access of these analog materials is digitization. The truth is, we are running out of time with many of these analog media, and digitization is the first step on the path towards future access, even if we can't yet figure out all the subsequent steps in detail.[6]

However, digital collections in cultural institutions not only come from self-initiated digitization plans. Born-digital materials have been around for more than twenty years and most institutions, not even close to solving the issues with analog media, have to deal with growing born-digital collections. Born-digital collections are present in almost every archive or library.

[6] A complete and extended evaluation of the longevity of physical audiovisual media, specifically for videotape, can be found on "The End of Analog Media: The Cost of Inaction and What You Can Do About It", Chris Lacinak, Nov. 8, 2013 http://www.avpreserve.com/wp-content/uploads/2013/11/Lacinak_COI_AMIA_2013_dist.pdf

This, added to the digitization of analog media has increased the growth of digital collections exponentially.[7]

Moreover, the overall fragility of digital collections puts in danger the only solution we have to save analog media. Software, hardware, processors, storage, they all can go obsolete quickly and fail unexpectedly. At a higher level, taking good care of digital collections requires even more considerations to take in account, that go beyond the preservation strategies applied so far for paper collections. To ensure digital preservation institutions need to establish systems and workflows that involve interdisciplinary team work, to comply with the three basic statements of long-term digital preservation: bit preservation, accessibility and usability, and sustainability.[8] As I mentioned before, all of them require institutions to develop comprehensive plans, strategies and policies that go beyond the solely custody of content.

Lack of metadata is also a hazard for digital collections. Metadata, i.e. data about data, is fundamental for the discoverability and access of digital content. For instance, technical metadata is truly critical for the retrieval of digital files since it includes the information used by machines to locate them on a storage device. But there are other types of metadata that provide other type of information about the digital objects, such as descriptive metadata.[9] Descriptive metadata provides information not about the digital file itself, but about its content. This kind of information is mostly utilized by users wanting to know what the file is about, if it's what they

[7] According to the International Data Corporation (IDC) from 2005 to 2020 the digital universe will grow by a factor of 300. "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East", IDC, Dec. 2012 http://idcdocserv.com/1414

[8] A complete description of the requirements for institutions to achieve successful digital preservation can be found in the "Reference Model for an Open Archival Information System (OAIS)", Consultative Committee for Space Data Systems, June 2012, http://public.ccsds.org/publications/archive/650x0m2.pdf

[9] More information about metadata standards and types of metadata on "Understanding Metadata", National Information Standards Organization (NISO), 2004 http://www.niso.org/publications/press/UnderstandingMetadata.pdf

are looking for. While not fundamental to the retrieval of the digital file, it is the only way to make digital collections discoverable for users – and sometimes also for institutions. And at the end, what is the point of having thousands of digital files and hundreds of digitized collections if users can't access them at all? Having a million files named "photo001.jpeg" together with a bunch of technical information such as "file extension: .PDF" will certainly not help researchers finding what they need. Bottom line: metadata is essential to make digital content available and descriptive information is fundamental to provide good access.

This is undoubtedly, a complicated scenario. Digital preservation not only requires changing the way things are being done so far, but it also requires a lot of resources and investment. Not many institutions can afford them. In terms of describing collections the problem is not far from complicated either. Having descriptive information unfortunately becomes a secondary need when dealing with digital files. Cataloging is a slow process and, at least until not many years ago, it required specialists who knew the complicated systems and standards. For instance, in the library world, cataloging books requires the knowledge of the MARC standard, which could only be applied by knowledgeable professional catalogers. There's no doubt that technology has given us tools to facilitate this process, but it still requires people in front of a screen.

In addition, digitization is normally faster than cataloging. For instance, the Museum of the City of New York started a digitization project of their photo collections to provide better access. However, catalogers couldn't keep up the pace of digitization, creating a huge backlog of digital materials that were not accessible to patrons.[10]

[10] Museum of the City of New York, "NEH Grant Final Report: Improving Digital Record Annotation Capabilities with Open-sourced Ontologies and Crowd-sourced Workers", April 30, 2013 https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-51480-11

Institutions need to find alternatives to describe their collections in order to provide good access, but also to keep up with the new technologies and to be able to communicate with users and patrons in the same language in the virtual world. Having limited resources, cultural institutions have searched for new online models, powered by the Social Web to try to solve these issues, such as crowdsourcing and Linked Data.

1.4 MAIN ISSUES WITH CROWDSOURCING AND LINKED DATA

Fortunately, even under the gloomy and fatalist reality of digital preservation described above, there is light at the end of the tunnel. Many institutions are looking for new ideas, systems and models to describe digital collections in order to improve access, optimizing time and resources. Crowdsourcing was one of them. I will provide a formal definition in Chapter three, but generally crowdsourcing is a participative online model for problem solving which entails mutual benefit, for users (the crowd) and organizers.

Crowdsourcing projects have been around for a relatively short time. Starting around 2006, this model was born as many other initiatives to benefit from the interactive World Wide Web. Despite its short existence and not long after its inception non-profit cultural institutions (and for-profits as well) realized its potential and have implemented several projects based on this now popular model.

While crowdsourcing can have many different applications, cultural institutions have applied the model primarily to identify and describe content that otherwise would be impossible to catalog because of time constraints, lack of resources or lack of staff. Thus, cultural organizations mostly crowdsource metadata.[11] As mentioned before, the creation of digital content has increased, enlarging the already huge backlog of descriptive metadata that collecting institutions have. Thus, crowdsourcing presents itself as a solution to tackle the urgency of basic description for growing digital collections in order to provide access.

The number of institutions applying the model has grown in the last years. This is because, in general, crowdsourcing can play several roles in the lifecycle of digital content inside cultural institutions, as highlighted by Johan Oomen and Lora Aroyo in *"Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges".* In this thesis I will only focus on crowdsourcing metadata, corresponding to the "description" stage of the image below.[12]

[11] Crowdsourcing metadata is not the only use of the model that can be beneficial for cultural institutions. Lora Aroyo and Johan Oomen describe six types of crowdsourcing: correction and transcription, contextualization, complementing collections, classification, co-curation and crowdfunding. Johan Oomen and Lora Aroyo, "Crowdsourcing in the cultural heritage domain: opportunities and challenges", in *Proceedings of the 5th International Conference on Communities and Technologies*, (New York, NY, USA: ACM, 2011), 138–149 http://www.iisi.de/fileadmin/IISI/upload/2011/p138_oomen.pdf

[12] An interesting environmental scan of the use of online resources and community participation to gather metadata was presented by the Online Computer Library Center in 2011. Karen Smith-Yoshimura and Cindy Shein, *Social Metadata for Libraries, Archives and Museums*, OCLC, Sept. 2011. http://www.oclc.org/content/dam/research/publications/library/2011/2011-02.pdf?urlm=162950

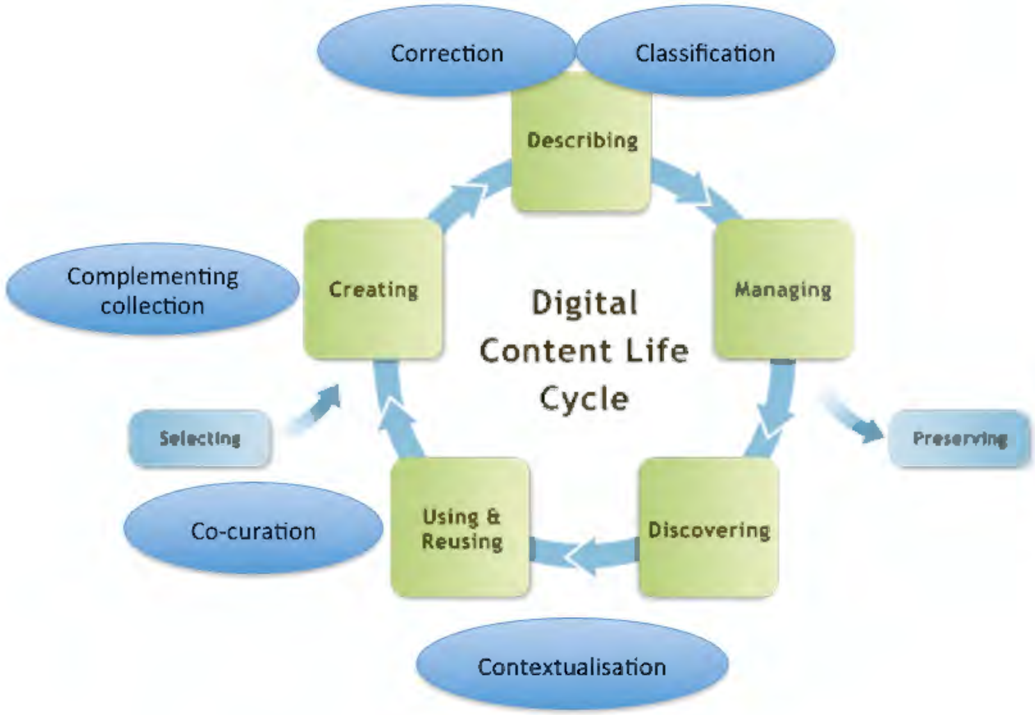


Figure 1. Digital Content Life Cycle and Crowdsourcing, based on the model created by the National Library of New Zealand.

However, the crowdsourcing model has some practical issues to consider. These issues have been a continued concern for researchers applying the model and, despite the difficulties, they are still working on improvements. This is undoubtedly why many of these projects never leave the prototype stage.

Crowdsourced metadata - regardless of the type of media it describes - presents a challenge when it comes to the validation of the information and how that information is integrated (or not) to the current descriptive systems, such as databases and catalogs, both locally and online. Since these initiatives are, most of them, open to the whole community to contribute, institutions need a system to ensure that the information is trustworthy and can be included in public catalogs. Cultural institutions are a source of trustworthy information and knowledge where community can find true answers to their inquiries. These issues will be further discussed in Chapter four.

The lack of vocabulary control is another problem in crowdsourcing projects. Professional catalogers and librarians follow standards and controlled vocabularies to describe collections, to improve access. However, when using the crowd, the data gathered, mostly in the form of tags, can be very messy; the people participating in these projects are of course not professional catalogers, and, even with the best of the intentions, issues such as folksonomies and synonyms arise.

Lately, some crowdsourcing projects have integrated the Semantic Web principles through Linked Data applications to solve these issues. Many institutions and other collaborative projects, such as Bibframe by the Library of Congress, Europeana and GeoNames[13], have been using Linked Open Data - free and open Linked Data available as data sets online - either by making data sets available online or directly collaborating with bigger Linked Data projects such as DBpedia and Freebase. Furthermore, although out of the scope of this thesis, it is also worth mentioning that the use of Linked Data for commercial purposes is growing everyday. However, despite the popularity and increasing use of both models, the application of Linked Open Data in crowdsourcing projects is a very new, quite unexplored but certainly promising match.

[13] Library of Congress Bibframe (Bibliographic Framework Initiative)
http://www.loc.gov/bibframe/
Europeana Linked Open Data Project http://pro.europeana.eu/linked-open-data
GeoNames, Geographical Database http://www.geonames.org/

2. CROWDSOURCING: Some Examples

During my research I found several examples of crowdsourcing in different cultural institutions worldwide. In order to present some examples I made a selection of them, considering the most important projects whether for its contribution to the field or other attributes I thought worth mentioning. I believe having this background will provide the reader an overview of the challenges and possibilities of the model, as well as providing a common ground for the discussions in the next chapters. I also included some crowdsourcing projects that exist in the commercial sector, to present the wide variety of applications of the model but also to understand the particular situation of non-profit initiatives.

It is worth noting that according to the latest discussions on definitions of crowdsourcing – which will be addressed in Chapter three - some of the projects described below may not be considered crowdsourcing projects. However, all of them represent initial efforts of engaging new communities using the interactivity provided by the web, as well as the new opportunity of using open online platforms to provide access to collections. In a normal situation, I would present the definitions before describing the projects, but I think that, in this particular case, it is easier to understand the attempts to describe the model if we first have some information about their implementation. Second, it is a way to reflect the historicity of this model – and probably of other models using online communities – where definitions and classifications are hard to determine in an ever-changing environment.

2.1 *THE COMMONS:* Library of Congress Flickr Project[14]

The Commons is a project created by the Library of Congress in 2008 with the aim of increasing access to their public domain photograph collection by using popular platforms online (in this case Flickr). Other goals include obtaining information through community input and participating in the interactive web by having a strong presence and growing community online. This is one of the most important projects in the realm of photograph tagging online and one of the first of its kind.

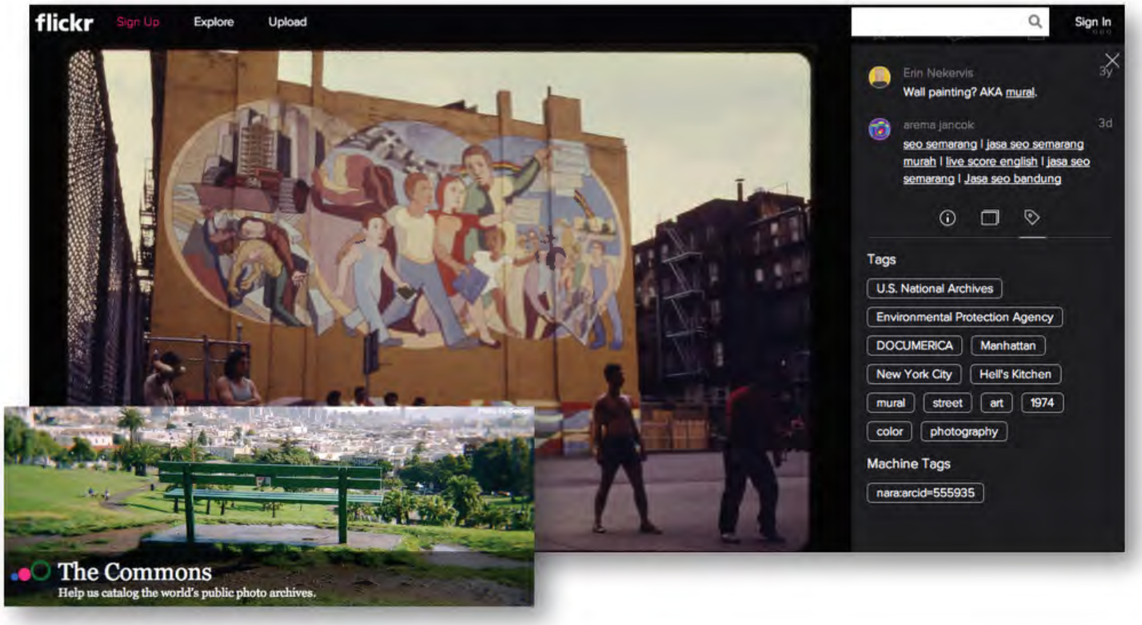
After choosing Flickr as an online venue for the project – a free image an video hosting service online owned by Yahoo Inc. - the Flickr team created a special area in the website for *The Commons* that would fit the need for user's input and copyright status of the collections available through the website.

At present, eighty institutions participate in the project and the site is open for further collaborations. Some institutions participating in this project are: The Royal Library of Denmark, NASA on The Commons, The Nationaal Archief of Netherlands, Bibliotèque de Toulouse, just to name a few. Institutions, previous to signing up, need to agree with the Terms of Service agreement with the company, create their own Flickr account and, most important of all, agree with the "no known copyright restrictions" policy of the project (which also includes the publication of a rights statement on the institution's website).

From the point of view of the users, addition of tags and comments are allowed, providing that they previously own a Flickr account. Photographs can be accessed, searched,

[14] *The Commons*' website in Flickr http://www.flickr.com/commons. Library of Congress, Various Authors, *For the Common Good: The Library of Congress Flickr Pilot Project,* 2008, http://www.loc.gov/rr/print/flickr_report_final_summary.pdf

used and tagged (or even untagged) by anyone. Clicking on a tag (see the list of tags on the right in Figure 2 below) will show you all the photographs including that tag, not only inside *The Commons* but in the whole Flickr website, which creates a cross-project platform. However, there is no information about any integration of the tagging system to a library or other online catalogs, in other words, tags are searched and displayed as such, thus there is no validation process going on.



Erin Nekervis 3y
Wall painting? AKA mural.

arema jancok 3d
seo semarang | jasa seo semarang murah | live score english | jasa seo semarang | Jasa seo bandung

Figure 2. The Commons *Project on Flickr.*

2.2 *TROVE:* The Australian Project[15]

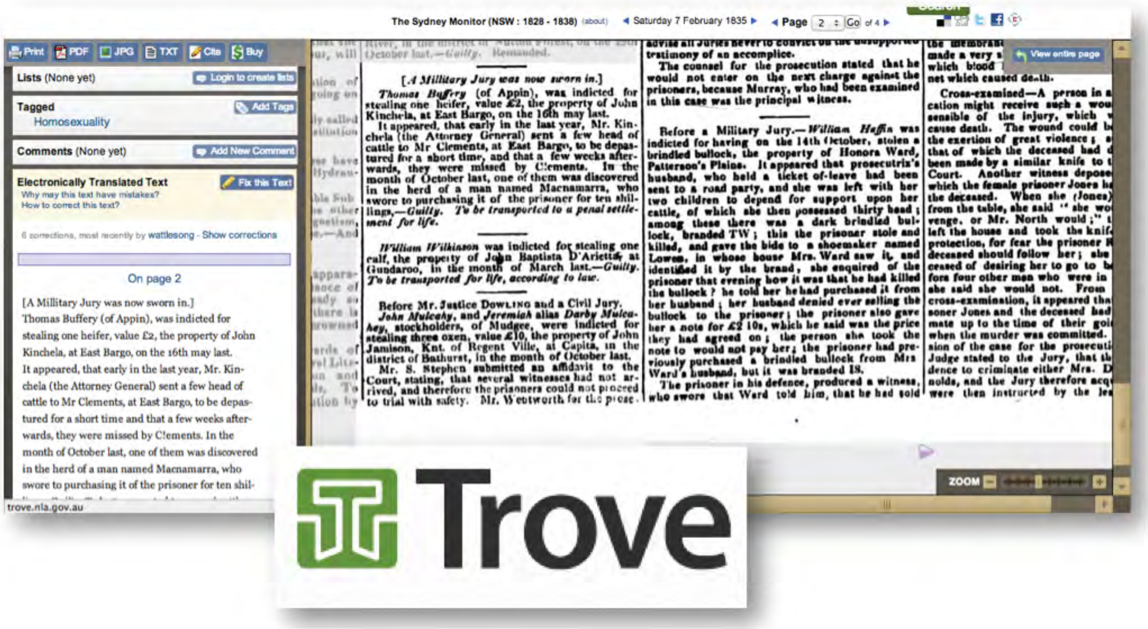
The first online access project by the National Library of Australia was *PictureAustralia*, created in 2006. Initially *PictureAustralia* was created to enable the community to upload their

[15] Trove Website http://trove.nla.gov.au/

own pictures to the Flickr site, as well as adding descriptions and tags (using crowdsourcing obtain metadata and enlarge digital collections). The aim of the project was to increase the contemporary photographs collections of the National Library and to engage new audiences.

Later, *PictureAustralia* was absorbed by the *Trove*, a huge common library online, where users can not only search the catalogs but also have access to all types of materials, from documents and newspapers, to photographs, music and video. The interesting part of this online library is that users can actively contribute by tagging, adding comments, uploading their own pictures – this was the previous *PictureAustralia* – and even correcting electronically translated text (machine-generated Optical Character Recognition, OCR, see Figure 3 below). Contributors are encouraged to provide enough descriptive information - either comments, a general description or tags - about the content uploaded by them. Some contributions can also be made without previous registration.

The *Trove* project also includes contribution from other institutions in Australia, other than the National Library, offering a common search platform for their content.



The Sydney Monitor (NSW : 1828 - 1838) (about) Saturday 7 February 1835 Page 2 Go of 4

Lists (None yet)

Tagged
Homosexuality

Comments (None yet)

Electronically Translated Text
Why may this text have mistakes?
How to correct this text?

6 corrections, most recently by wattlesong - Show corrections

On page 2

[A Millitary Jury was now sworn in.]
Thomas Buffery (of Appin), was indicted for stealing one heifer, value £2, the property of John Kinchela, at East Bargo, on the 16th may last.
It appeared, that early in the last year, Mr. Kinchela (the Attorney General) sent a few head of cattle to Mr Clements, at East Bargo, to be depastured for a short time and that a few weeks afterwards, they were missed by C!ements. In the month of October last, one of them was discovered in the herd of a man named Macnamarra, who swore to purchasing it of the prisoner for ten shil-

October last.—*Guilty.* Remanded.

[*A Millitary Jury was now sworn in.*]

Thomas Buffery (of Appin), was indicted for stealing one heifer, value £2, the property of John Kinchela, at East Bargo, on the 16th may last.

It appeared, that early in the last year, Mr. Kinchela (the Attorney General) sent a few head of cattle to Mr Clements, at East Bargo, to be depastured for a short time, and that a few weeks afterwards, they were missed by C!ements. In the month of October last, one of them was discovered in the herd of a man named Macnamarra, who swore to purchasing it of the prisoner for ten shillings,—*Guilty. To be transported to a penal settlement for life.*

William Wilkinson was indicted for stealing one calf, the property of John Baptista D'Arietta, at Gundaroo, in the month of March last.—*Guilty. To be transported for life, according to law.*

Before Mr. Justice Dowling and a Civil Jury.

John Mulcahy, and *Jeremiah* alias *Darby Mulcahey*, stockholders, of Mudgee, were indicted for stealing three oxen, value £10, the property of John Jamison, Knt. of Regent Ville, at Capita, in the district of Bathurst, in the month of October last.

Mr. S. Stephen submitted an affidavit to the Court, stating, that several witnesses had not arrived, and therefore the prisoners could not proceed to trial with safety. Mr. Wentworth for the prosec.

testimony of an accomplice.

The counsel for the prosecution stated that he would not enter on the next charge against the prisoners, because Murray, who had been examined in this case was the principal witness.

Before a Military Jury.—*William Haffin* was indicted for having on the 14th October, stolen a brindled bullock, the property of Honora Ward, Patterson's Plains. It appeared that prosecutrix's husband, who held a ticket of-leave had been sent to a road party, and she was left with her two children to depend for support upon her cattle, of which she then possessed thirty head; among these there was a dark brindled bullock, branded TW; this the prisoner stole and killed, and gave the hide to a shoemaker named Lowes, in whose house Mrs. Ward saw it, and identified it by the brand, she enquired of the prisoner that evening how it was that he had killed the bullock? he told her he had purchased it from her husband; her husband denied ever selling the bullock to the prisoner; the prisoner also gave her a note for £2 10s, which he said was the price they had agreed on; the person she took the note to would not pay her; the prisoner had previously purchased a brindled bullock from Mrs Ward's husband, but it was branded IS.

The prisoner in his defence, produced a witness, who swore that Ward told him, that he had sold

made a very
which blood
net which caused death.

Cross-examined—A person in a
cation might receive such a wou
sensible of the injury, which
cause death. The wound could b
the exertion of great violence;
that of which the deceased had
been made by a similar knife to
Court. Another witness depose
which the female prisoner Jones h
the deceased. When she (Jones)
from the table, she said "she wo
venge, or Mr. North would;"
left the house and took the knif
protection, for fear the prisoner
deceased should follow her; she
ceased of desiring her to go to b
fore four other men who were in
she said she would not. From
cross-examination, it appeared tha
soner Jones and the deceased had
mate up to the time of their goi
when the murder was committed.
sion of the case for the prosecuti
Judge stated to the Jury, that th
dence to criminate either Mrs. D
nolds, and the Jury therefore acq
were then instructed by the

Figure 3. Trove*, Transcription Platform.*

2.3 WIKIMEDIA FOUNDATION

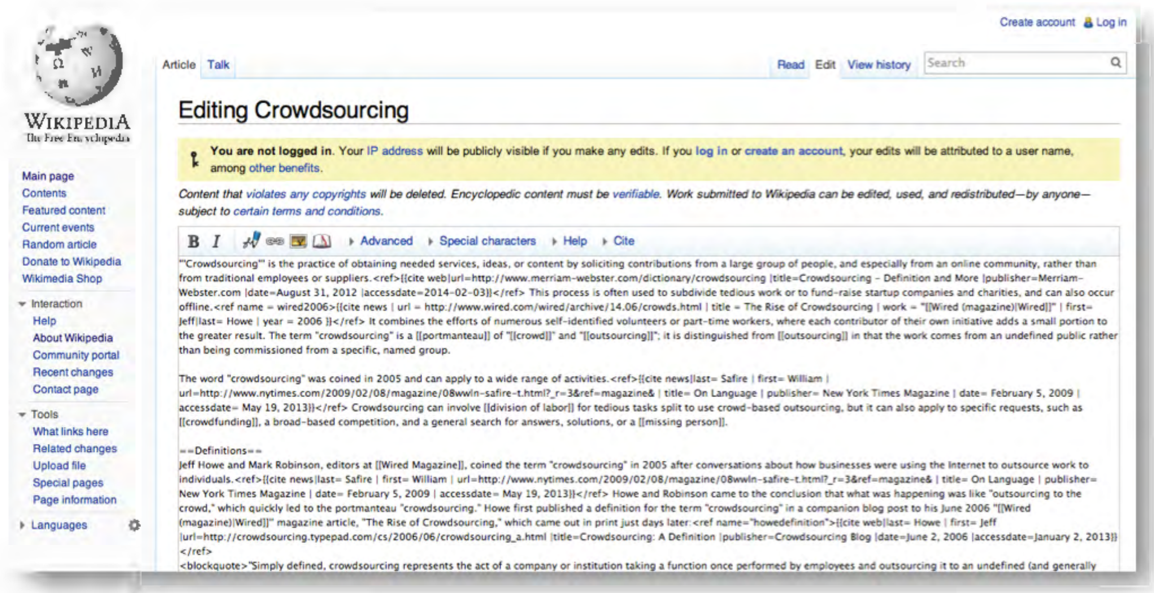
Widely known by almost all Internet users is the *WikiMedia* network of encyclopedias online, including *Wikipedia, Wikisource, Wiktionary, Wikiversity*, and others. It is one of the largest and most popular encyclopedias online, with the participation of more than 25 million volunteers-authors around the world, published in more than 285 languages. In its ten years of existence it has become the largest collection of shared knowledge in the history. In their website they state "*The people who support it are united by their love of learning, their intellectual curiosity, and their awareness that we know much more together than any of us does alone.*"[16]

In *Wikipedia*, anyone can edit or add new articles - as long as they follow the project's guidelines: Five Pillars of Wikipedia - except in some specific cases, where these actions are

[16] Wikimedia http://wikimediafoundation.org/wiki/FAQ/en

restricted to prevent disruption or vandalism. The integrity of *Wikipedia* is kept thanks to the so called “administrators”, i.e. editors who have been granted special permissions, such as blocking specific IP addresses, delete and undelete pages and other privileges. Anyone can apply to be an administrator.

As I will discuss in the next chapter, according to Brabham’s definitions of crowdsourcing, *Wikipedia* can’t be considered one of them. He argues that crowdsourcing implies that “the locus of control” of the creation of goods and ideas must reside between the organization and the users. However, regardless of the status of *Wikipedia* as a crowdsourcing platform, the truth is the level of participation and commitment of the users is a clear example of the *crowd* contributing free content and information online.



Editing Crowdsourcing

You are not logged in. Your IP address will be publicly visible if you make any edits. If you log in or create an account, your edits will be attributed to a user name, among other benefits.

Content that violates any copyrights will be deleted. Encyclopedic content must be verifiable. Work submitted to Wikipedia can be edited, used, and redistributed—by anyone—subject to certain terms and conditions.

```
'''Crowdsourcing''' is the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community, rather than from traditional employees or suppliers.<ref>{{cite web|url=http://www.merriam-webster.com/dictionary/crowdsourcing |title=Crowdsourcing - Definition and More |publisher=Merriam-Webster.com |date=August 31, 2012 |accessdate=2014-02-03}}</ref> This process is often used to subdivide tedious work or to fund-raise startup companies and charities, and can also occur offline.<ref name = wired2006>{{cite news | url = http://www.wired.com/wired/archive/14.06/crowds.html | title = The Rise of Crowdsourcing | work = "[[Wired (magazine)|Wired]]" | first= Jeff|last= Howe | year = 2006 }}</ref> It combines the efforts of numerous self-identified volunteers or part-time workers, where each contributor of their own initiative adds a small portion to the greater result. The term "crowdsourcing" is a [[portmanteau]] of "[[crowd]]" and "[[outsourcing]]"; it is distinguished from [[outsourcing]] in that the work comes from an undefined public rather than being commissioned from a specific, named group.

The word "crowdsourcing" was coined in 2005 and can apply to a wide range of activities.<ref>{{cite news|last= Safire | first= William |
url=http://www.nytimes.com/2009/02/08/magazine/08wwln-safire-t.html?_r=3&ref=magazine& | title= On Language | publisher= New York Times Magazine | date= February 5, 2009 | accessdate= May 19, 2013}}</ref> Crowdsourcing can involve [[division of labor]] for tedious tasks split to use crowd-based outsourcing, but it can also apply to specific requests, such as [[crowdfunding]], a broad-based competition, and a general search for answers, solutions, or a [[missing person]].

==Definitions==
Jeff Howe and Mark Robinson, editors at [[Wired Magazine]], coined the term "crowdsourcing" in 2005 after conversations about how businesses were using the Internet to outsource work to individuals.<ref>{{cite news|last= Safire | first= William | url=http://www.nytimes.com/2009/02/08/magazine/08wwln-safire-t.html?_r=3&ref=magazine& | title= On Language | publisher= New York Times Magazine | date= February 5, 2009 | accessdate= May 19, 2013}}</ref> Howe and Robinson came to the conclusion that what was happening was like "outsourcing to the crowd," which quickly led to the portmanteau "crowdsourcing." Howe first published a definition for the term "crowdsourcing" in a companion blog post to his June 2006 "[[Wired (magazine)|Wired]]" magazine article, "The Rise of Crowdsourcing," which came out in print just days later:<ref name="howedefinition">{{cite web|last= Howe | first= Jeff |url=http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html |title=Crowdsourcing: A Definition |publisher=Crowdsourcing Blog |date=June 2, 2006 |accessdate=January 2, 2013}}
</ref>
<blockquote>"Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally
```

Figure 4. Editing screen for the article about “Crowdsourcing” on Wikipedia.

2.4 *GALAXY ZOO*: Clasifying galaxies[17]

The precedent of this very particular crowdsourcing project began with the Sloan Digital Sky Survey (SDSS), which in 2000 gathered one of the first and largest digital maps of the Universe. This survey resulted in millions of unclassified images of galaxies.

Born in 2007 and now operated by Zooniverse[18], this project uploaded the images on their website for people to classify. It relies on the fact that you can derive a lot of information about the galaxies with only determining its shape. The validation of the information is done by multiple entries of the same shape; in other words, images that have the same shape classifications from all the volunteers are more trustworthy.

In the new version of the project in 2012, the site reached almost 70,000 classifications an hour within the first 24 hours of launching. This second phase also included a more complex set of tasks for volunteers: determining the number of spiral arms and the size of the bulges of the Sloan galaxies. All this information has been used in many studies and scientific papers in astronomy.[19]

The most interesting thing about this project is that, despite what many would believe, people did not need to be experts, the site provides guidelines and educates users to do the job correctly. This project is a great example of the incredible possibilities the crowd gives, if tasks and project scope are well defined and well communicated.

[17] Galaxy Zoo website http://www.galaxyzoo.org/

[18] Zoouniverse is an online platform for citizen science projects. https://www.zooniverse.org/

[19] A list of published papers that used GalaxyZoo's Data: http://www.galaxyzoo.org/#/papers

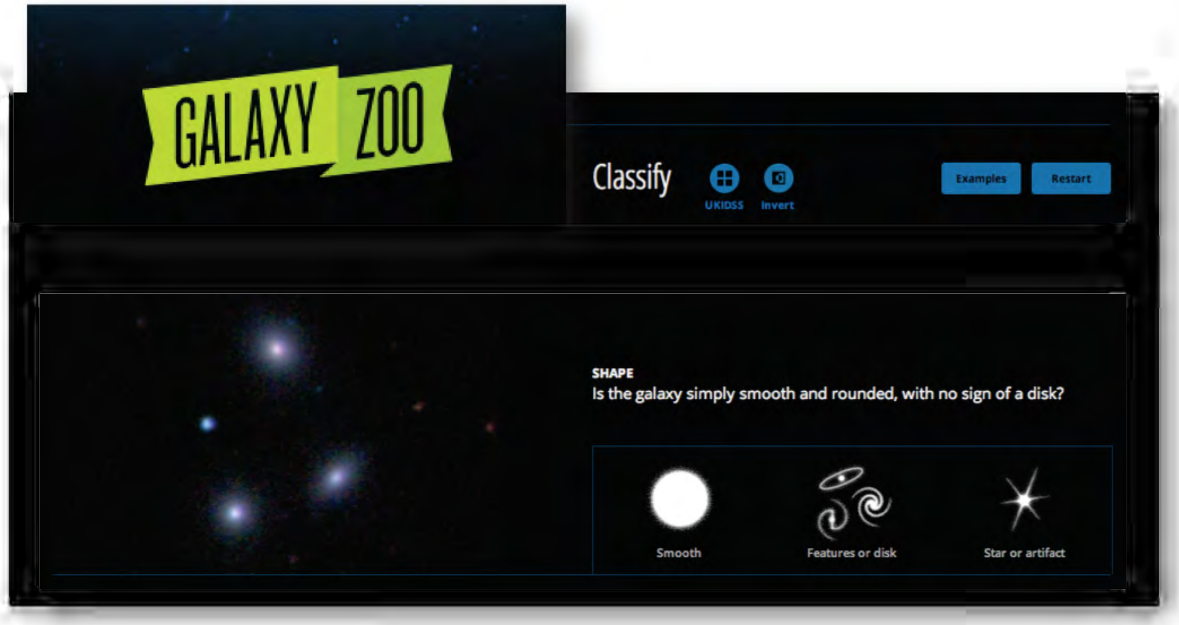


Figure 5. Galaxy Zoo's Online Platform.

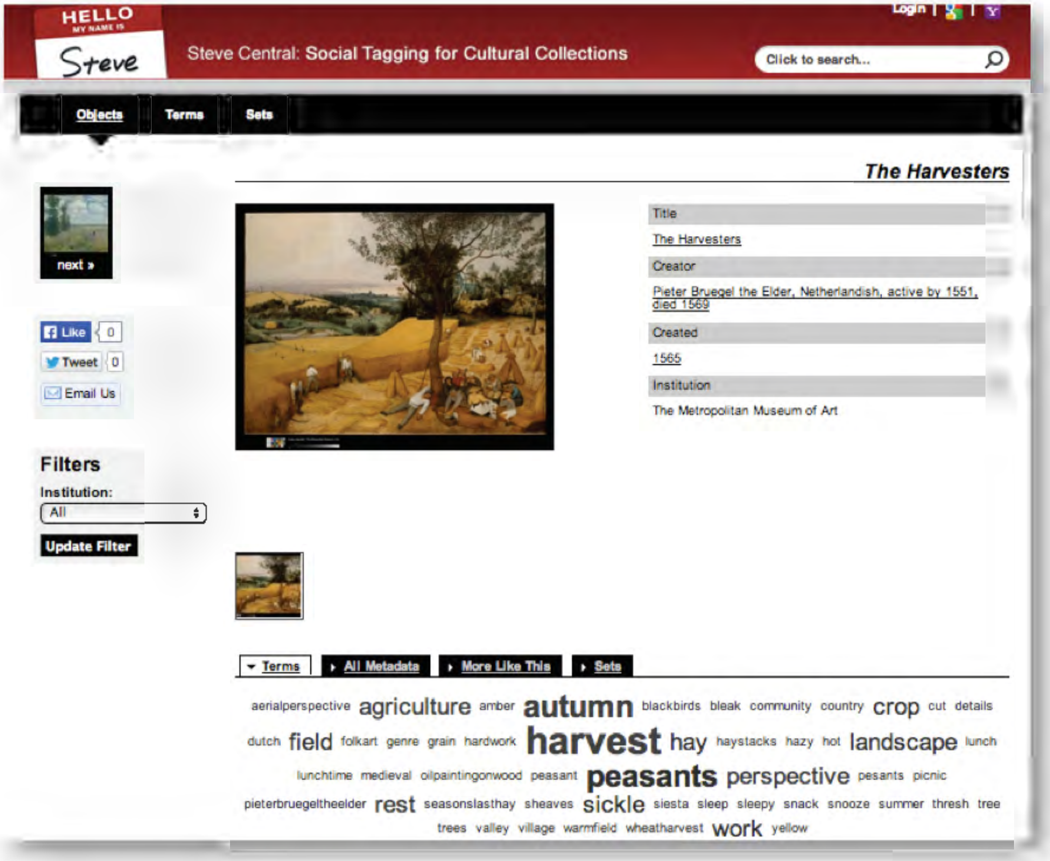
2.5 STEVE.MUSEUM PROJECT

Steve is "*a collaboration of museum professionals and others who believe that social tagging may provide profound new ways to describe and access cultural heritage collections and encourage visitor engagement with collection objects"*[20] by doing research about tagging projects in museums, developing software to gather and manage tags and engaging coomunities in discussions about tagging projects in museums. The project started in 2005 and it was partially funded by the U.S. Institute of Museum and Library Sciences and had collaborators in different institutions, such as New Media Consortium, University of Maryland, the Indianapolis Museum of Art, the Walket Art Center, among others.

[20] Steve Project website http://www.steve.museum/

Their research focused on institutional affiliation, tagging motivations on-site and online, folksonomies and usefulness of tags for institutions. The biggest projects were *Steve in Action* and *T3: Text, Tag, Trust*. They developed an online tool called *Steve Tagger* which allows people to help museums describe their collections using tags online. They also developed the *Steve Software*, which is an open source downloadable version of the application.

Unfortunately the project ended in 2011, but it's worth mentioning since it was one of the first research efforts to study tagging systems for cultural institutions. This, of course, is not a crowdsourcing project, but it's certainly relevant for the topic of this thesis, especially in the development of downloadable open tools.



The Harvesters

Title
The Harvesters

Creator
Pieter Bruegel the Elder, Netherlandish, active by 1551, died 1569

Created
1565

Institution
The Metropolitan Museum of Art

aerialperspective agriculture amber autumn blackbirds bleak community country crop cut details dutch field folkart genre grain hardwork harvest hay haystacks hazy hot landscape lunch lunchtime medieval oilpaintingonwood peasant peasants perspective pesants picnic pieterbruegeltheelder rest seasonslasthay sheaves sickle siesta sleep sleepy snack snooze summer thresh tree trees valley village warmfield wheatharvest work yellow

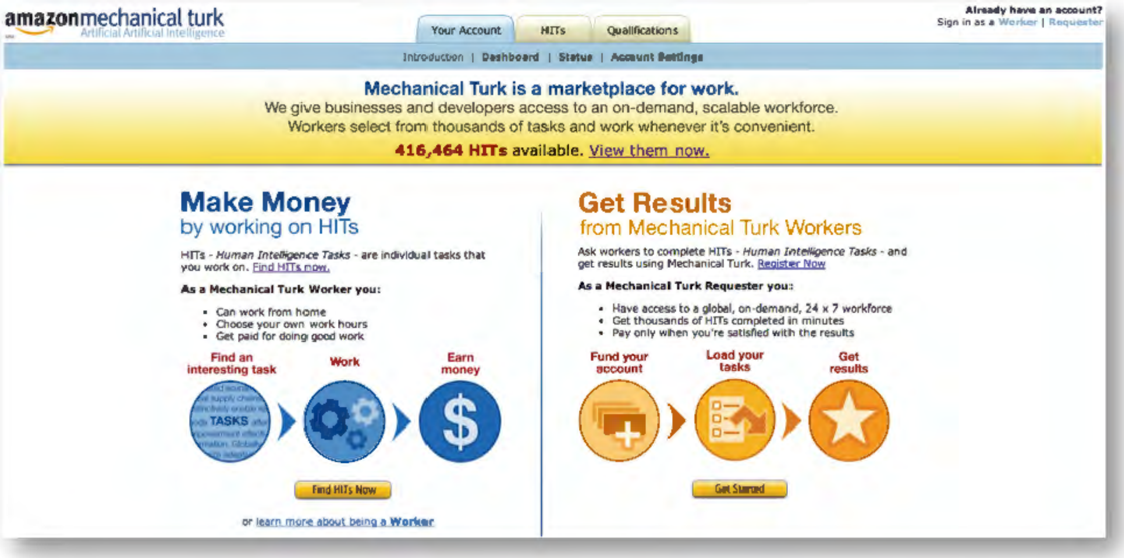
Figure 6. Steve Tagger*, Online Platform. Collections can be accessed online.*

2.6 MECHANICAL TURK[21]

Mechanical Turk is an online platform owned by Amazon, which allows companies (requesters) to outsource discrete tasks (human intelligence tasks, HIT) to be completed by the crowd (workers). Mechanical Turk is interesting because it is a crowdsourcing project that employs people instead of using an open call for volunteers. It is also interesting to mention because it divides the work in small tasks that can be performed in short periods of time and where workers can choose from a long list of very different topics. This raises many questions about the organization of the crowdsourced work in single tasks as well as about the effectiveness of paid workers in crowdsourcing projects versus volunteers, which I'll address in Chapter four.

As I'll explain later, this platform was used by Tagasauris (see below) and the Museum of the City of New York as a user interface for their photo-tagging project. Tagasauris still uses Mechanical Turk for some of the tasks they offer on their website.

[21] Amazon's Mechanical Turk Website: https://www.mturk.com/mturk/welcome



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.
416,464 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

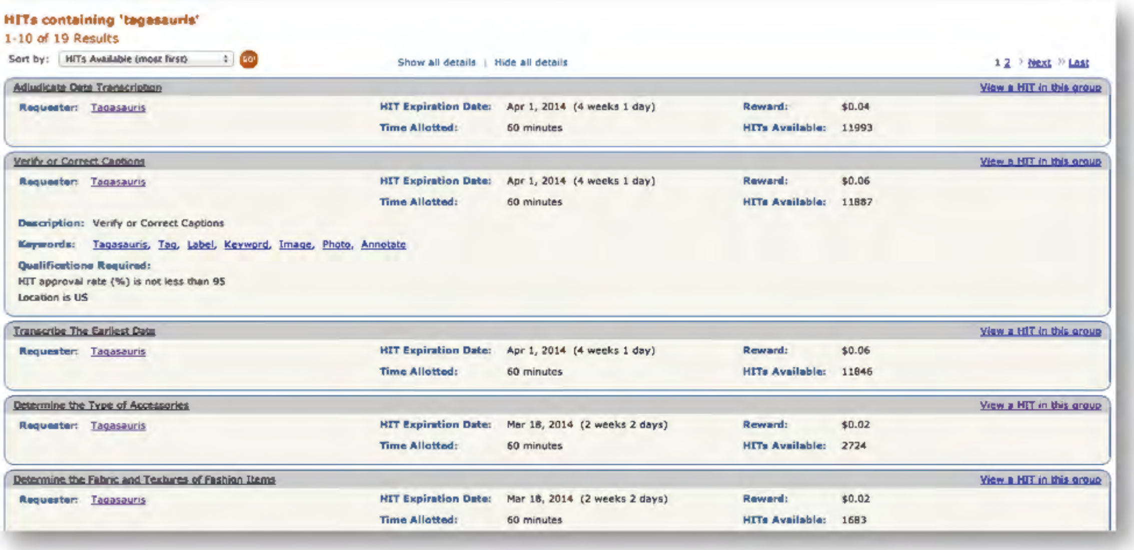
Get Results
from Mechanical Turk Workers

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

Figure 7. Mechanical Turk's Website


Sort by: HITs Available (most first)
GO!
Show all details | Hide all details
1 2 › Next » Last

HITs containing 'tagasauris'
1-10 of 19 Results

Adjudicate Date Transcription — View a HIT in this group
Requester: Tagasauris
HIT Expiration Date: Apr 1, 2014 (4 weeks 1 day)
Time Allotted: 60 minutes
Reward: $0.04
HITs Available: 11993

Verify or Correct Captions — View a HIT in this group
Requester: Tagasauris
HIT Expiration Date: Apr 1, 2014 (4 weeks 1 day)
Time Allotted: 60 minutes
Reward: $0.06
HITs Available: 11887
Description: Verify or Correct Captions
Keywords: Tagasauris, Tag, Label, Keyword, Image, Photo, Annotate
Qualifications Required:
HIT approval rate (%) is not less than 95
Location is US

Transcribe The Earliest Date — View a HIT in this group
Requester: Tagasauris
HIT Expiration Date: Apr 1, 2014 (4 weeks 1 day)
Time Allotted: 60 minutes
Reward: $0.06
HITs Available: 11846

Determine the Type of Accessories — View a HIT in this group
Requester: Tagasauris
HIT Expiration Date: Mar 18, 2014 (2 weeks 2 days)
Time Allotted: 60 minutes
Reward: $0.02
HITs Available: 2724

Determine the Fabric and Textures of Fashion Items — View a HIT in this group
Requester: Tagasauris
HIT Expiration Date: Mar 18, 2014 (2 weeks 2 days)
Time Allotted: 60 minutes
Reward: $0.02
HITs Available: 1683

Figure 8. Tagasauris' HITs on Amazon Mechanical Turk

2.7 TAGASAURIS[22]

Tagasauris is a New York based company founded in 2010. They created an online human assisted computing platform to tag, categorize, annotate, label and organize media. It allows including human and collective intelligence to solve problems that computers can't solve by themselves. They believe in the need of several layers of information and descriptions that will allow media to be more discoverable online. They offer these services for both individuals (using their online platform) and enterprises (who have the possibility of using their Application Programming Interface, API, to integrate these services in their own platforms).

Tagasauris collaborated with the Museum of the City of New York in their tagging project (presented in Chapter six) by providing the tagging platform through the use of Mechanical Turk. All Tagasauris' projects include technologies based on the Semantic Web model. Currently the company is researching the possibilities of expanding their services to time-based media through the use of the W3W Media Fragments URI Standard.[23]

[22] Tagasauris' Website http://www.tagasauris.com/

[23] The W3W Media Fragments URI Standard is a series of syntax specifications for the construction of media fragment URIs and their use in the HTTP protocol. "Media Fragments URI 1.0 (Basic)", W3C Recommendation, Sept. 2012 http://www.w3.org/TR/media-frags/ Interview with Todd Carter, January 7, 2014.

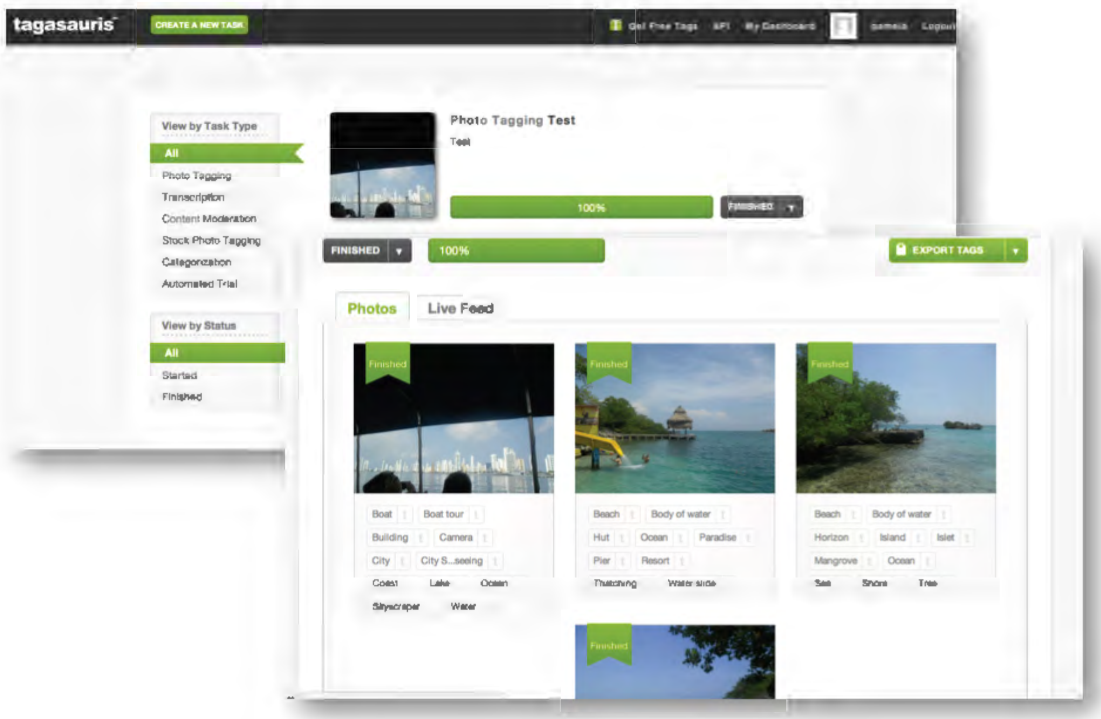


Figure 9. Tagasauris' Online Platform for photo tagging.

3. CROWDSOURCING AND LINKED DATA: Basics

3.1 CROWDSOURCING: History and Definitions

As I previously mentioned, the highly interactive and participative Web 2.0 has become a hub for the development of open projects with the participation of the community since the beginning of the 21st century. Ironically this change was not the consequence of the evolution of the tool itself, since codes, protocols and standards did not evolve tremendously during this transformative stage. Rather, the change came from the way in which Internet was being used to present the content, from our intrinsically social behavior, and from the change in the sources of information, i.e. common users were able to publish information and develop new online platforms and resources. Many platforms and websites with these characteristics appeared during this time. This new way of using the Internet is known as Social Web or Web 2.0. In other words, our desire to connect with each other, together with the decentralization of the information sources allowed this change.

Certainly many technological and socio-economic factors were involved in the evolution of the platform. Internet connections were getting faster and many more people had access to a computer connected to the web. Moreover, the launching of social websites such as Facebook, Flickr, YouTube, Vimeo, etc. have helped people realizing the benefits of the web beyond the idea of only accessing information.

3.2 CROWDSOURCING: Looking for a Definition

The exponential growth, broad scope and variety of collaborative projects online during the past decade have made it difficult to stop, observe and evaluate the phenomenon. Defining models and even classifying online activities is not easy due to its constant evolution. Crowdsourcing is not the exception.

Additionally to this fast paced growth and proliferation, many of the models implemented online - because of their impact, scope and technological features – require an interdisciplinary team to study them, and regular communication among those disciplines can be very cumbersome. In the case of crowdsourcing, several analyses have emerged from very different areas of study, such as computing (the most prolific one thanks to their need to produce application tools), social sciences, ethics and business. In the words of Brabham:

> "*The empirical research on crowdsourcing is untidy because it is developing within various disciplinary silos that are not in conversation with one another. And when untidy scholarly discourses mix with arbitrary popular media usage about crowdsourcing, the result is unkempt theory and practical crowdsourcing applications with shaky foundations*."[24]

Regardless of the lack of substantial information and the apparent variety of opinions, cultural institutions have already been experimenting with the crowdsourcing model having

[24] Daren C. Brabham, "Crowdsourcing" (Cambridge, Massachusetts ; London, England : The MIT Press, 2013), Preface.

different results[25]. However, in the case of library sciences the study of crowdsourcing has been very directed towards the application of the model in conjunction with the technical sciences – which have helped improving these applications to a certain extent - but not many institutions or members of the community have stopped to evaluate the effectiveness and benefits of the model for the field. There are some quantitative studies and some hints of study of the social impact of the use of crowdsourcing, but not a substantial body of knowledge around it.

Nonetheless, the use of this model by cultural institutions has grown in the past 5 or 6 years. Many studies point out that non-profit organizations have a special advantage based on the way community projects are normally built. According to Rose Holley, the Manager of the Trove project in Australia, "*Volunteers are much more likely to help non-profit making organisations than commercial companies, because they do not want to feel that their work can be commercially exploited.*"[26]

My research on crowdsourcing started more than a year ago. One of the first goals I had was to find a clear definition of what crowdsourcing was. Because of the reasons I just explained, that goal was very hard to accomplish. Even more, one year later I found that the problem of finding a definition still remains. However, there is a clear history of the evolution of those many definitions. I will not attempt to explain them all, but rather, to highlight the ones that have prevailed or influenced others.

[25] Some examples of cultural institutions applying the model: New York Public Library "What's on the Menu?" (http://menus.nypl.org/), Brooklyn Museum "Tag! You're It!" (http://www.brooklynmuseum.org/opencollection/tag_game/start.php) and "Click! A Crowd-Curated Exhibition" (https://www.brooklynmuseum.org/exhibitions/click/), The National Archives UK "Operation War Diary" (http://www.operationwardiary.org/).

[26] Rose Holley, "Crowdsourcing: How and Why Should Libraries Do It?" on *D-Lib Magazine*, March/April 2010, Volume 16, Number 3/4
http://www.dlib.org/dlib/march10/holley/03holley.print.html

Crowdsourcing was a term coined in 2006 by Jeff Howe, to describe the sharing website Flickr.[27] The first definition of crowdsourcing appeared in 2006 on Wire Magazine in an article written by Jeff Howe titled "The Rise of Crowdsourcing":

> *"Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers."*[28]

Many other definitions followed, such as the one presented by Daren C. Brabham, also in 2006, which encompasses the broadness of crowdsourcing: *"Crowdsourcing is an online, distributed problem-solving and production model."*[29]

Howe's definition has been declared insufficient in many cases because it uses the word "outsourcing" which is a slightly different concept[30]. Moreover, the definition lacks the very spirit of what crowdsourcing represents, the idea of community. Nevertheless, this definition is historically important since it was the first attempt of theorizing this online phenomenon. On the other hand, Brabham's idea of calling it a "model" fits better than using the word "act", which explains what crowdsourcing is today, a collaboration model based on community collaboration.

[27] Flickr website: http://www.flickr.com/about/

[28] Howe, Jeff, "Crowdsourcing: A Definition", 2006, http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html

[29] Brabham, Daren C., "Crowdsourcing as a Model for Problem-Solving: An Introduction and Cases", 2006, http://www.clickadvisor.com/downloads/Brabham_Crowdsourcing_Problem_Solving.pdf

[30] Outsourcing, a term used widely in business models, is the contracting of an internal process to a third-party organization.

Brabham dedicated his doctoral work to the evaluation of the model and, as many of us, he noticed the need for a clear definition. In his book, *Crowdsourcing*, he explains how he encountered many different definitions coming from different disciplines and with different approaches. However, he highlights the contribution made by Enrique Estellés-Arolas and Fernando González-Ladrón-de-Guevara, who surveyed and evaluated more than forty different definitions of crowdsourcing to finally provide a unified one:

> *"Crowdsourcing is a type of participative online activity in which an individual, an institution, a non profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken."*[31]

In the presence of this comprehensive definition, Brabham adds that in crowdsourcing projects "the locus of control of the creation of goods and ideas" must reside between the organization and the users. As I will address below in the next chapter, institutions must be

[31] Enrique Estellés-Arolas and Fernando González-Ladrón-de-Guevara, "Towards and Integrated Crowdsourcing Definition" on the *Journal of Information Science (JIS)*, Volume XX, pp. 1-14 http://jis.sagepub.com/content/38/2/189.abstract

aware of the characteristics of the model before its application, being "control and guidelines" one of them.

Going back to Estellés-Arola and González-Ladrón-de-Guevara's study, it is worth noting that they produced an eight-step evaluation for the compliance of projects with the new definition. This list can be useful to determine if we are in the presence of a crowdsourcing project, which would ultimately help describing the scope and final goals of the endeavor:

(a) There is a clearly defined crowd

(b) There exists a task with a clear goal

(c) The recompense received by the crowd is clear

(d) The crowdsourcer is clearly identified

(e) The compensation to be received by the crowdsourcer is clearly defined

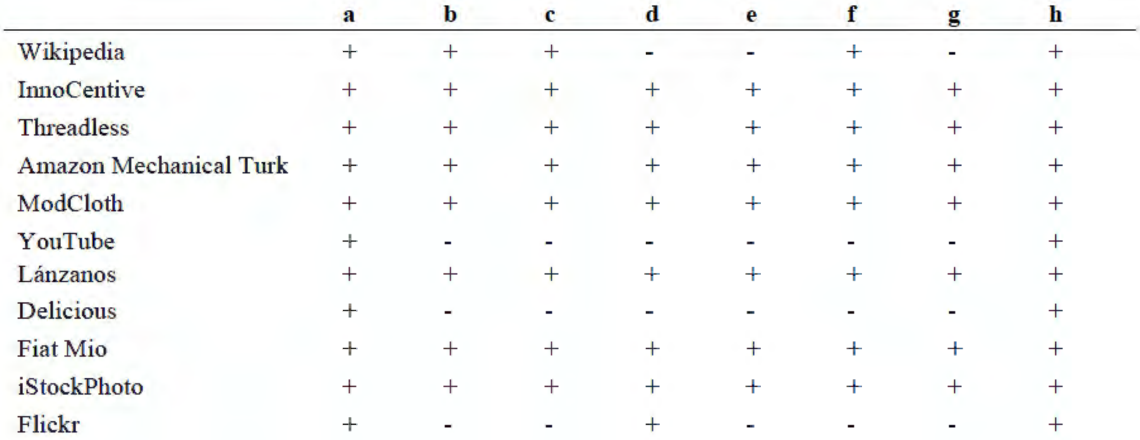
(f) It is an online assigned process of participative type

(g) It uses an open call of variable extent

(h) It uses the Internet

This evaluation helps defining the previously blurry lines of crowdsourcing. Under this microscope, we can clearly say that many projects that have been called crowdsourcing in the past, lack one or more of these eight conditions, among them open source initiatives, common-based peer production, market research and brand engagement, and crowdfunding.

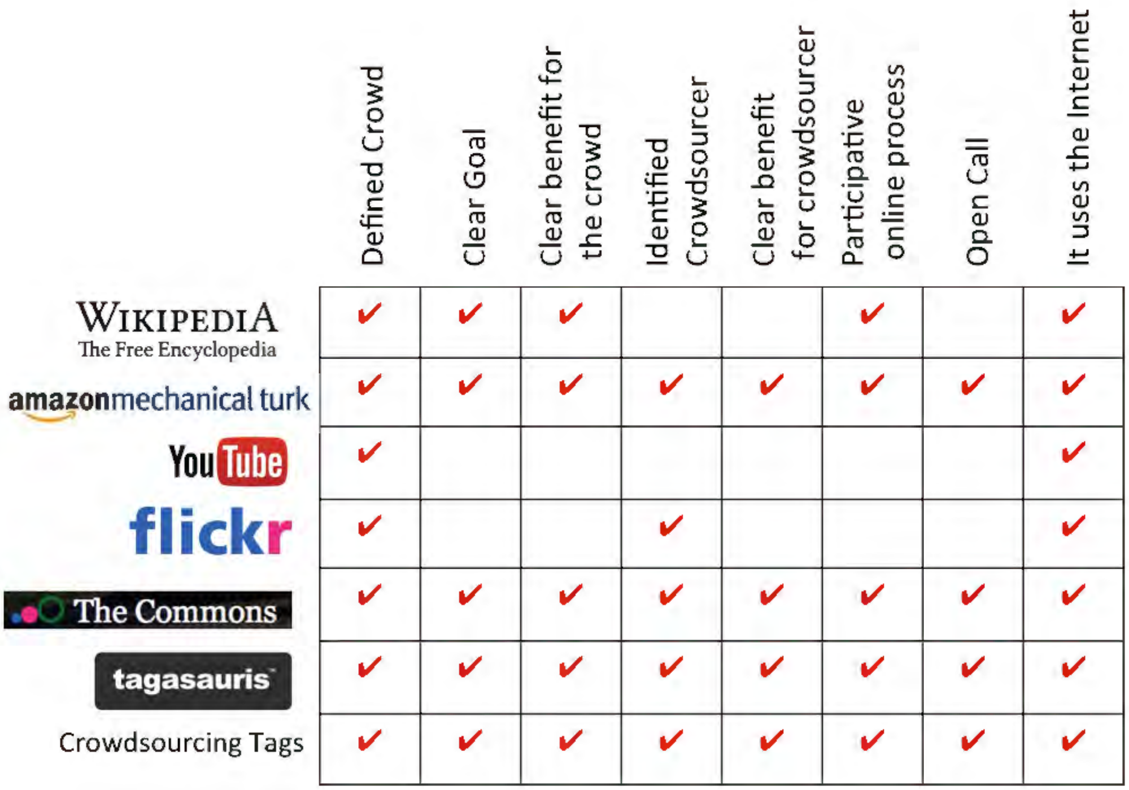
As part of this study, Estellés-Arola and González-Ladrón-de-Guevara also evaluated several online projects and compared them to the new definition. The following table shows the final results:



	a	b	c	d	e	f	g	h
Wikipedia	+	+	+	-	-	+	-	+
InnoCentive	+	+	+	+	+	+	+	+
Threadless	+	+	+	+	+	+	+	+
Amazon Mechanical Turk	+	+	+	+	+	+	+	+
ModCloth	+	+	+	+	+	+	+	+
YouTube	+	-	-	-	-	-	-	+
Lánzanos	+	+	+	+	+	+	+	+
Delicious	+	-	-	-	-	-	-	+
Fiat Mio	+	+	+	+	+	+	+	+
iStockPhoto	+	+	+	+	+	+	+	+
Flickr	+	-	-	+	-	-	-	+

Figure 10. Evaluation of 11 online projects according to Estellés-Arola and González-Ladrón-de-Guevara's definition of crowdsourcing. Letters (a) through (h) correspond to the list provided above.

On this table we can see that some projects that were born and widely known as crowdsourcing do not comply with the requirements of this definition. It is interesting to note, for example, that Wikipedia and Flickr are in this group. On one hand, Wikipedia functions more as an open source model than as a crowdsourcing one. Additionally, as Brabham argues, in Wikipedia the balance of the locus of control is greatly inclined towards the users who control most of the output of the site, and the organization itself only provides the platform that allows this interaction. On the other hand, while Flickr can not be considered as a crowdsourcing project by itself, it could be - if used adequately and in addition to other platforms – a starting point for projects of this nature. For instance, *The Commons,* the crowdsourcing project started by the Library of Congress, used Flickr as a platform. It provided a task with a clear goal (tagging photos of public photo collections), a recompense for the crowd (allowing better access to the collections), a compensation for the organization (complying with their mission statement of providing access to collections), it was an online participative process and it used an open call (all members of the community could contribute). In the same vein, *The Trove, Galaxy Zoo* and the transcription projects all fit with this definition (see Figure 11 below).



	Defined Crowd	Clear Goal	Clear benefit for the crowd	Identified Crowdsourcer	Clear benefit for crowdsourcer	Participative online process	Open Call	It uses the Internet
Wikipedia The Free Encyclopedia	✓	✓	✓			✓		✓
amazonmechanical turk	✓	✓	✓	✓	✓	✓	✓	✓
YouTube	✓							✓
flickr	✓			✓				✓
The Commons	✓	✓	✓	✓	✓	✓	✓	✓
tagasauris	✓	✓	✓	✓	✓	✓	✓	✓
Crowdsourcing Tags	✓	✓	✓	✓	✓	✓	✓	✓

Figure 11. Classification of some of the crowdsourcing projects presented in Chapter two according to the definition and table developed by Estellés-Arola and González-Ladrón-de-Guevara.

While not all cultural institutions apply the crowdsourcing model in the same way, many of the projects have some similarities that are worth mentioning. First of all, most organizations have the mission of providing access to collections in some way. They have a responsibility with the community not only in those terms, but also in creating participative communities. Crowdsourcing projects offer a unique opportunity to make this happen. Second, most crowdsourcing projects in cultural institutions are related to either describing collections (gathering metadata, mostly social metadata) or enlarging collections in areas or subjects in which the institution has a particular interest (gathering new materials). To do so, it is a priority to have some creative or intellectual input from the community. Generally speaking, the crowdsourcing model fits cultural institutions and can be very beneficial when their goals and implementations are well defined.

I would like to highlight the term I used in the previous paragraph, referring to the type of metadata that institutions collect during the application of crowdsourcing projects: social metadata. According to the recent survey conducted by the Library of Congress about this practice, social metadata is *"Additional information about a resource resulting from user contributions and online activity —such as tagging, comments, reviews, images, videos, ratings, recommendations —that helps people find, understand, or evaluate the content."*[32] Thus, social metadata is the information about an element gathered through the use of the crowdsourcing model. I think that keeping in mind this definition is very important, especially because I believe it will have a huge impact in how this type of information is integrated to the current metadata systems.

Now, focusing only in projects that use crowdsourcing to enhance content description for better access, which represent the most challenging part of crowdsourcing for digital collections; they mostly use the process of tagging (although some of them include comments as well). According to Breslin et al. tags are *"A keyword annotation that acts like a subject or category*

[32] Karen Smith-Yoshimura, Cyndi Shein.

for the associated content."[33] Initially, tags were used in social media to annotate the content of the elements people wanted to share, in order to allow other users to have easy access to them. Tags are normally *free-form keywords,* ideally of only one word, although some users add symbols such as "." or "_" between words to use more complex concepts.[34] The schematic visualization of a group of tags, for example in one particular website, is called *tag cloud,* which not only allows to see all the tags used, but it also highlights the most popular terms by making them bigger or bolder.

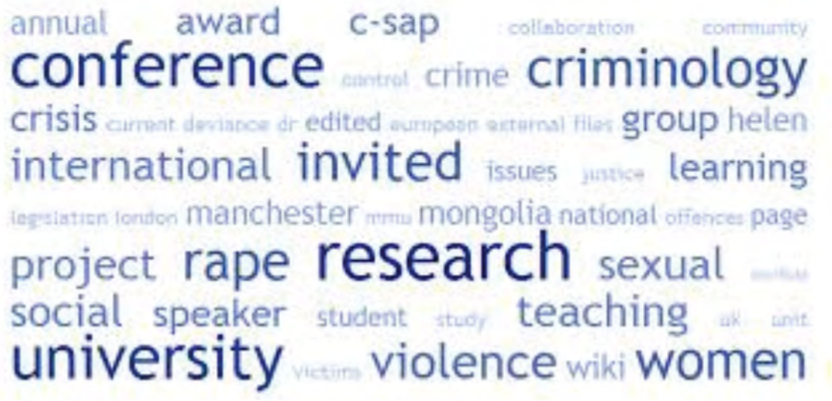

annual award c-sap collaboration community
conference control crime criminology
crisis current deviance dr edited european external files group helen
international invited issues justice learning
legislation london manchester mmu mongolia national offences page
project rape research sexual
social speaker student study teaching uk unit
university victims violence wiki women

Figure 12. Example of a tag cloud,

http://criminology.pbworks.com/w/page/12518021/Tag%20Clouds

One of the advantages of using tags is that normally these words are searchable, when kept in the same platform they were created, allowing easy access to content, even allowing accessing a list of content tagged with the same term. Another advantage of using tags, as

[33] John G. Breslin, Alexandre Passant, Stefan Decker, *The Social Semantic Web* (Heidelberg ; New York : Springer, 2009)
[34] Golder and Huberman present a classification of tags, divided in seven types. Scott A. Golder, Bernardo A. Huberman, *The Structure of Tagging Systems*, http://arxiv.org/pdf/cs.dl/0508082.pdf?origin=publication_detail

described in *The Social Semantic Web,* is that anybody can annotate content and not previous knowledge in standards or subject structures are needed. *"Users can use exactly the words they deem necessary without limitations."* But, in the cultural world and more specifically in the archival world, this advantage turns into a disadvantage when it comes to the integration of tags to existing metadata schemas.

Under this particular environment, folksonomies arise. Folksonomy, a term coined by Vander Wal in 2007 is a *"social collaboratively generated, open-ended, evolving and user-driven labeling systems that enable users of social websites to categorize their content using the tag system and to thereby visualize popular tags usages via tag clouds."*[35] The value of folksonomies not only resides in the fact that anyone can use them, but also in that they represent a certain community, group of people, association, cultural, political and social scenarios that describe a particular place and point in time.

3.3 SEMANTIC WEB: History and Definitions

The idea of the Semantic Web, contrary to what many might think, it is not new and did not begin with the Social Web. In fact is as old as the Web. The first idea of the web, created by Tim Berners-Lee already included concepts related to linking data beyond the use of only URLs. However, the Semantic Web as a project was first started by the World Wide Web Consortium to transform the "web of content and documents" in the "web of data". But why? What is the problem with the current system? After all, we have all enjoyed the benefits so far. What is it lacking?

[35] John G. Breslin, Alexandre Passant, Stefan Decker, pp. 140.

Up until now, websites are a set of documents stored on a server and displayed in a certain way, following certain standards and protocols (HTTP and HTML). Content can also be linked to other websites using the Uniform Resource Locators (URL), i.e. the particular directory in the server where the documents are stored. That's the web of content. In spite of allowing connecting websites and content, these are not interoperable and linking information is quite difficult. In other words, the web of content doesn't allow machines to understand the language in order to improve the use of the web. Machines know that a certain URL connects to a certain website, but what is inside the websites falls out of the capabilities of computers. Some standards have been created to solve some of these issues, such as RSS, FOAF or SIOC[36], but the problem requires more than standards, it requires rethinking the way information is published and shared.

The Semantic Web allows *"large scale integration of, and reasoning on, data on the Web"*[37] meaning that systems can now establish relationships between data that were not possible to establish before. Put in simple words, the Semantic Web is a way of connecting, sharing and reusing data, so that it is understandable for machines and humans. The method or model used to make this possible is called Linked Data, the name Tim Berners-Lee gave to this idea in 2006. To make this possible, the web must have the following characteristics[38]:

[36] Rich Site Summary (RSS) is an xml-based format used mainly to distribute news headlines online, in other words, provide more information about the content of the website.
Friend of a Friend (FOAF) is an ontology that provides relationships between people and their activities.
Semantically-Interlinked Online Communities (SIOC) is a project that provides links between discussion platforms such as blogs, forums, etc.

[37] W3C, *What is Linked Data?*, http://www.w3.org/standards/semanticweb/data

[38] Breslin et al., "The needs for semantics" in *The Social Semantic Web.*

(a) Entity Identity: each element, subject or object (see below) must have a unique identifier. In the Semantic Web, this identifiers are called Unique Resource Identifiers (URI).

(b) Relationships: to be able to link content, the relationship between entities must be established.

(c) Extensibility (flexibility and adaptability): to be able to share and reuse information, systems must be compatible.

(d) Vocabularies (Ontologies): having basic vocabularies is key in the way data is structured to allow complex relationships. In the case of the Semantic Web, vocabularies can be combined, since they all use the same standardized sharing model: Resource Description Framework (RDF).

But, why use Linked Data? What are the benefits? Having structured and linked data, in general, can help making content easily discoverable. It also allows computers to answer more complex queries, since the use of structured vocabularies allows establishing relational and hierarchical links between data. Data is reusable, since it is shared in a widely adopted standard. For cultural institutions, Linked Data is a way of using other people's or other organization's data, saving time and resources. Moreover, since it provides a clear structure, it can be considered an equivalent of controlled vocabularies and authority lists.

To make this possible data must be structured and linked in a very particular way. Linked Data's basic conceptual structures are called statements or triples. The structure of a triple is:

subject – predicate – object,

where the subject is the element we want to describe, the object is what we want to say about it and the predicate is the relationship between both. Objects can also be the subject of another triple and subjects can be linked to many other objects using other predicates. This is the fundamental principle of Linked Data, something like a huge relational database.

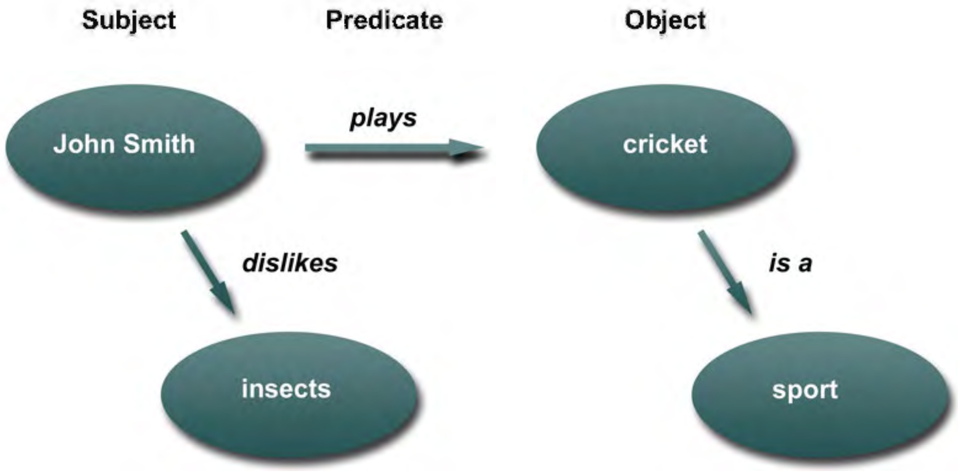


Figure 13. Example of triples, subjects, predicates and objects.

http://www.ansta.co.uk/blog/semantic-web-technologies-part-3-94/

In Figure 13 we can se an example of how data is linked. In this specific case we have three triples: “John Smith plays cricket”, “John Smith dislikes insects” and “Cricket is a sport”. Note that subjects can also be objects of other triples. As you can see, endless relationships can be drawn from these structures. As mentioned before, each of these entities (subjects, objects and predicates) has an associated URI, thus, every time we need to link to it, we can point to the exact resource. In other words, we don’t need to define *John Smith* every time we refer to him, because he’s already in the “database” (in Semantic Web datasets), we just need to point to it using the correct predicate. This is extremely important for the model, since it's the base of shared data, i.e. if I link my data to a public dataset I am using all the entities and relationships in it, aggregating much more value to my own data but also saving time and ultimately resources.

Now, how are things linked? Each element in a triple has a URI. URIs used in a particular data set can also come from other data sets. For example, if a library wants to establish the following triple:

Jane Eyre – written by – Charlotte Brönte,

they can use the URIs for each of the elements in the triple provided by Library of Congress, without inventing a new one, meeting one of the objectives of Linked Data, sharing and reusing. Another benefit of URIs is having a unique identifier for that particular subject, object or predicate that differentiates it from any other subject, providing disambiguation. For instance, I can have two different subjects named Harry Potter, but one referring to the book and the other referring to the movie.

Now, all these things are possible if we have access to the datasets. That is the main difference between Linked Data and Linked Open Data. According to Tim Berners-Lee, Linked Open Data is "*Linked Data which is released under an open license, which does not impede its reuse for free."* Mr. Berners-Lee also developed a 5-Star rating system, to encourage people and institutions to exchange and reuse data. Any institution can make their datasets available for free use, however, true Linked Open Data must be linked to other people's data to provide context.[39]

[39] Tim Berners-Lee, *Is Your Linked Open Data 5 Star?*, 2009
http://www.w3.org/DesignIssues/LinkedData.html

This aspect is key for the success of the model; everyone uses everyone else's data, avoiding redundancy.[40]

[40] More details about Linked Open Data in Appendix 1.

4. UNDERSTANDING THE SOCIAL DIGITAL CULTURE: Communities' online behavior

This new scenario, enabled by the particular characteristics of the Web, where the origin of production of content has shifted from institutions – whether commercial or not – to users and where connectedness has reached high levels, we can't help but wonder what will be the impact, on our society, our communities and our institutions (and by extension our roles as professional inside these organizations).

The impact and effects of crowdsourcing can't be studied in isolation, since the model is only one type of implementation enabled by the social web. To understand crowdsourcing we must be capable of having a broader view of the phenomenon and study instead the characteristics of social media. However, when approaching this fascinating subject there are some things we must understand that go far beyond technology and some particular tools or implementations. The origin of social media has its roots in social behavior; this doesn't mean that technology is to be disregarded (in fact, it will be addressed in the next chapter). But, having a clear understanding of the reasons why people participate in these projects will help us evaluate and plan future tools in a more efficient way.

This area of study, though it will help understand crowdsourcing better, can be applied to many other online tools and initiatives, since it tries to explain the foundations of our behavior online, which at the same time has its roots on our human behavior which is by default social.

This chapter will explain the social side of digital media, analyzing the basic characteristics that make the social web what it is from the individual point of view (personal motivations) to group behavior (membership and online groups). I will also explore the implications of such behavior for institutions and organizations, from the market and also from the non-profit point of view.

4.1 CLAY SHIRKY'S THEORY: Means, Motive and Opportunity[41]

As old as the web and Internet are, finding a clear and definitive explanation of the sources of our highly interactive response as a society to these new models is not easy. I believe mostly because, as I explained before, its origins are in our social behavior, which is highly complex, influenced by many factors and it is permanently changing. Different theories approach the issue in different ways and at different degrees of depth.

As I initially started my research, I found some interesting approaches – although contradictory - coming from members of institutions and organization. For example, studies behind the tagging game project called *Waisda?* created by the Netherlands Institute for Sound and Vision (which will be discussed and explained in more detail in Chapter 6) showed that altruism was an important motivation to contribute with tags.[42] However, other sources indicate that there is certain degree of selfishness when participating in these projects. For instance, Dean Stringer from the Waikato Centre for eLearning says in his notes after attending a workshop about design of creative communities "*Derek [Powasek] noted that people tend to participate for selfish reasons, but that this can be good in a wisdom of the crowds fashion. For example, people don't create hyperlinks for altruistic reasons, but when aggregated the links support Google's*

[41] Clay Shirky, *Cognitive Surplus* (New York: Penguin Press, 2010)

[42] Netherlands Project, Various Authors, "Emerging Practices in the Cultural Heritage Domain, Social Tagging of Audiovisual Heritage" in *Proceedings of the WebSci 10: Extending the Frontiers of Society On-Line (*Raleigh, NC: US April 26-27th, 2010) http://journal.webscience.org/337/2/websci10_submission_23.pdf

pagerank algorithm. Likewise, people tag in Flickr for 'selfish' reasons, but when aggregated these tags become a powerful tool."[43]

But to find a more complete study of human behavior in the digital world I had to go beyond the world of cultural institutions: cultural studies, mass media and communications. Clay Shirky, professor at the Interactive Telecommunications Program (ITP) at NYU, in his book *Cognitive Surplus* illustrates a very complete explanation of why and how social media is what it is and how much of that success can be attributed to the advance in technology and human social behavior.

Analyzing media and human behavior around it from the invention of printing press by Gutenberg through the culture of television and now Internet, Shirky's theory focuses on how technologies enabled social media (and social digital culture) to appear, thanks to our intrinsic social behavior, rather than giving technology all the credit. In his book he also explains that behavior and how tools can be optimized if their creators acknowledge the importance of such conduct. The three factors that enabled growing online communities are means, motive and opportunity.

In his study of human behavior, a large portion of his work focuses on *cognitive surplus*, which is, put in very simple words, people's free time. For the past fifty years we have been using that free time consuming media (from television to paper publications), which limited our social activities. However, as Shirky says: "*…people like to consume, but they also like to produce, and to share."* Such dormant conduct has been awakened by new media, which not

[43] Dean Stringer wrote a blog publication after attending the workshop *Sustaining Creative Communities* taught by Heather Champ and Derek Powazek at the Webstock Conference 2009 in New Zealand. http://librarytechnz.natlib.govt.nz/2009/03/designing-sustaining-creative.html

only allows people to communicate, but to create, share and mold the tools they use, changing the locus of production from organizations and companies to common people.

4.1.1 Means

One of the main explanations of the growth and importance of social media, according to this theory, is the availability of *means* allowing social behavior and cognitive surplus to manifest, and that means was the Internet. Cognitive surplus and social behavior were there before this new technology, but they only came to life thanks to this tool that facilitated connectivity and the use of free time in the most varied activities. The Internet is so big and so diverse that gives a chance to everyone to produce and share content and to connect with people who have the same interests. As important as technology is, it is also important allowing this behavior to occur; in that sense Shirky highlights that *"Human character is the essential component of our sociable and generous behaviors… technology enables those behaviors, but it doesn't cause them."* This also coincides with Brabham's theory of the Internet as a participatory culture, where crowdsourcing can only exist online, since it enables scalability and crowds' social behavior.

Thus, one of the most important characteristics of Internet that enabled such conduct was scale, which would be the equivalent to what Brabham defines as *reach*, mentioned previously in Section 1.1. Previous to the emergence of the Internet, people would use their cognitive surplus in small-scale activities, whether they involved social communities or not: watching TV, going to church, repairing a car, etc. Now people can not only use this surplus to do such things, but also to connect with other people who like watching TV or repairing cars, allowing them to share information, opinions and help each other. An originally local community can now be a huge community online, which reaches people from all over the world. This is no longer a group of people; this is a *crowd.* The advantage of such large scale is that "…*with a large enough crowd, unpredictable events become predictable.*" Thus, it is much more likely to find people with the same interest or needs, and at the same time finding more people willing to help and give their free time to a particular project. Now the question is, why would people be willing to give away their free time?

4.1.2 Motive

As immersed as we are in a highly consumerist and market-driven culture, it is very hard for us to believe that people would be willing to give away their free time. Why would anyone do anything for someone (or some organization) without getting paid or at least without a reward? The answer to that question is very simple: because we like spending time doing what interest us. We've always done that, but not at this large scale.

Now, as I mentioned earlier, there are many theories that explain the phenomenon by attributing crowd's interest to selfish motivations. The truth is, they are not far from reality, although selfishness wouldn't be the most accurate word in my opinion. Edward L. Deci's theory about human motivation explains that humans experience two sources of motivation: extrinsic and intrinsic[44]. Examples of intrinsic or personal ones can be autonomy and competence. Thus, human beings can be intrinsically motivated by activities that promote the development of self-

[44] Edward L. Deci developed the Self-determination theory (SDT) based on the human needs of competence, relatedness and autonomy, as well as human intrinsic and extrinsic motivations. http://www.selfdeterminationtheory.org/

conducted practice and learning. Activities that somehow let the individual do, create and improve his or her skills in a somewhat unconstrained environment. As Shirky and Brabham point out, social media and online interactions create the perfect atmosphere to allow users to explore and to find their own personal motivators. These intrinsic motivations (in the very personal realm) can be reinforced through membership and sharing (in the social realm).

Of course human beings also have external motivations, which not necessarily involve market or commercial transactions. Receiving payment for one's work could be one of them. However, as Deci's study pointed out, these so called extrinsic motivations can affect our intrinsic ones, inhibiting them. I will explain further this in Section 4.5

4.1.3 Opportunity

The third and last component of Shirky's theory is opportunity. It is not enough to have a means and a motive, we – we meaning the crowd but also institutions – have to create the right opportunities for people to exploit their own intrinsic motivations, being these new tools or new platforms. "*Given the right opportunities, humans will start behaving in new ways."* It is precisely because of this that social media is highly changeable. Tools can be created with a very particular purpose, but the final results are given by the crowd; they will be the ones determining how they use what is presented to them.

4.2 MEMBERSHIP AND DIGITAL CULTURE

Thanks to the characteristics of the Internet and the new opportunities it presents for users, our old behavior of pure consumption is rapidly shifting towards a more interactive way of using online tools. The expectations of the "audience" are much more complex every day, as the tools get more and more complex as well. In general terms, people want to participate and feel they belong to a community where what they do or say really matters. However, each community has its own particular needs, not only about what they expect from the tools but what they expect from organizations – either small groups or formal institutions – and other members. This is what Shirky defined as the *culture*. Culture plays a big role in determining the path a certain project will follow and its final results.

However, online social projects are highly fickle due to the complex features of society, both online and outside the web, therefore those expectation are very difficult to predict. Thus, online social projects are certainly a double-edged sword; we can take advantage of the scalability and of the highly committed members of our communities, but if they want to take our project somewhere else they will. This is not completely a negative thing, after all this characteristics is what made social media what it is now, but this implies that there are many issues to think about before starting a crowdsourcing project.

4.3 FEWER RULES, BETTER RESULTS?

While many people would associate online communities with freedom and uncontrolled activity (or in some cases even the exact opposite!), the truth is, having online communities does not mean that rules are absent. The reason why people have such a different opinion about it is because different platforms provide different levels of *moderation*[45]; i.e. the level of control over the use of the platform presented in the form of rules or terms and conditions. Therefore some sites can be very restrictive, or moderate, about the use of the resources presented, and others can be fairly open about it. However, this factor has a huge impact in the way the crowd interacts not only with the site, but with each other. According to the article written by George Oates (creator of *The Commons* on Flickr) in 2008: *"Given fewer rules, people actually behaved in more creative, cooperative, and collaborative (or competitive, as the case may be) ways."*[46]

What Oates noted was that, what actually happens is that communities, when presented with fewer rules, tend to communally create guidelines to moderate the platform. More than rules, these are norms for correct social behavior online. As Shirky noted, in some online communities for example "*Failure to give credit where credit is due is the crime in this community, a violation not of property rights but of deeply held ethical norms about credit."* Punishment for the violation of these norms are even worse than a fine, and it is equivalent to social rejection. This mechanism is what Shirky defines as *self-governance* and what Elinor Ostrom had defined as *joint governance* as early as 1993.[47] This self-governance enhances our idea of membership and it reinforces our intrinsic motivations of autonomy and competence.

[45] According to Smith Yoshimura: "*Sometimes sites are heavily moderated, and others not at all. The moderated sites tend to have fewer contributions than those that are not. Strict credentialing can be a barrier to more broad-based participation."*

[46] Oates, George, *Community: From Little Things, Big Things Grow*, 2008, http://alistapart.com/article/fromlittlethings

[47] Elinor Ostrom was a political economist and leading theorist behind the Commons movement.

4.4 COGNITIVE AUTHORITIES

Certainly the new technologies have disrupted the idea of our historically traditional trustworthy sources of information. Who do we trust when we want to learn about the world? This issue is present in our daily life, from home, schools, newspapers, to cultural institutions, which manage and provide access to certain information; and it is of course also present in any social media project. Do we trust that the photograph about a public event uploaded on Facebook and taken by a common person is real? How is that photograph different from what the New York Times made public on its website? This is a matter of cognitive authorities.

A cognitive authority is a person or entity that influences my thinking in relation to a sphere of interest at different degrees. Cognitive authorities involve relationships between at least two people, since it involves certain levels of credibility and trustworthiness. For instance, for most people libraries represent a cognitive authority, since they keep books, which are historically our primary source of information and sometimes truth. But different libraries may represent different levels or degrees of authorities and cover different spheres of interest or topics For example, to do research on surgical medicine I might want to go to a specialized library, as opposed to my local branch library.

But libraries are not the only place where we go to learn and know the world. According to Patrick Wilson[48], our way of learning about the world is through first-hand and second-hand knowledge. First-hand knowledge involves all our personal experiences, that of course not only relate to formal education. For example, I can have a very specific knowledge about the best

[48] Patrick Wilson, *Second-hand Knowledge: An Inquiry into Cognitive Authority*, Westport, Conn. : Greenwood Press, 1983.

route to go downtown from my neighborhood based on my personal experiences. On the other hand, second-hand knowledge is everything we learn from other sources, whether they are people or institutions. Most of our knowledge comes from these sources, since we can't experience all the things we need or want to know. Therefore, we need to establish votes of trust to determine which people or institutions are trustworthy for us, since, as mentioned above, a cognitive authority is a two-way relationship.

Normally you would trust educational and cultural institutions because most of them have a history of consistent study and transparency, and because they are supported and/or formed by professionals specialized in whatever they do. However, social media has challenged that way of thinking. In some situations we do trust in the crowd, in my opinion mostly because we think of that knowledge as acquired by first-hand. For example, as trivial as that information might seem, we read reviews online about restaurants, stores and people, and we create systems to validate people's information or opinions. A great example of this is the Ebay star system, where buyers assign stars (from 1 to 5 being 5 the best) to sellers according to the quality of the transaction, which serve as reference for other buyers. As Shirky says: "…*we increasingly learn about the world through strangers' random choices about what to share…"* Undoubtedly, self-governance and community ethics play a huge role in the amount of trust we give to our peers' production. Inside a strong online community, providing fake or malicious information will put you under social judgment: here's where self-governance plays a huge role in determining trustworthiness.

This certainly changes the way we think about institutions, professionals and amateur participation. In some cases you do want a professional to assist you, but in some cases what the crowd provides, or what you can do with the help of the crowd, can have much more value for individuals. Online reviews would be an example; people don't want the opinion of the retail store about their own products because, even if they could be saying the truth, we would always believe that their main interest is to sell the products. First-hand knowledge about that product coming from another person has much more value for me when I'm considering purchasing it. In other situations, platforms create their own systems of validation, which could be equivalent to "online cognitive authorities". Such is the case of the example of the star rating system used by Ebay: as users of the site we *choose* to believe that high-scored people are trustworthy to make transactions with. In online games platforms, for example, users with high scores can be seen as authorities since they have successfully completed several stages in the game. After all, cognitive authorities are a lot about logic.

4.5 MARKETS, THE POSITION OF NON-PROFITS AND PROFESSIONALISM

Crowdsourcing and more widely social media have had a huge impact both in the commercial and non-profit world. Some commercial initiatives have had great results using this model, by providing monetary rewards to participants. Some non-profit and cultural institutions have also implemented crowdsourcing solutions that include, in some cases, the participation of paid crowds.

However, the way in which commercial value is given could take projects from promising to complete failure. The main reason is because money, as an extrinsic motivator, can sometimes negatively affect our intrinsic ones. Deci's study on human motivation found that "… *an extrinsic motivation like being paid can crowd out an intrinsic one like enjoying something for its own sake."*

Crowdsourcing, as a model highly supported by the human theory of intrinsic motivations, can be affected by the abuse of extrinsic ones. Crowdsourcing's most important feature is that through them people can reach as many people as possible who are interested or willing to give away their free time for a particular cause. Paying the crowd to achieve the project's goal can crowd out intrinsic motivations, transforming what once was a voluntary contribution to the community into a market transaction, where the sense of community and self-governance not longer apply.

Contrary to what many people would believe, this doesn't mean that these community models are not applicable to commercial institutions, as long as they know how to keep balance between intrinsic and extrinsic motivations or if they clearly show the crowd the objectives of the endeavor. It is worth noting, though, that non-profit institutions do have an advantage: keeping crowds motivated to work for free for a non-profit is easier, since they tend to better fulfill our need of autonomy, competence and membership. As explained by Rose Holley: "*Volunteers are much more likely to help non-profit making organisations than commercial companies, because they do not want to feel that their work can be commercially exploited.*"[49] As I mentioned in the previous paragraph, it doesn't mean that crowdsourcing is useless for commercial endeavours, but keeping crowds intrinsecally motivated can be a bigger challenge for them.

Now, taking market and social online behavior further; what happens when the crowd is no longer composed only by amateurs? In some occasions community online projects are intended to find specialists in the crowd. As I mentioned before, the scalability of the Internet provides the opportunity to reach out to those professionals; that is the case of many open source

[49] Holley, Rose.

projects for example. There have been a lot of debates regarding amateurism, professionalism and ethics in social online platforms, especially regarding projects that look for revenues. Clay Shirky's theory is highly based on amateurism and Daren Brabham talks about ethical issues on his book *Crowdsourcing.*

There are no clear answers yet about this complicated topic, if there will ever be since the openness of online social projects allows highly diverse audiences. But I think that we must keep in mind that (a) volunteers or amateurs fulfill their intrinsic motivations by participating in collaborative projects, (b) paid activities can crowd out intrinsic motivations and (c) depending on the goals of your projects you may want to reach professionals and/or amateurs and build the right platforms to increase their participation as well as keeping your goals transparent.

4.6 GAMES WITH A PURPOSE

Although game theory is out of the scope of this thesis, two of the cases studied in Chapter six were based on a previous research project on games, which I thought worth mentioning.

Many of the crowdsourcing platforms used to gather metadata as tags use game systems. They rely on Luis von Ahn and Laura Dabbish's idea of Games with a purpose (GWAP), which was developed at the School of Computer Science of Carnegie Mellon University through the game *ESP*, a photo tagging game.

It is based on the fact that computers still need human aid to perform some tasks, being tagging one of them. However, through games people are asked to perform these tasks without noticing, since they do so by playing and having fun. According to von Ahn and Dabbish:

"People play not because they are personally interested in solving an instance of a computational problem but because they wish to be entertained."[50]

This project was also based on the fact that many people spend their free time, i.e. cognitive surplus, playing games. As I mentioned previously, engaging users in activities that enhance their intrinsic motivations, such as competitiveness and desire for social interaction, can help develop platforms which output is socially meaningful, but also free from errors. Thus, these systems must provide an enjoyable and challenging environment as well as keeping high quality results. Such attributes can be developed by studying game features such as timed response, score keeping, player skill level, high score lists and randomness.

In the case of *ESP*, the output quality was kept by random matching of players, repetition of tasks to ensure same output and a taboo lists of outputs. They also studied different game systems that allow quality assurance: Output agreement games (players must have the same output to score, which at the same time validates the information), inversion-problem games (one player describes an input and the other guesses a word) and input agreement games (players must determine if they have been given the same input).

This theory is further explored in Chapter six, where I provide two examples of its application.

4.7 QUALITY AND LONG-TERM VALUE

With every day growing online communities, we can't help but wonder about the quality of the information we get as a result of social media projects. This is also a concern since, even

[50] Luis von Ahn, Laura Dabbish, "Designing Games With a Purpose" in *Communications of the ACM*, Aug. 2008, Vol. 51, No. 8 https://www.cs.cmu.edu/~biglou/GWAP_CACM.pdf

though some participants are professionals, most of them are not, creating a very diverse poll of results. In the case of cultural institutions this is an issue of outmost importance, given the fact that they are still considered as cognitive authorities by their target communities. That same outreach and diversity that crowdsourcing projects look for can be also the origin of its failure. In those terms, institutions must think about a way to validate the information they get through these means. I will explore this further in the next chapter.

Regarding long-term value of this type of projects, Shirky proposes a four-level scale, based on the impact of the contributions for the society as a whole. He divides projects from personal, communal, public to civic according to their level of contribution. Personal refers to activities that only reward the intrinsic motivations of an individual, such as sharing a picture on Facebook. A contribution of communal value would be the one involved in an activity that enhances membership and generosity, for instance sharing an article on a particular group on Facebook to keep members informed. Activities of public value are those that involve the creation of a public resource for the whole community and civic ones are those which impact affect or seek to transform the whole society. Activities that have a public and civic value, are harder to create and maintain. In that sense, crowdsourcing projects in cultural institutions should always strive to go for at least activities of public value, which would not only go along their mission statement but also to hold their position as cognitive authorities.

5. CROWDSOURCING METADATA: Issues and challenges

As addressed in the previous chapter, from a social point of view, crowdsourcing and other social media related projects, when taken in the right direction, can be of great value for small and large communities, and even for the society as a whole. Surprisingly and contrary to what many people thought – and still may think – our human social behavior is the catalyst of crowdsourcing platforms, not the platforms by themselves. But that doesn't mean we must disregard any analysis about the way technological tools are developed to precisely allow that to happen. According to Shirky, the right tools create the right opportunity for people to participate in online social projects, that's what this chapter is about. Here I will study the pros and cons of developing crowdsourcing projects - specifically for metadata - from a practical point of view, in other words, what technical issues must be considered and which problems must be addressed.

In the case of cultural institutions, and most precisely archives, crowdsourcing has mostly been used either to enlarge digital collections or to gather information about collections in the form of tags. In the first case, the incorporation of new digital content to the organization's digital repository will depend on the systems they have in place for preservation, acquisition and access of digital content, thus it will vary in each case. It is also worth mentioning that copyright issues are associated to this practice and I will talk about them later in this chapter. The second case, however, and in which I'll be focusing on, brings a lot more questions and issues to solve. How do we incorporate that information to our current metadata schemas? Which standard do we use? How do we provide access to that information? As a cognitive authority, how do I ensure data quality and accuracy?

Many institutions have started a crowdsourcing project to gather tags for their photograph collections and have ended up with tons of information that can't be accessed by end users or thousands of tags that can't be sorted because the characteristics of crowdsourcing interfaces didn't allow to establish any control over the crowd's input or because the platforms where not thought to perform that task. A good example is the case of the Museum of the City of New York (which project will be explained more in depth in the next chapter). After their tagging project with the company Tagasauris ended, they couldn't find a way to integrate the tags to their current online systems because it was very hard for them to determine how this could be done without compromising the museum's authority as a source of trustworthy information.[51]

The implementation of platforms that allow users to contribute with their voluntary work adding tags, in a way that is actually useful for institutions in terms of accuracy and flexibility, and also for the community in terms of future use is still a matter of discussion and it is still a highly experimental area. This is only the tip of the iceberg if we consider crowdsourcing projects that attempt to describe time-based media, such as video or audio.

5.1 TAGGING TIME-BASED MEDIA

Time-based media represents a huge challenge, not only for crowdsourcing projects, but for cataloging in general. The biggest problem is that, in order to describe the content, the cataloger must watch it. I always like to illustrate this issue using the comparison with photographs. A photograph contains only one "frame", and to describe it I need to see one image

[51] Interview with Lacy Schutz, Project Director of the Museum of the City of New York Project "Improving Digital Record Annotation Capabilities with Open-sourced Ontologies and Crowd-sourced Workers". Dec. 3, 2013.

only. A film, for example, normally contains 24 frames per second, without considering the audio embedded (that contains a whole different stream of information). An NTSC video contains 29.97 frames per second plus audio. Each frame and second of audio could potentially contain useful information for users.

The time for full description is then also multiplied, being extremely time consuming for catalogers, leaving no other option than describing each element in a very general way or by item. This is a solution, but there's a lot from the content that doesn't get any description; information that could be extremely useful for a researcher or in broadcasting or production companies for example, where users need very specific information about certain segments - e.g. a piece of a speech - to be able to reuse the content. Tagging systems can be an easy way to bring that obscure information to light. In this case tags can provide a first layer of information, which can be very helpful to allow further and deeper description in the future.

Although out of the scope of this document, I would like to mention a system that could be an alternative to tagging, or maybe a complement for it: the system called keyframe indexing, which consists in extracting representative frames for every X number of seconds or scenes – something like thumbnails but on a timeline. Although this provides an overview of the content of a video, because it essentially allows one to "browse" a video, it doesn't provide any description of the content, which has to be done by a human.

5.2 CONTROLLED VOCABULARIES

Controlled vocabularies have been the inseparable companion of catalogers for decades. They are a list of words and phrases that are systematically assigned to works with the goal of having easy retrieval of information as well as some hierarchies. They solve ambiguity, such as synonyms, homonyms and homographs, for example. Some examples are subject headings, thesauri and taxonomies. Controlled vocabularies are carefully organized and their application can only be done by trained staff, especially if applied in conjunction with other structure standards such as MARC (the Machine Readable Cataloging). From the point of view of the user, in many cases, the assistance of a trained librarian is needed to find the resources needed, since a simple keyword search could provide thousands of results.

The case of tags is exactly the opposite; they have no rules, nor hierarchy. Tags are messy. So, what's the benefit of using them? Oomen and Aroyo asked themselves the same question, and from their point of view, tags can help bridging the semantic gap between formal catalogs and users, in other words, they allow users to actually find what they're looking for since materials are described using their "untrained language". Through the *Waisda?* project (explained in the next chapter), Oomen and Aroyo's idea was corroborated: many tags were not in the thesaurus used by the institution, but they legitimate words, thus words that could potentially be used by patrons to find content. From my point of view, this doesn't mean that we can leave aside professional cataloging, but tags can be a great complement to enhance discoverability. Moreover, the MCNY photo-tagging project found that with the platform working, catalogers could use their time performing other important tasks, such as prioritizing collections.[52]

But, what if we can actually give tags certain hierarchy and allow disambiguation? Linked Data could be the answer to that question.

[52] Museum of the City of New York, "NEH Grant Final Report: Improving Digital Record Annotation Capabilities with Open-sourced Ontologies and Crowd-sourced Workers", April 30, 2013 https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-51480-11

5.3 TAGS AND THE SEMANTIC WEB

The semantic web has opened the possibilities of using and reusing data in many different fields. In the case of tag crowdsourcing, it provides common global databases that can be used as the basis of these projects. This is possible because these hierarchical open online databases provide something comparable to a huge controlled vocabulary, which can be freely accessed through RDF files, RDF endpoints or APIs (See Appendix 2, Glossary). The next chapter will study some projects that used this model, providing a more deep insight of how this is done.

John Breslin et al. list the issues of tagging systems that can be improved using Linked Data:

(a) Ambiguity: as a simple string of characters, a tag can have different meanings depending on the context. For example, if a person uses the tag *Harry Potter*, he or she could be talking about the book, or the movie, or even the new section in Universal Studios' theme park. Linked Data allows being specific about which *Harry Potter* the user is referring to. It is also useful with words that have various meanings depending on the context. For example, pool (to swim) and pool (the game).

(b) Heterogeneity: on the other hand, in some cases many tags can be used to refer to the same term. For example, *restroom, bathroom, toilette, comfort room,* etc. By using specific online data sets as a controlled vocabulary these problems can be avoided.

(c) Lack of organization: tags on their own don't have any organization or hierarchical structure, making sometimes difficult to establish relationships between tags. Linked Data, by definition has a specific structure (ontology), thus any project using it gets the benefits of its structure.

5.4 DIVERSITY, A BLESSING AND A CURSE

There have been different implementation approaches to solve the issues associated with tagging, such as defining very clear and discrete tasks for volunteers or implementing games as a way of validating the veracity of the tags. But there's still something we can't change: volunteers in front of the keyboard having the freedom of typing whatever they think is the best word (or words) to describe the picture they see on the screen. Even with the best of the intentions, subjective terms and inherent language characteristics can't be avoided.

Freedom, the same attribute that makes crowdsourcing an effective way of engaging users, becomes a problem in the realm of cataloging and description. When tagging people have the freedom to choose any term they deem useful for the image or sound they are presented with, which makes an excellent opportunity for the use of folksonomies, for example. The problem with folksonomies is when they represent only a small portion of the community or, in some cases, they can only be understood by some users. For instance, what word would you use to describe the image of a soft drink? People on the North East of the US would say "soda", in Texas many would say "coke" and in the North West it would be "pop". But they're all talking about the same thing; they are synonyms, but they're also folksonomies, since these terms are not formal words for the actual elements they refer to (the formal terms would be soft drink) and because they have a different meaning (or not meaning at all!) depending on the region.

Another big problem with tags is that some systems do not allow tags that include more than one word, or people can have different criteria about describing an image. For example, to describe a photograph of a white cat, people can use white, cat, whitecat, white cat, white-cat, white&cat and even folksonomies such as the name of their pet or a cartoon or misspellings like witecat.

Because of the inherent characteristics of tags explained above, they are incompatible with most current standards and systems used by archives to store and use metadata. Tags contributed by volunteers are messy.

In spite of all the issues and difficulties that systems need to sort, crowdsourcing is still a possibility that is worth exploring. Aside from all the benefits that social media brings for institutions in terms of connecting organizations and collections to communities, from a practical side, tagging has several advantages. Some of them constituted the foundations of the *Waisda?* project: bridging the semantic gap, enrich collections with factual information, increase connectedness and defining the future workflows of digital content (see Chapter 6).

Another very important benefit is the fact that tags allow related content to be linked. In other words, digital elements that have similar content can be tagged with the same term or word. Most platforms that include tagging systems allow users to browse content that has the same tag, by only clicking on it. This is certainly an advantage compared to traditional systems, where identifying and linking related content requires time and personnel. For example, to determine relationships between several videos, the cataloger needs to sit and watch each of them, to later add the relationship information on the cataloging system. While not completely accurate, for reasons I will explain below, tagging allows a first layer of relationship that can be used as a primary search system. In the same vein, thanks to that characteristic, tags allow one to find related content faster through the presentation of tag clouds. Users only need to click on the word to see a list of all the files tagged with the same word.

Systems that include tags are also easily searchable through keywords, acting somehow as subjects or topics. Though, most of them unfortunately don’t have a hierarchical structure that would make the search even easier. But this problem could be solved with the use of Linked Data.

Additionally, tagging projects are a very good way to engage community participation because it is a very straight-forward task, in which users can easily engage on without requiring too much time or specialized knowledge. Thus, tagging projects are perfect for the development of micro-tasks, i.e. small and straight-forward jobs, with different levels of difficulty that allow several levels of commitment, which translates to high flexibility in the time users spend performing the task. Many, if not all, tag-driven crowdsourcing projects use this system. It is worth noting that micro-tasks must be carefully designed, to allow enough flexibility but also commitment, inviting people to step in and out, but also to come back regularly.

Finally, crowdsourcing projects allow reaching large and very diverse groups and communities, thanks to its scalability. For the particular case of archives, this is a huge advantage, since the likelihood of having more accurate tags increases as well as the chance of describing or identifying unknown or very obscure content.

5.5 PROFESSIONAL METADATA VS. USER METADATA

One of the main concerns behind the use of crowdsourcing to gather metadata about collections online is the difference between professional cataloging and user cataloging. Institutions and catalogers have developed in the last fifty or sixty years systems to describe collections in a way that information can be retrieved easily, to find the elements on shelves for example, but also to describe them in a way that allows users to find faster what they are looking for. Controlled vocabularies are one of them, but several organizations have also created different cataloging systems and conventions with the same objective as well as to allow sharing information between institutions. Some of those systems have been updated over the years to allow new technologies and new uses, but they have not really been re-envisioned for our new interactive way of using technology. Those systems and standards were designed to be cooperative and interoperate among cultural institutions, but not cooperative in the broad sense of the word – involving the public.

It is true; people do not apply conventions when using these platforms and we can't expect that to happen. However, contributions made through these channels can be very valuable and unique, so how do we manage them in order to make them compatible with the existing systems? I ask this question because many institutions using tagging systems agree that both users and professional tags must be kept, as Marieke Guy and Emma Tonkin say "*We agree with the premise that tags are no replacement for formal systems, but we see this as being the core quality that makes folksonomy tagging so useful.*"[53]

The archival community must think in a way to integrate these annotations in a way that is compatible with the current systems and standards.

In an interview with Sukdith Punjasthitkul, project manager of Metadata Games (see Chapter six), he pointed out that for them this issue has always been a concern and that they believe tags should be integrated to formal descriptive systems to ultimately make them public in catalogs. Currently they recommend saving the information using the Dublin Core Metadata Schema under the field "subject" or the Metadata Object Description Schema (MODS) using "subject" and "authority attributes". Later this information can be mapped to other schemas and standards if necessary.

[53] Guy, Marieke, Tonkin, Emma, "Folksonomies, Tidying up Tags?" in D-Lib Magazine, Volume 12, Number 1, 2006, http://www.dlib.org/dlib/january06/guy/01guy.html

5.6 IMPROVING AND VALIDATING TAGS

Because of the way previous archival standards have been developed and also because of the value community gives to cultural institutions in terms of trust and accuracy, validation is a big issue for cultural institutions, thus any system applied to ingest tags to the formal descriptive system used by the institution must not only be effective but accurate as well, because of their status of cognitive authorities.

The first method that some institutions have suggested to improve tagging, or at least to avoid the problems mentioned above, is to educate users, teaching them how to add tags that can be useful for future searches, for example avoiding plural terms (using cat and no cats), teaching them some conventions for example when using a personal noun (using Marilyn Monroe instead of M. Monroe or Monroe, Marilyn) or avoiding the use of symbols such as # or /. This would require a lot of communication with users and the problem would still persist. After all, with that philosophy we are still expecting that users are committed enough to pay attention to the quality of their tags and again, going back to the previous chapter, we would be making the experience more and more complicated for the user which could finally limit the quantity of collaborators. The case of the MCNY and Tagasauris is a great example. They developed an online list of micro-task descriptions, which included examples and videos.[54]

Some institutions have implemented online games to gather and improve crowdsourced tags. Such is the case of the Institute for Sound and Vision of the Netherlands and their project *Waisda?*, and the project *Metadata Games* from the Dartmouth College. Both projects based on the Games with a Purpose concept explained in the previous chapter.

[54] List and videos available here http://instructions.tagasauris.com/

Another viable solution to improve tags is using software such as Open Refine (ex-Google Refine), a "free, open source, powerful tool for working with messy data".[55] This tool allows cleaning, transforming and reformatting data. It also allows linking data with other databases and reconcile[56] terms against Freebase. However, the process is still somewhat manual, not really suited for large amounts of data.[57]

The other solution, and until today the most effective one, is cleaning tags – or even gathering tags - using the semantic web. Systems using Linked Data allow users to select the words they deem useful for an image or video straight from a dataset, reducing any manual process afterwards. This can be easily complemented with other algorithms to filter misspelled words or meaningless strings of characters.

During the interview with Sukdith Punjasthitkul, he also mentioned that Metadata Games is currently working on a project proposal to research the possibility of validating tags in a system similar to the Oxford Dictionaries Online[58], where tags could be validated by an institution according to the number of search matches, in other words, according to the number of times patrons use them to find content. To do so, however, data must be made available first.

[55] Documentation and software available here http://openrefine.org/

[56] Reconciliation is a semi-automated process in which an application (usually an API) provides a list of suggested terms from a particular Linked Data dataset, which are matched with a particular word in the institution's dataset. Doing this allows linking your data with the data cloud of that particular Linked Data project.

[57] Google Refine 2.0, *Data Augmentation, 3 of 3.* This video shows how to reconcile data using Google Refine https://www.youtube.com/watch?v=5tsyz3ibYzk

[58] http://www.oxforddictionaries.com/us

5.7 COPYRIGHT AND LEGAL ISSUES

In order to publish content on the web for people to tag, institutions must be aware of the copyright laws that regulate this type of activities. While the US Copyright Law does not include any mention to crowdsourcing, all institutions holding copyrighted materials know that uploading content to the web without the corresponding permissions or licenses is illegal. However, in the case of many "orphan works",[59] the chance of uploading them is a great opportunity to know more about the work and eventually about the creators. When I first started this research I was hoping to find more projects including this type of collaboration. I was surprised to find that institutions are very cautious when it comes to making orphan works public. However, some of them do publish these works under a "no known copyright restrictions" policy.[60]

An interesting issue regarding copyright is the one placed by the materials uploaded by users. Since these platforms are considered to be open and used by everybody, the current copyright law is not adequate to define the legal framework for community uses. In that sense, institutions have chosen to include in their policies and/or guidelines the use of Creative Commons Licenses[61] that are normally accepted when the users register in the websites or suggested in the guidelines. Such is the case of the project PictureAustralia, which states on its website: *"While this is not a condition for contributing to this group, we suggest you consider licensing your images with a Creative Commons License like "Attribution-NonCommercial."*

[59] A definition of orphan Works can be found in the document published by the Society of American Archivists, *Orphan Works: Statement of Best Practices,* 2009, available at http://www.archivists.org/standards/OWBP-V4.pdf
[60] Flickr The Commons Project http://www.flickr.com/commons/usage/
[61] More about Creative Commons Licenses http://creativecommons.org/licenses/

Some institutions also set a minimum number of conditions to upload materials, in order to have enough information about the source and the content. For example the Trove Project publishes in the website:

> *Potential acquisition:*
>
> *To have your [the user's] images considered for acquisition by a Trove contributor:*
>
> *- include detailed captions, descriptions and tags;*
>
> *- upload high resolution images e.g. 2500 x 1900 pixels;*
>
> *- allocate a creative commons or all rights reserved license;*
>
> *- include alternative contact details or check Flickr mail regularly.*

This is a perfect example of a well thought workflow for new material, especially when the process is expected to be massive, in order to avoid having problems in the future with orphan content.

5.8 PLATFORMS FOR CROWDSOURCING TAGS

Now, I've talked about the characteristics platforms must have to allow crowd participation and, as much as possible, tag accuracy. However, there are only a few institutions that can afford to develop these platforms; it requires specialized professionals, economic resources and interdisciplinary teams that understand technology and metadata implementations.

That's why some organizations have developed open source crowdsourcing software. Such is the case of *Metadata Games*, *Waisda?* and a couple of years before the *steve.museum* project. These software's documentation and code are available for download online.

Metadata Games even designed a software that doesn't require much technical knowledge: institutions have to simply upload their content (photos, videos, etc.) to be tagged, which is administered by Tiltfactor. As a result institutions receive the tags for each element as .csv files. By doing so, institutions agree to have the tags and metadata under a DPLA public domain license, for future re-use.

Another interesting thing to mention, in relation to the previous section, is that these software, being open source, offer the possibility of developing an on-site application or version of the software to allow the playback of content that can't be made available online for copyright reasons, allowing users to tag content inside the institutions. Mary Flanagan - director of Tiltfactor - in an interview with Trevor Owens mentioned that this is also possible by applying restrictions to certain IP addresses.[62]

[62] Trevor Owens, *The Metadata Games Crowdsourcing Toolset for Libraries & Archives: An Interview with Mary Flanagan*, April 3, 2013
http://blogs.loc.gov/digitalpreservation/2013/04/the-metadata-games-crowdsourcing-toolset-for-libraries-archives-an-interview-with-mary-flanagan/

6. CROWDSOURCING METADATA: Three Case Studies

I will present and study in this section three different projects that implement systems to gather metadata as tags using online interfaces. While these are not the only institutions implementing crowdsourcing projects in the cultural world, they certainly represent an interesting sample of what current implementations do, their results and future applications and possible improvements. The three of them are in an experimental stage, thus their input is extremely valuable to study the effectiveness of the crowdsourcing metadata model for cultural institutions.

The first project I'll introduce is *Metadata Games*, developed by the Tiltfactor Laboratory at Dartmouth College, led by Mary Flanagan, Sukdith Punjasthitkul, Max Seidman and Geoff Kaufman. *Metadata Games* is a free and open source software focused on gathering metadata for photo, audio and video through different game styles.

The second one is the project developed by the Museum of the City of New York (MCNY) in collaboration with the company Tagasauris. The goal of this project was to gather metadata of a collection of photographs as tags using Mechanical Turk's online workers through the development of very specific micro-tasks. While this project didn't use time-based materials, it is interesting because the tags were derived from Linked Open Data projects online and because they developed a monitoring system to evaluate the worker's performance.

The last project included in this study is *Waisda?,* developed by the Netherlands Institute for Sound and Vision in collaboration with the VU University of Amsterdam. *Waisda?* is an online platform that collects metadata as tags for videos through a game. This project is particularly interesting because it was one of the first institutions in experimenting with this type of platforms for time-based media. Because of that, the project has evolved and its features have been adjusted over the years, showing an interesting record of the development of the crowdsourcing model, which is also possible thanks to its rich online documentation.

These three projects will provide a practical point of view to the study of crowdsourcing implementations for cultural institutions.

6.1 METADATA GAMES

Metadata Games is a project funded by the National Endowment for the Humanities (NEH) and the American Council of Learned Societies (ACLS), and developed by Tiltfactor Laboratory at Dartmouth College, defined by its members as a "*socially conscious game design laboratory*".[63] The project is led by Dr. Mary Flanagan, an artist, author, educator and designer.

Tiltfactor, founded by Flanagan in 2003, is a design studio dedicated to the study of games and the generation of new knowledge using the psychological principles of games with the aim of creating an impact on users' thoughts and behaviors as well as to develop social values and civic engagement, transforming individuals into what they call the *citizen archivist.*

The main objective of the project is to create free and open source customizable software (FOSS) to gather metadata about audiovisual media through the crowdsourcing model. The ultimate goal is to save digital cultural artifacts worldwide by providing tags that would allow communities to have access to innumerable collections. Currently, *Metadata Games* offers 5 different games: *Zen Tag, Nextag, Guess What!, One Up* and *Pyramid Tag*. As a research

[63] http://www.tiltfactor.org/

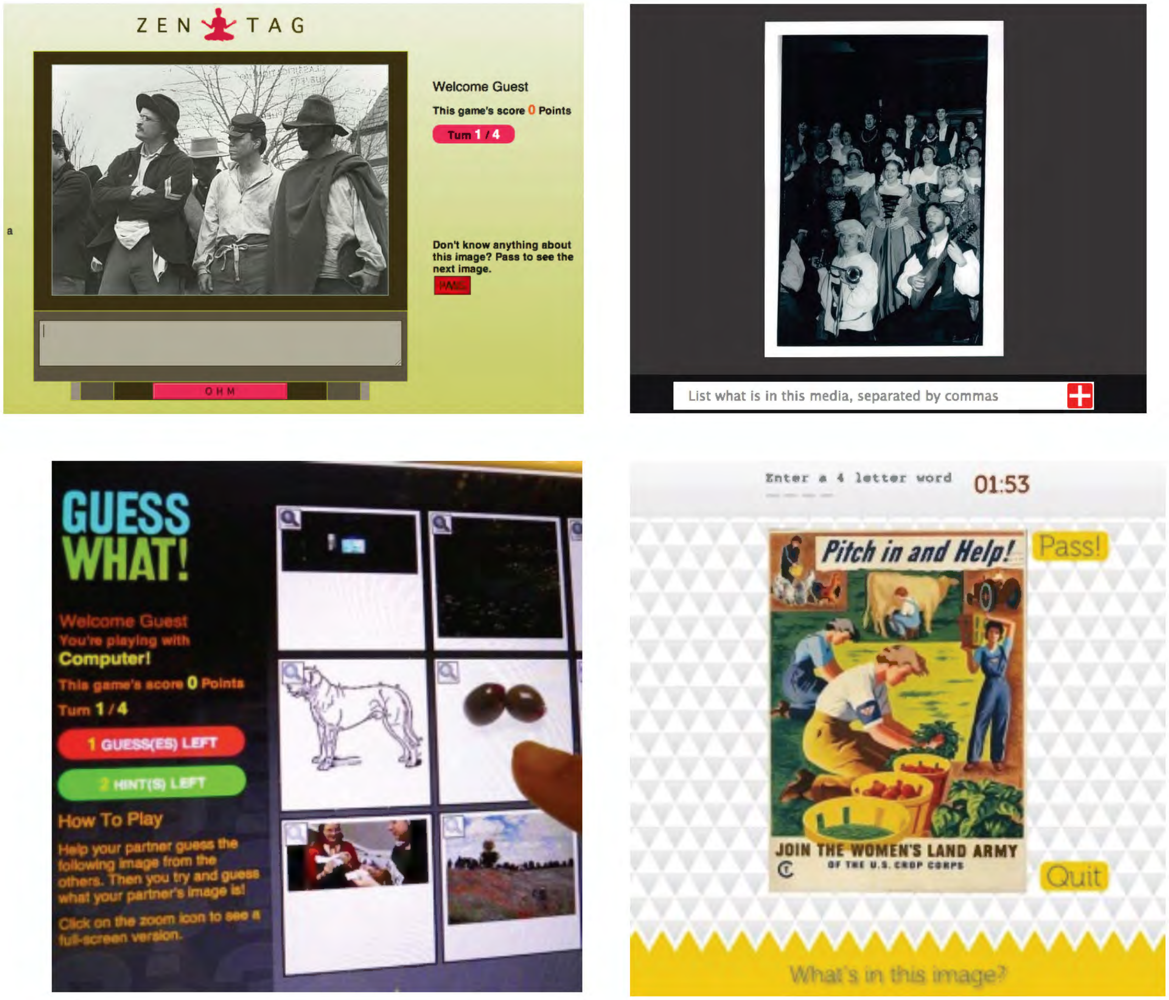
project, all versions of the games are constantly improving, based on users' feedback and internal research projects about the effectiveness of the platforms.

Based on Luis von Ahn and Laura Dabbish's "games with a purpose" (GWAP), *Metadata Games* takes advantage of user's intrinsic motivations; mainly altruism, love for a subject area and the desire to compete and be challenged. The design aims to use unique game techniques, such as mechanics, dynamics and reward systems, to collect accurate tags while ensuring high levels of player engagement.

6.1.1 Zen Tag, Nextag, Guess What! **and** *Pyramid Tag*

Zen Tag is a basic one-player photo tagging game, where the user is required to enter as many tags as possible, to describe the image accurately. Tags must be entered separated by commas. The player gets points for each tag and extra points for new tags, i.e. never entered before for that photograph. *Nextag* is the improved version of *Zen Tag*. It works in the same way, but *Nextag* also allows tagging audio and video. *Guess What!* is a two-player game. The first player must provide enough information about an image through description so the other player can guess what's on it.

Finally *Pyramid Tag* is a one-player game in which the user must guess as many previously recorded tags as possible for that image. This game is available as a mobile app, for both android and mac systems.



Don't know anything about this image? Pass to see the next image.

List what is in this media, separated by commas

Help your partner guess the following image from the others. Then you try and guess what your partner's image is!

Click on the zoom icon to see a full-screen version.

Figure 14. Captions of Metadata Games. *From left to right:* Zen Tag, Nextag, Guess What! *and* Pyramid Tag

6.1.2 One Up

This game has been studied in detail by the Tiltfactor team, both in terms of platform design as well as in its effectiveness in providing good quality and accurate metadata. *One up* is a mobile game app where players get points for submitting single-word tags against a friend or random player. The application is customizable and users can enter their areas of interest, which will be considered when selecting collections to tag. However, the game is geared towards experienced and non-experienced users, by making it enjoyable even if the subject matter is not familiar to that person. This mobile version was launched in January 2014.

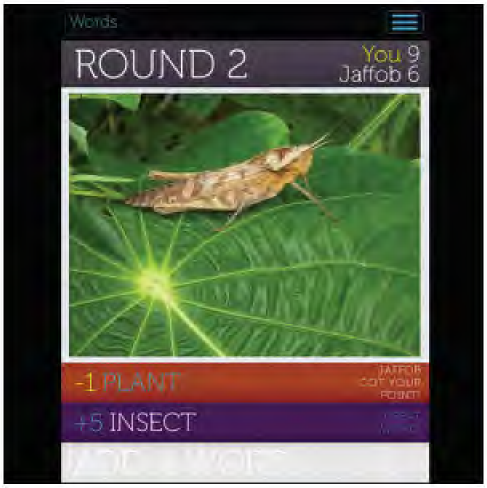


Figure 15. One up, *Mobile Interface*

The game is divided in three rounds:

(a) Each player enters three tags per image, one point is rewarded for each tag and additional points are rewarded – called accuracy bonus - if they enter previously recorded tags, i.e. tags already stored in the system, which were provided by other players in previous games.

(b) Accuracy bonuses increase. However, if players enter a tag the opponent previously entered in round one they are penalized with one point and the opponent is awarded with one (you have been one-upped!).

(c) Both accuracy bonuses and penalties increase. After the three rounds the player with the higher score wins.

One of the main concerns of the developing team was to determine ways to avoid people cheating by entering inaccurate metadata. This concern had its origins in the commonly developed peer-to-peer validation games, in which the tags are compared to the opponent in order to determine its accuracy. The problem with this type of validation is that players could agree on the tags entered to earn more points without really contributing with good information. This problem was solved as follows:

(a) Normally in these applications players are paired randomly, which guarantees that players don't have other interaction outside the game. However, this method inhibits the social experience, which is fundamental in the success of online platforms. Thus, the team resolved to compare and match tags only with previous games, in other words, with the game's database where all tags are stored for each image. This minimizes collusion but at the same time allows users to play with friends, strengthening social behavior.
(b) All good games have a risk, which not only encourages players to do their best, but also increases competitiveness. Some games implement, for example, a list of taboo words, which are displayed along with the image on screen. This increases the challenge by limiting the words players can use, encouraging them to think of more specific and accurate words that would finally enrich the information about the image. But these lists can't be very long since they could discourage users to play the game. In *One up*, the list

of taboo words is unknown, because those words are the ones entered by the opponent in the previous rounds.

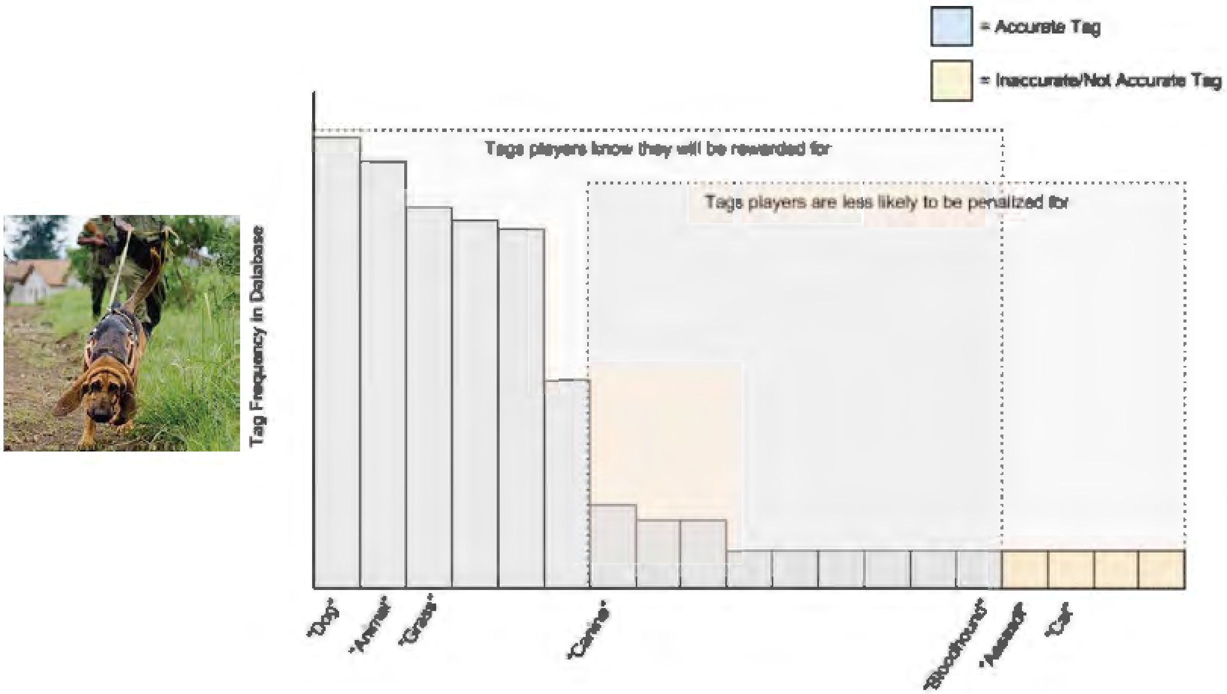


Figure 16. The intersection shown is "The Holy Grail" of metadata, i.e. tags that are very specific and accurate.

Finally, players don't enter inaccurate tags, because to earn more points they must also match their tags to previously entered tags. However, they also avoid entering obvious words, because they represent a risk to have matches with the opponent - which would end up in the possibility of losing points. Thus players try very hard to be as specific as possible. With this methodology the game intends to increase the amount of useful tags, which fall under the so called "Holy Grail of metadata": in Figure 16 the intersection between tags players know will be rewarded for and tags that are less likely to be penalized for.

In order to test the effectiveness of the game as described above, they compared tags gathered through *One-up* with tags collected by *Zen Tag*. Tags generated by *One up* had better accuracy and specificity than *Zen Tag*'s tags.

As an open source project, the code is available for download on Github under a AGPL v3+ license. As pointed out by Dr. Flanagan in an interview with Trevor Owens from *The Signal* of the Library of Congress, the software is flexible enough to allow different institutions with different budgets and collections using this tool to improve the description of digital collections. During the design stage they had to consider the different tools and states of technological development of many institutions; many of them recognized they were still using outdated technology, both for websites and databases. She adds, "*We went with a solution that is familiar for now, and can be upgraded later through a plug-in architecture.*"[64]

Another flexible feature of the software is that many institutions have concerns about copyright issues with their collections, which makes impossible for them to upload material online, even for these purposes. Dr. Flanagan states that "*The fact that the institution can 'own' their own data is essential for most of our affiliates who aren't legally allowed to share some of the collections on the internet due to copyright restrictions and such. Institutions can use the system in-house if desired, or restrict Metadata Games to a particular IP address.*"

Currently, the developing team is working on a new game called *Stupid Robot*, in which the player has to show the world through words to a robot that "sees everything but understands nothing". The idea for this new game was born after the DPLAfest in October 2013, which gave

[64] Trevor Owens, "The Metadata Games Crowdsourcing Toolset for Libraries & Archives: An Interview with Mary Flanagan" in *The Signal*, April 3, 2013 http://blogs.loc.gov/digitalpreservation/2013/04/the-metadata-games-crowdsourcing-toolset-for-libraries-archives-an-interview-with-mary-flanagan/

the team insightful feedback about *Pyramid Tag*. According to users' comments, matching expert tags was frustrating for them when playing the game.

6.2 MUSEUM OF THE CITY OF NEW YORK AND TAGASAURIS

The Museum of the City of New York (MCNY), together with the New York based company Tagasauris embarked on a NEH funded project to increase the accessibility of their digital collections - mainly photographs - through the use of a platform that combines two models: crowdsourcing and Linked Open Data. The idea began after the institution realized that their already existent digitization project was creating more digital objects than what catalogers could describe, thus making thousands of photographs indiscoverable, not only for the users but for the museum as well. In addition, this situation was creating a huge digital backlog. Catalogers could generally describe collections and provide basic description for each photo element, but the museum realized that they needed more basic information about each photograph - such as number of people in the photo, if the image was horizontal or vertical, nigh or day, etc. – to fulfill patrons needs, in order to provide straightforward sorting and search services.

Unfortunately, the institution was unable to hire more catalogers to do this because of space and budget constraints. Under these conditions, the team considered this was an ideal project to implement the crowdsourcing model, since it could be divided in discrete and straightforward tagging tasks. This project, as opposed to the other two initiatives discussed in this section, the main focus was not to create a generic application or software, but rather to implement a particular solution for this institution. However, the results not only helped the museum to have more granular information about this collection, but also to improve the services that Tagasauris offers online. It is also an interesting project since it used paid workers, not volunteers, to gather tags.

In order to make this project viable and to make sure that the online data sets used would fulfill the project´s needs, Tagasauris, in charge of the technical part of the project, first reconciled and/or merged MCNY´s data sets with Freebase. Reconciliation is a semi-automated process in which an application provides a list of suggested terms from a particular Linked Data set – in this case Freebase - that is matched with a particular word in the institution's data set. This would allow avoiding the repetition of entities as well as contributing with new terms to this crowdsourced, free and open online database. This was initially possible thanks to the system previously developed by Tagasauris to communicate the crowdsourcing platform with the museum´s Cortex[65] digital asset management system.

Through the use of online crowdsourced marketplaces provided by Amazon´s Mechanical Turk, Tagasauris implemented an online interface, which included 15 micro-tasks. These micro-tasks, and their associated actions, were discrete and very straightforward tagging tasks, which were divided mostly by type. For instance, there was a task dedicated to the description of gender on the picture, other to count the number of people, other for location, etc. (See Figure 17 below) This initial sorting was key to the results of the project, since every task was associated with a determined dataset on Freebase, decreasing the chance of error by the workers.

[65] http://www.orangelogic.com/

Micro-task	**Description**
Descriptive Tags	Identification of main subject(s) and action(s) in the image.
Caption Tags	Relevant tags extracted from the caption (if available).
Tag Day/Night	Day or night.
Tag Indoors/Outdoors	Indoors or outdoors.
Tag Photographic Styles	Aerial, bird's eye, streetscape, background, cross-processed, full frame, full length, head and shoulders, landscape, lens flare, looking at camera, low angle view, multiple exposure, panoramic, portrait, scenic, still life, studio shot, urban scene, wide angle, X-ray image.
Tag People Count	None, 1, 2, 3, 4, group.
Tag Gender	Male, Female
Tag Age	Infant (birth to age 2), child (3 to 12), teen (13 to 19), young adult (20 to 35), mid adult (36 to 65), senior adult (66 and older).
Tag Geolocation	Currently location is mapped to Google Maps using Named Entity Recognition or human judgment.
Tag Emotion	Joy, trust, fear, surprise, sadness, disgust, anger, anticipation, optimism, love, submission, awe, disapproval, remorse, contempt, aggressiveness, etc.
Quality Control	Crowd-sourced workers voted on the appropriateness of fellow workers' tags.
Semantic Reasoning Module	Presents the possibility to reconcile similar terms.

Figure 17. List of micro-tasks, based on MCNY's Report to the NEH, 2013.

Each worker, then, would choose from all these micro-tasks the one they felt more comfortable with. This decision was made after the project team realized that when workers choose their tasks the results are better in terms of quality and productivity. It is worth noting that, in this case, the theory about intrinsic motivations doesn't fully apply, since workers are getting paid, in other words, they have an extrinsic motivation to do the job. The fact that productivity increases could or could not be related to intrinsic motivational factors. For example, even if a person is getting paid, she could enjoy tagging photographs by geolocation because she loves geography. However, all participants in this project were paid workers, thus there's no point of comparison between the performances of volunteers versus workers in terms of tag quality.

In order to improve the performance of the online workers, Tagasauris provided direct communication with them via Skype, chat and instant messaging. All tasks were also thoroughly described on videos online.[66]

To provide a better evaluation of the project as a whole and in its individual parts (micro-tasks), Tagasauris also developed a monitoring tool. This tool provided statistical information that would be useful in the future to evaluate the effectiveness of the combined model. With this information, the team was able to assess the performance of each worker as well as compare crowdsourced metadata with professional metadata previously recorded by the museum´s staff.

[66] Videos can be found here http://instructions.tagasauris.com/tagging-main-subject

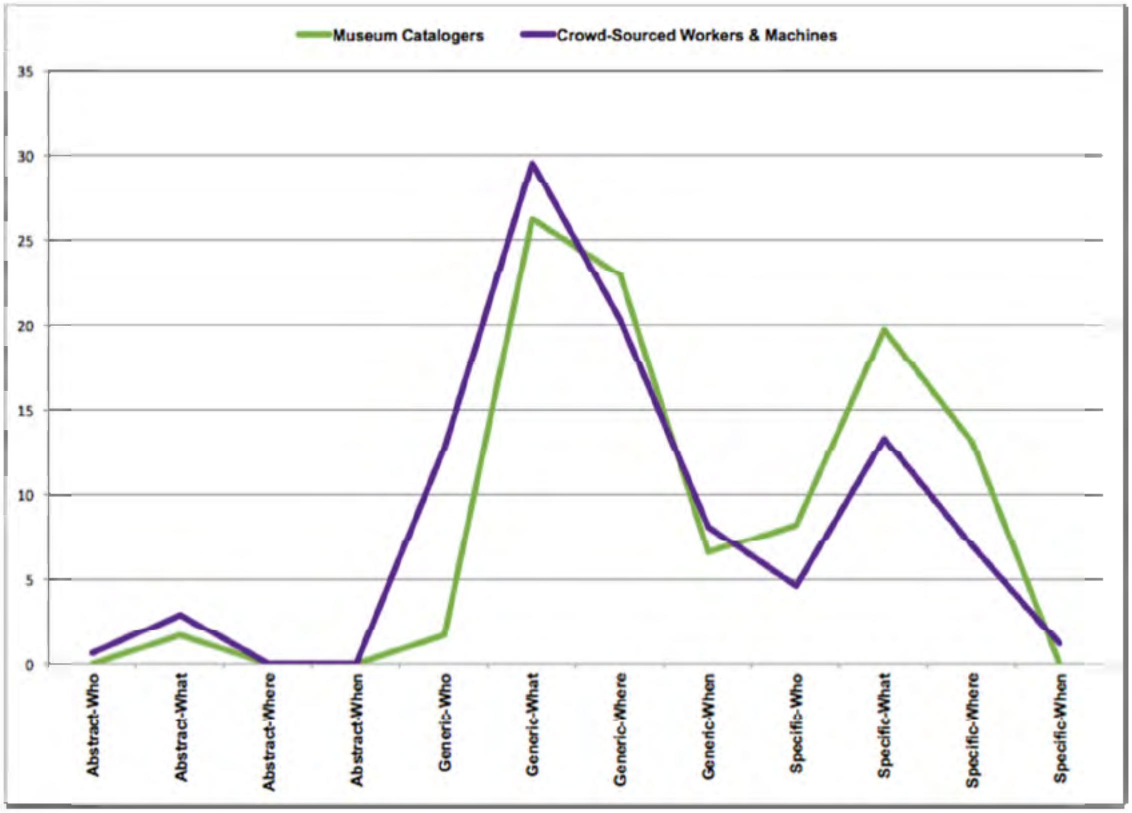


Figure 18. Statistical information about the type of tags and level of description for professional catalogers and crowdsourced workers.

The results, in terms of quantity, were somehow expectable: online workers provided more tags per photograph than professional catalogers. Additionally, to assess the performance of the project in term of quality the museum developed a model based on Panofky-Shatform matrices, a hybrid model that basically divides the tags in two levels of complexity: generic/specific/abstract and who/what,/when/where. In this sense, the results were surprising: crowdsourced tags had a similar quality compared to the professional annotations. It is worth noting though that professional annotations were basic descriptions of the photographs, which intended to provide searchable and sorting terms for each object with the aim of making it discoverable in the future (see Figure 18). This evaluation also showed that professional catalogers can sometimes provided more complete information, because they have knowledge of the background of the collections (Figure 19).

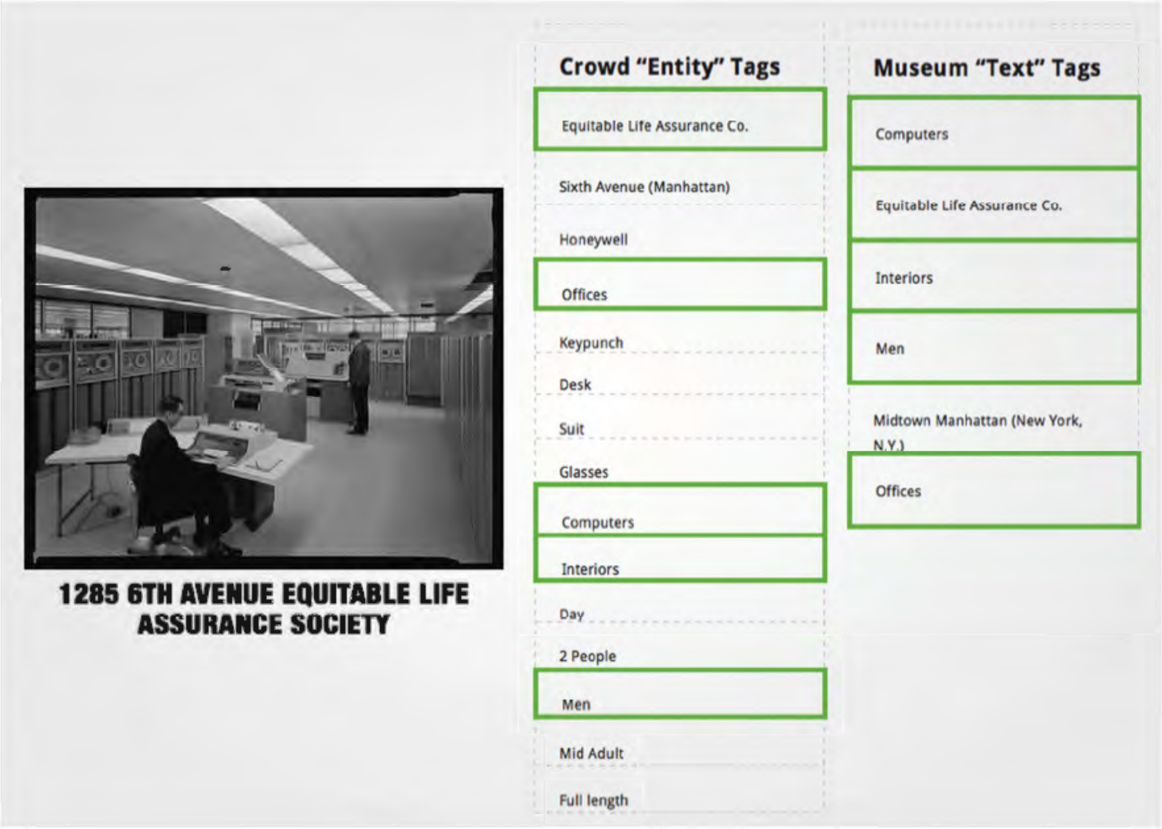


Figure 19. Comparison between tags provided for the same photo by the crowd and professional catalogers.

Far from showing that the work of catalogers can be replaced, this project showed that library and museum professionals can take advantage of these tools and models to redirect their efforts to supervise, design and overview their performance, adding more value to their work.

Now that the project is over, and with the positive evaluations, the museum and Tagasauris hope to continue improving the model, especially to enhance the quality of tags.

Tagasauris was able to improve their online service as well as develop plans for the creation of new online tools that will allow, in the future, to use this system to tag time-based media.

For MCNY, a problem that still remains unsolved is connecting the crowdsourcing platform to their online *Collections Portal*, in order to actually provide access to the collections to their users. In an interview with Lacy Schutz, the project director, she manifested that this issue was discussed after the project was over, but they couldn't find a proper way to display the information online, since for the Museum validation of the crowdsourced information is vital before making it public.

6.3 WAISDA?

Implementing tag systems to gather descriptive metadata for video elements is a particular daunting task. As I mentioned in section 5.1, in terms of technical requirements, the system must be capable of not only saving the tags, but also saving the moment in which the tag was added, otherwise it loses meaning. Not many institutions have the resources and trained staff to implement these types of platforms, which more often than not are very complex; but also not all of them have the time and funds to implement research projects around this very new way of cataloging audiovisual materials. There have been several projects that experimented with tagging systems for time-based media such as *VideoTag*[67], or *TagATune*[68] but most of them had a hard time sustaining them in the long-term. The project *Waisda?* has been so far the exception.[69]

The Netherlands Instituut voor Beeld en Geluid (The Netherland Institute for Sound and Vision) in collaboration with the VU University of Amsterdam and KRO Broadcasting was one of the first institutions that took the challenge of doing research around tagging systems for the recollection of metadata for time-based media and certainly the longest so far (2009 - 2013). In a joined effort they created an online platform - developed by the software development company Q42 - called *Waisda?*, a crowdsourcing project for audiovisual tagging based on a game.[70]

This project is not only interesting for the novelty of tagging videos but also because it involved a lot of research about the validation, usability and searchability of the annotations entered by the users. The decision of making of this project a game was not only driven by the idea of engaging communities with collaborating projects in an entertaining way, but also because it was a way to validate the information. The highly interdisciplinary team, the wide range of aspects of crowdsourcing studied, its long existence, the flexibility of the tools developed and the openness of the research and results give this project a notoriety that is certainly worth exploring.

[67] http://www.videotag.co.uk/

[68] Edith L. M. Law , Luis Von Ahn , Roger B. Dannenberg , Mike Crawford, *TagATune: A Game for Music and Sound Annotation*, http://www.cs.cmu.edu/~elaw/papers/ISMIR2007.pdf

[69] It is worth noting that this project is part of the NWO Project Agora and PrestoPRIME and that the institution also collaborates with other projects in Europeana, such as EUScreen, which shows the institutional support that has made possible to carry out this ambitious research.

[70] Waisda? website http://woordentikkertje.manbijthond.nl/

Background studies for this project revealed that the advantages of social tagging for cultural institutions are:[71]

(a) Bridging the semantic gap**:** as opposed to what many people would think, professional cataloging - which has very clear structures, hierarchies and specific vocabularies – can generate a distance between users and catalogs, which can create difficulties in the search of resources. Common users are not familiar with controlled vocabularies, which sometimes can be very far from what people think an element should be described as. "*Research has proven that tagging can indeed bridge the semantic gap and consequently increase the findability of archival objects."* In this particular case, the lack of hierarchical structure can be beneficial, since users do not require previous knowledge about the description of the resources.

(b) Enrich collections with factual information: cumulative knowledge, i.e. the information gathered through the participation of larger crowds, can provide a wide range and high diversity of tags that can complement professional cataloging. Some people may think that having a large number of tags can be very hard to manage and not worth the effort, but according to the *Waisda?* team *"the possible disadvantage of an excess of tags can be countered by the argument that this just provides more possibilities for people to retrieve archival materials."* Furthermore, *Waisda?* has proven that storing and managing time-based tags is possible and effective.

[71] Netherlands Project, Various Authors, "Emerging Practices in the Cultural Heritage Domain, Social Tagging of Audiovisual Heritage" in *Proceedings of the WebSci 10: Extending the Frontiers of Society On-Line (*Raleigh, NC: US April 26-27th, 2010) http://journal.webscience.org/337/2/websci10_submission_23.pdf

(c) Increase connectedness**:** tagging systems are a good way to connect communities with collections, engaging them in activities that fulfill their intrinsic motivations by making them realize the long-term benefit and usability of their own contributions. In other words, people play because it's fun, but also because it will help them and other people find the information they need in the future, a genuinely altruistic motivation that has its roots on social and community behavior.

(d) Defining the future workflows: as digital collections are growing and digital preservation is becoming more and more a big issue for collective institutions, social media and the interactive web is becoming more important for the description and access of digital collections.

The online video game platform, which is the main output of the project, works as follows: two participants are presented with the same video at the same time. Every time they use the same tag to describe a part of the video, within a time frame of 10 seconds, they receive points. This method of gathering metadata is also based on the idea of Games with a Purpose (GWAP), where peer-to-peer validation is used to assigned rewards as well as to ensure that the information entered is accurate, based on the theory that if two people, who are not in communication during the game, enter the same tag the probability of the tag to be valid is very high. If no other users were online, people could play against the so-called "bots", which were pre-recorded game sessions of that particular video.

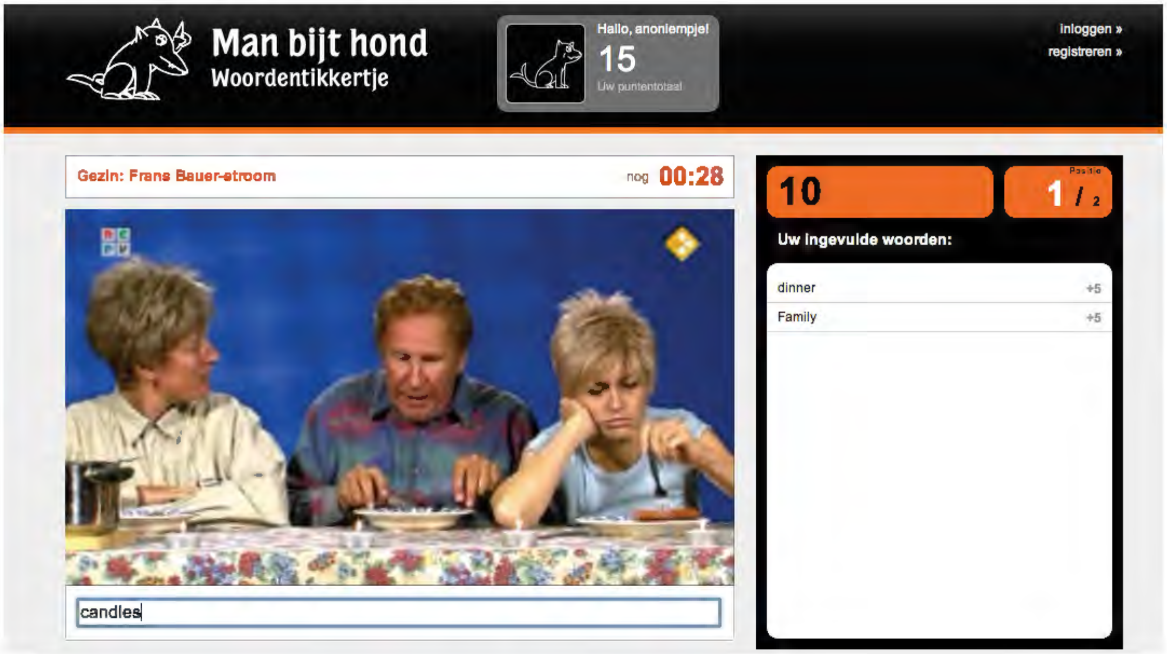


Figure 20. Waisda? *online interface*

The project has been developed in two phases. Phase 1 included the development of the abovementioned platform as well as several activities to evaluate and improve it, such as questionnaires, focus groups, and usability tests to evaluate interface design. Connection with social networks, such as Facebook and Twitter, and other targeted promotional activities such as TV promotions was also used, especially with the KRO collections used for this phase of the project, which allowed gauging audience interest.

The first phase of the project showed interesting results in many different aspects. Several studies were conducted to determine users' behavior, which showed that altruism was an important motivation to participate in the project, as well as the content of the videos, the natural competitiveness of playing a game and having a nice interface to enhance motivations and let players enjoy. They also found that users preferred to play against each other than against the bot and that short videos had better user responses than longer ones (which explains the effectiveness of micro-tasks for the crowdsourcing model).

In terms of qualitative and quantitative results, overall the project collected 340,551 tags, 40% of them were deemed useful (i.e. validated tags through peer-to-peer validation) and 42,000 of them were unique tags.

This phase of the project also focused on the value of the tags gathered, by comparing them to professional thesauri and semantic databases. In this particular case, they compared the tags to the Gemeenschappelijke Thesaurus Audiovisuele Archieven (GTAA, which translates to English as Common Thesaurus Audiovisual Archives) used by professional catalogers at the Institute for Sound and Vision. They also used the Cornetto (Combinatorial and Relational Network as Toolkit for Dutch Language Technology), a lexical semantic database for Dutch language. The study determined that, conservatively - i.e. leaving aside plural versions of a word or conjugations - 5.8% of the tags had matches on GTAA and 23.6% had matches on Cornetto, which shows that many of the words used by players are valid Dutch words that are not present in professional descriptions. Additionally, unique tags are almost non-present in GTAA and Cornetto.

These results were complemented with the evaluation of a professional cataloger, who found a high percentage of useful tags. In their own words:

> *"The senior cataloguer noted that in general the useful tags describe the material in a different way than keywords that cataloguers add do. Firstly because the tags focus on*

> *describing what is seen and heard within a programme, while the professional metadata for audiovisual content focuses on the topical subjects that a programme refers to."*[72]

To further develop the results of the studies showed above, other studies were also focused on the type of words used as tags, whether they described nouns, adjectives or subjects. To do so, they chose 5 videos: Two of the best-tagged, one average-tagged, two low-tagged. The results showed that tags are mostly nouns and adjectives, which means that users describe mostly what they see and hear. The team also studied the similarities between the tags and audio information, by comparing them to the video's subtitle. They found a large overlap of tags with subtitles. Users figured out that by entering information that was already in the audio they would have a higher chance to match the other player's tags, thus getting more points. This led the team to improve the rewards system by assigning more points to the player when tags were not related to audio content, which allows to balance obtaining audio and image description as well as making the game more challenging.

To study and verify the usefulness of tags in terms of searchability, the team applied phrase search in Google.[73] They found that 84% of the unverified tags returned positive hits, the rest were misspellings, garbled text or morphological variations (mixed letters and symbols, slangs, etc.).

Phase 2 of the projects was mainly focused on improving the online platform, using the results obtained in the previous phase as well as to improve the game experience by boosting

[72] Riste Gligorov, Lotte Belice Baltussen, Jacco van Ossenbruggen, Lora Aroyo, Maarten Brinkerink, Johan Oomen, Annelies van Ees, *Towards Integration of End-User Tags with Professional Annotations* http://journal.webscience.org/363/2/websci10_submission_65.pdf

[73] Phrase search is a type of text search in which the user looks for an exact string of words or characters.

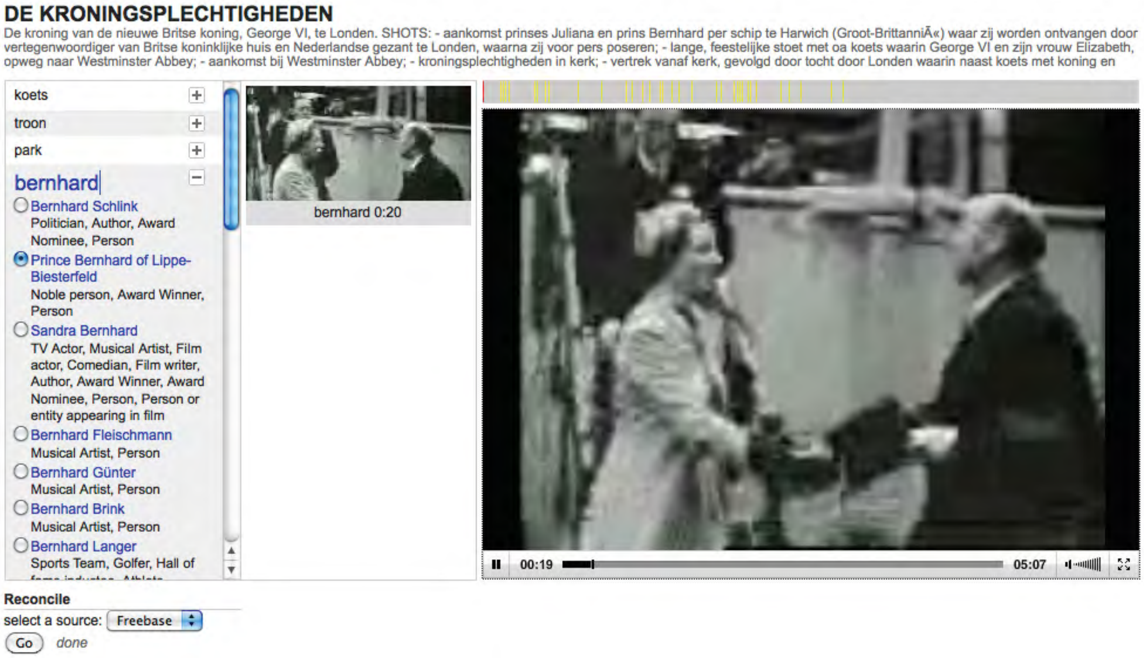
features that would increase users' intrinsic motivations, such as competitiveness. To improve the accuracy of the tags entered – and also the originality of the content provided by the users – tags newly entered (never used in that video before) are rewarded with extra points.

The team also did further work on the study of searchability of tags versus other types of search. The results speak for themselves:

> *"Our findings show that search based on user tags alone outperforms search based on all other metadata types. Combining user tags with the other types of metadata yields an increase in search performance of 33%. We also find that the search performance of user tags steadily increases as more tags are collected."*[74]

The last part of the project, mostly carried out in 2013, was focused on researching the possibilities of applying the semantic web to provide a controlled vocabulary for the tags, as a way to improve their quality – in terms of the different meanings that can be assigned to one word - as well as to connect the project to some Linked Open Data (LOD) projects. The main goal now was to create a semi-automated system to clean tags using the LOD cloud, only as a prototype. Initially they thought of using only Open Refine's reconciliation API - which links to Freebase, to provide controlled vocabulary - but they realized that they needed an embedded player, to provide context for reconciliation. Thus, they built a new reconciliation and search interface, which shows the final reconciliations, the video embedded in the center and information from the reconciled terms/tags. All time-based tags for that video can also be seen in a timeline.

[74] Riste Gligorov et al.



DE KRONINGSPLECHTIGHEDEN

De kroning van de nieuwe Britse koning, George VI, te Londen. SHOTS: - aankomst prinses Juliana en prins Bernhard per schip te Harwich (Groot-BrittanniÃ«) waar zij worden ontvangen door vertegenwoordiger van Britse koninklijke huis en Nederlandse gezant te Londen, waarna zij voor pers poseren; - lange, feestelijke stoet met oa koets waarin George VI en zijn vrouw Elizabeth, opweg naar Westminster Abbey; - aankomst bij Westminster Abbey; - kroningsplechtigheden in kerk; - vertrek vanaf kerk, gevolgd door tocht door Londen waarin naast koets met koning en

koets +
troon +
park +
bernhard −

○ Bernhard Schlink
Politician, Author, Award Nominee, Person

● Prince Bernhard of Lippe-Biesterfeld
Noble person, Award Winner, Person

○ Sandra Bernhard
TV Actor, Musical Artist, Film actor, Comedian, Film writer, Author, Award Winner, Award Nominee, Person, Person or entity appearing in film

○ Bernhard Fleischmann
Musical Artist, Person

○ Bernhard Günter
Musical Artist, Person

○ Bernhard Brink
Musical Artist, Person

○ Bernhard Langer
Sports Team, Golfer, Hall of

Reconcile
select a source: Freebase

Figure 21. Waisda? *Reconciliation Interface*

After studying other Linked Data projects that offered reconciliation services, they determined that Freebase was the best suited because it provides better ranking quality during reconciliation, i.e. more and better suggested terms for each tag.

The interface used SPARQL as communication standard to link with Open Refine's reconciliation API[75]. They also decided to include GTAA and Cornetto. Cornetto was mapped to WordNet, its equivalent in English, to access the English-based Freebase. This was necessary since the *Waisda?* project and all the tags were in Dutch.

The interface allows the user to see the tags and the video, which can be played back at any time. For each tag users select the most suited data set, then the system provides a list of

[75] SPARQL is an RDF query language that allows to retrieve data stored as Resource Description Framework (RDF) format. http://www.w3.org/TR/rdf-sparql-query/

recommended terms to be reconciled. Users then select the best term for that individual tag. This prototype was only tested inside the institution.

The final evaluation showed that local databases (GTAA and Cornetto) provide better results for reconciliation, i.e. provided more accurate terms, however all databases were complimentary to each other. Cornetto was better for subjects and GTAA and Freebase for people and places. Disagreements between participants were very subtle, and terms selected were always related. The team also noticed that users quickly realized which dataset was best for each term, saving time during reconciliation.

Although this project showed that Linked Data can be successfully applied to tagging systems, they still need to test the prototype on a larger scale. They also plan to do more research on the reputation of tag authors and provenance information with the aim of improving search results. Currently, the *Waisda?* code is available online as an open source software on Github.[76]

[76] Waisda?'s code and documentation can be found here https://github.com/beeldengeluid/waisda

7. CONCLUSIONS

Below you'll find my reflections on crowdsourcing metadata inside cultural institutions. I believe that having a distance from the actual projects helped me study the subject in a more objective way. However, because of that same reason these reflections may sound a bit too general for someone who is looking for very straightforward answers or actual implementations. The truth is, from my perspective, reflecting on our actions help us look things differently; it allows us to understand what we've done to project effective solutions.

However, when dealing with systems that are fueled by communities – using models that were created around communities – there are many factors involved, like in any complex social system. We can't expect people to immediately like and use whatever platform we design, but we can design it in a way in which we take advantage of people's behavior online. As Shirky said: "*The trick for creating new social media is to use those lessons to weight the odds in your favor. Rather than as a set of instructions that guarantees success."*

From my point of view, the results shown by the team working on *Waisda?* are very straightforward answers to several of the main doubts around crowdsourcing metadata and the usefulness of tags. They showed that tags can be beneficial for the retrieval of information and content searchability.

That being said, I do believe that there are some important considerations we must embrace - both as a community and institutions individually – to be prepared not only for crowdsourcing projects, but for the impact of social media on cultural organizations.

7.1 BUILDING ONLINE COMMUNITIES

Communities are not there to be taken. You can find groups of people interested in the same topics, issues or ideas. Taking them and creating a community out of them, that is the challenge. Communities must be built over time. One thing is to have many people participating, the other is to make them feel they belong. You do not belong to a community by just clicking on buttons once in a while, you are part of it when you do it with the commitment of contributing to something bigger, with greater goals, with public impact.

Furthermore, people like to see results; they will not continue participating in a project that doesn't seem to be going anywhere, how would that fulfill their intrinsic motivations? In the case of volunteer-based projects, how does the organization create a social environment in which users feel motivated enough to come back? I believe that is one of the reasons why crowdsourcing projects in general tend to have a limited duration. It is true that highly exploratory projects or in the stage of prototype are not considered long-term solutions, but maybe we should start thinking about that possibility, with the creation of public or civic value in mind. A great example that this is possible is *The Commons*; this project has been around for several years, in my opinion because people can enjoy the results immediately (they can access the collections online, tag them and keep using the tags to find things).

In order to develop communities, keeping communication with users and being available to answer questions or comments is key to the success of these projects. In other words, the institution must also belong to the community. In that sense Oates says *"Treat your place like your home: welcome people, fix them a drink and make them feel comfortable. Before you know it, your guests will be chatting amongst themselves, the party will be pumping, and people will be* *making plans together."*[77] This was also noted by the team who developed the *Waisda?* project, in which results were better when they were able to advertise it through television and social networks. In the case of the MCNY, they have built a solid online community through Facebook, where they post images weekly for people to help identifying photographs (although this activity was not connected to the crowdsourcing project described in this document).

Shirky also mentioned that personal motivations and social motivations *"amplify each other in a feedback loop".* If people see a successful online platform, they will want to join it, probably because their friends are also there, and if they see that many people are very motivated about a project, they'll feel motivated too. In the same vein, institutions must be careful to keep the balance between personal and community interest, as well as preferring guidelines instead of rules, acting more like a moderator. However, platforms loose absolute meaning when no rules are presented, remember it is always about taking and giving, therefore users must understand their place in the community and the collaborative nature of it. In solving these problems many site administrators have decided to loosen the rules – normally written as strict legal documents - by presenting *guidelines.* Guidelines are less strict, are written in a simple language and are drawn upon the basis of the crowdsourcing model: mutual benefits. It can be thought like a code of conduct, "don't give to others what you don't like to receive". This principle is a very important part of these online communities, because it has its roots precisely on the feeling of community, of belonging.[78]

[77] George Oates, *Community: From Little Things, Big Things Grow*, May 6, 2008 http://alistapart.com/article/fromlittlethings
[78] Breslin et al. present some interesting guidelines for social behavior online. "Be Careful Before you Post" in *The Social Semantic Web,* p. 42.

This can be hard, especially if the results we want are very specific: gathering accurate metadata. The good news is that, when really committed to a project, crowds can find the way to govern themselves, for the sake of the greater good.

Additionally, I believe that building online community is essential for any social media project, and that is something institutions must cultivate, not only for crowdsourcing but for other online endeavors as well. One thing I believe is key for the success of this projects is to show people the final results: if you ask them to tag collections, show them the tags and how they've been useful for themselves and other members of the community. People want to know that they've been using their free time on something that is worthwhile, and this will bring them back to help with other projects.

7.2 IN SEARCH OF THE PERFECT PLATFORM

Based on complex and ever changing social systems, the design of crowdsourcing platforms is complex as well; there's no recipe. What works for one community might not work for another.

Now, in relation to projects involving crowdsourcing of metadata, there are certain practical recommendations that seem to be working for many implementations. One of them is the use of micro-tasks, which align perfectly with projects based on games or with Linked Data initiatives. Crowdsourcing metadata is already a very straightforward task, but by narrowing it even further the results could be even better and tags could be more and more specific and ultimately more useful for institutions and users.

Micro-tasks offer the opportunity to apply Linked Open Data more effectively, since tasks can be separated according to different datasets, making the process of tagging easier for the user and more effective in terms of the final results. In the MCNY's project the separation of the tasks ended up being very useful to narrow the scope of what workers were doing and ultimately for the implementation of the interface. In my opinion, this sorting should be thought very carefully in other projects, if they were to follow the same model, because it could determine the success or failure of the project.

In the particular case of time-based materials, tasks should be divided in short clips to prevent people from getting bored. Also, dividing tasks also allows people to focus on different areas of interest, levels of complexity, or even different levels of commitment, which would increase the diversity of the community, thus also the diversity of the information provided by them.

In terms of building an attractive platform and engaging users to contribute, games seem to be a good choice, since they naturally trigger intrinsic motivations, even more if it's for a good cause. However, developing interesting games can be very daunting, time-consuming and expensive. The development of open source software, such as the one provided by *Metadata Games* or *Waida?* is a step forward; they are flexible enough to implement in different systems inside different institutions.

There is certainly no doubt that being an active entity online brings notoriety to the institution to their communities, engaging them with the collections, which is enhanced with the participation of volunteers instead of paid workers. After seeing the case of the Museum of the City of New York and interviewing the project manager, Lacy Schutz, I can say that paid workers not necessarily improved the quality of the work since their results were comparable to other projects using volunteers. Lacy even mentioned that they have a quite big community online of people who have some connection with the museum, which even includes historians and specialist whose expertise could be very valuable in crowdsourcing projects. We can say then that paid workers could limit the reach of the project – a main and desirable characteristic of online platforms - since only people registered in these online jobs could have access to participate. In addition, the extrinsic motivation of getting paid can inhibit institutions building communities, which is of course always a desired consequence. Moreover, in my interview with Todd Carter, CEO of Tagasauris, I asked him what was the approach of the company towards volunteers. He answered that they are aware of the differences between paid workers and volunteers and that they were exploring the possibility of finding a way to include volunteers in their online tagging systems for organizations and enterprises.

In summary, I think both options are valid, either inviting volunteers or paying people to participate, as long as there is transparency and a clear objective. For instance, a company may want to tag their pictures to make them more discoverable online, and they can use a paid platform for that. However, paid workers certainly don't contribute to the creation of an online community, which is both desirable for institutions for this type of projects (and other similar social media projects) but also for the institution as a whole. Having a large community could translate into more visitors, either online or on-site, as well as in more resources to develop new projects and strategies: it's ultimately a feedback loop.

7.3 INTEGRATING TAGS

Having analyzed the current state of validation and harvesting systems which demonstrated to be in a very early stage of development, the next question that arises is, how are institutions managing tags if they are not including the information in their cataloging systems?

As I mentioned above, I think the answer to those questions is present in the long lasting projects online, such as *The Commons* in Flickr. This type of projects are gathering the information online and keeping it online. In the same way that people tag items they find items described under the tags presented on the website. The interaction is then happening in both directions simultaneously, users tag and search at the same time. And this is not the only project going in this direction, after all, that is one way of building community. The question is then: do we even want to include tags in our formal descriptive systems if users are already using them?

In the case of combining linked data with crowdsourcing, this problem would be partially solved, since tags can be easily stored as RDF files, where each tag has an associated URI in a particular dataset. In my opinion this is definitely a good solution, since linked data is very rapidly gaining ground, both inside cultural institutions and among for-profit companies. In the mean time, tags can be stored in regular databases and be displayed online as such. The proposal of *Metadata Games* of integrating tags using Dublin Core and MODS is also a great solution if Linked Data is not involved. Ultimately, any solution that allows displaying tags for people to use would be beneficial.

Perhaps we are in the turning point of cataloging systems and this new paradigm will force institutions to rethink the current practices; not adapting new models to the current ones but moving forward towards the creation of systems that are more compatible with online interactions.

7.4 VALIDATION OF TAGS AND LINKED DATA

Regarding the validation of the information, there are three concerns: validation of tags, use of controlled vocabularies and validation of datasets.

For the first problem, platforms designed as games seem to be a good option, especially when using peer-to-peer validation. This could be considered a first layer of validation, together with certain algorithms that help filtering words from non-sense character strings. In other words, algorithms help determining which tags are actually words and the game can help determining if the tag is actually related to the content on screen.

However, even if tags are validated, there are still no controlled vocabularies or a hierarchical structure that allows searching tags more effective. Linked data could have that role, in addition to provide context for the tag, especially in the case of synonyms or homonyms as well as structure and hierarchies.

Validation of datasets can be a concern if using Linked Open Data, especially those data sets that are gathered via crowdsourcing, or built collaboratively such as Freebase or DBpedia (which is derived from Wikipedia). These datasets are permanently changing, which could be an issue in keeping the information updated (or for example the links between tags and the correct URIs). My concern is: is Linked Data mature enough, complete enough, or stable enough? For example, two of the projects studied that used Linked Data (MCNY and *Waida?*) used their own data sets in addition to other publicly available datasets, finding both more useful their own metadata. Is that why these institutions are ingesting their own? Maybe in the future, when public datasets are big enough institutions will not need to contribute with their own and they'll be able to reconcile their terms with existing datasets. In the mean time, and for the sake of contributing with these projects, institutions should make their datasets available for free use.

From a practical perspective, reconciliation apps for time-based media are complicated to provide, especially if tags are associated to a timeline. Additionally, it seems to be very important to provide context for reconciliation and having an interface to do that. Open Refine, for example, could maybe be a solution for reconciliation done by museum staff, because they know the collections and have easy access to them, which eliminates the need of a contextualization tool. However, in many other cases such interface would be needed, limiting the number of institutions that can afford to have specialized people to provide this kind of tools, since there is no interface available online yet.

In the particular case of the *Waisda?* project, I think one of its most remarkable characteristics is the fact that language was not a limitation for the implementation of this hybrid model. It is, no question, a huge advantage to be able to say that this systems can be implemented in many countries, which again reinforces the spirit of Linked Open Data: being able to easily share and reuse information.

7.5 LEAVING THE PROTOTYPE STAGE BEHIND

As I mentioned before, this area is a very new and exploratory field. Yet, some projects are launching free open source prototype versions online with the hope that other institutions would bring their energy and contribute in the improvement of these tools. My recommendation is to download and use free software that these projects are offering. In particular the Metadata Games software is designed to not require that much technical knowledge. I think that large-scale use is the way to leave the prototype stage behind: the platforms available now are not perfect, but using them is the only way to provide feedback to create strong tools suitable for everybody.

APPENDIX 1: More about Linked Open Data

In section 3.3 I covered the basic concepts and structures behind the Semantic Web, Linked Data and Linked Open Data (LOD). However, there's much more to it than that. This appendix intends to complete that basic description. It is, by no means, a comprehensive explanation of the model, but it will give the reader a more general perspective of what Linked Data is beyond its relationship with crowdsourcing.

As I previously mentioned, LOD is Linked Data made publicly available for reuse. There are several LOD projects online – anyone can make their data available – but certainly the biggest ones are Freebase, DBpedia and Europeana, the latter very important in the cultural community.

Freebase, as defined in their website is a *"community-curated database of well-known people, places and things."*[79] It was originally created by Metaweb in 2007, a company later acquired by Google in 2010. Any content contributed to or used from Freebase is under the Creative Commons Attribution (CC-BY) license. The data available on Freebase was originally gathered by the Freebase team from open data sources online. Today, the database, in other words the data itself, can be corrected and provided by anyone as long as they follow their Contribution Guidelines.[80]

Freebase, as a semantic web project is based on triples. However, the organization of the information is different from DBpedia; they have different ontologies, i.e. different way of structuring data and hierarchies. Roughly explained, Freebase stores data using nodes (equivalent to subjects/objects) and edges (equivalent to predicates). Nodes represent people, places and

[79] Freebase website http://www.freebase.com/

[80] Freebase Contribution Guidelines http://wiki.freebase.com/wiki/Contribution_guidelines

things, and some nodes can also be considered topics depending on their importance or the amount of data they connect to. For instance, an artistic movement such as Romanticism or a person like Dalai Lama can be considered topics. In addition, each topic can be assigned a type in case they relate to many definitions. For example, the topic Leonardo da Vinci has several types assigned: painter, sculptor, architect, etc. Types are also grouped into domains, thus the type "sculptor", for instance, can me under the Fine Arts domain.

From the practical point of view, institutions (or the public in general) can have access to the database using either the Freebase APIs (Application Programming Interface) available in RDF format (N-Triples) using the MQL protocol or by downloading the raw data dumps from the website. Data is also easily searchable on the website. To contribute with Freebase the only requirement is to sign up. However, the use of this tool requires previous understanding of how the project works.

DBpedia is a *"crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web."*[81] However, DBpedia is also linked to other data sets online. DBpedia also has a special ontology based on OWL (Web Ontology Language), which includes classes, properties and datatypes. As opposed to Freebase, contributing or editing data to DBpedia appears to be more restricted and controlled. Although the information in Wikipedia can be edited by anyone – which is from where DBpedia takes its data - the data itself can't be edited. In terms of using DBpedia, data sets can be accessed through semantic web browsers (such as LOD Browser Switch or LodLive), the SPARQL endpoint or downloadable RDF dumps.[82]

[81] DBpedia website http://dbpedia.org/About

[82] A comparative table of the characteristics f DBpedia and Freebase can be found here: http://wiki.freebase.com/wiki/DBPedia

Europeana is a project that started with the idea of providing open access to millions of resources from several European institutions through a unique portal. Europeana's Linked Open Data project provides open metadata about all the objects included in this original project. The data sets are available online under CC0 Public Domain Dedication License and under the terms of Europeana's Data Exchange Agreement. Data sets are accessible through a SPARQL endpoint and also as downloadable data dumps. This project, however, is in a pilot stage.[83]

A very important concern arises from this open data projects, which bring the old question of trusted information. How can institutions validate the information available on these data sets? There's no way to do that yet, since the spirit of these projects is actually based on this openness, however, all of them have different levels of control, so that's definitely an option for institutions looking for more controlled metadata.

Another thing to have in mind is that Linked Open Data projects, as open services, provide data that is constantly growing, migrating and changing, Institutions using this service would have to reconcile data often to keep links and information updated. It definitely depends on the level of interaction with the data set, for example only extracting very specific information or connecting crowdsourcing projects to it and also they way in which this data sets are accessed, either by an API, SPARQL endpoint or just using the downloadable raw data.

[83] Europeana Linked Open Data Project http://pro.europeana.eu/linked-open-data

APPENDIX 2: Glossary

Application Programming Interface (API): set of routines, protocols and tools for building software applications. APIs specify how software components should interact with each other.

Cognitive authority: person or entity that influences my thinking in relation to a sphere of interest or knowledge at different degrees. Cognitive authorities involve relationships between at least two people, since it involves certain levels of credibility and trustworthiness

Cornetto: Combinatorial and Relational Network as Toolkit for Dutch Language Technology. Cornetto is a lexical sematic database for Dutch language.

Crowdsourcing: type of participative online activity in which an individual, an institution, a non profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken.

Descriptive metadata: metadata that describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.

Dublin Core: structural metadata schema of fifteen basic elements. Element Set available here: http://dublincore.org/documents/dces/

First-hand knowledge: knowledge acquired through our personal experiences, not only related to formal education.

Folksonomies: a social collaboratively generated, open-ended, evolving and user-driven labeling systems that enable users of social websites to categorize their content using the tag system and to thereby visualize popular tags usages via tag clouds.

Friend of a Friend (FOAF): a machine-readable ontology that provides relationships between people, their activities and their relations to other people and objects.

Games with a purpose (GWAP): concept developed at the School of Computer Science of Carnegie Mellon University through the game *ESP* (a photo tagging game). It is based on the fact that computers still need human aid to perform some tasks, being tagging one of them. However, through games people are asked to perform these tasks without noticing, since they do so by playing and having fun.

Gemeenschappelijke Thesaurus Audiovisuele Archieven (GTAA – Common Thesaurus [for] Audiovisual Archives): Specialized thesaurus used by the Netherlands Institute for Sound and Vision. It contains approximately 160,000 terms related to television.

Keyword search: search algorithm for finding an item that contains the same character string entered by the user.

Linked Data: a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF.

Linked Open Data (LOD): Linked Data that is released under an open license, which does not impede its reuse for free.

Metadata Object Description Schema (MODS): is a schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications. The standard is maintained by the Network Development and MARC Standards Office of the Library of Congress with input from users.

Metaweb Query Language (MQL): is a specialized semantic web query language used to make queries to Freebase using JSON (JavaScript Object Notation). It is analogous to SPARQL.

Micro-tasks: small and straightforward activities separated by subject or type of activity that allow users to focus in short and very direct requests.

Ontology: in the information sciences, ontology represents knowledge as a hierarchy of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts.

Phrase search: type of search that allows users to search for documents containing an exact sentence or phrase opposed to being limited to keywords.

Second-hand knowledge: knowledge acquired from external sources, whether they are people or institutions.

Semantically-Interlinked Online Communities (SIOC): project that provides links between discussion platforms such as blogs, forums, etc. SIOC ontology is an open-standard machine-readable format for expressing the information contained both explicitly and implicitly in Internet discussion methods.

Semantic Web: collaborative movement led by international standards body the World Wide Web Consortium (W3C). The standard promotes common data formats on the World Wide Web. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web, dominated by unstructured and semi-structured documents into a "web of data". The Semantic Web stack builds on the W3C's Resource Description Framework (RDF).

Social metadata: Additional information about a resource resulting from user contributions and online activity —such as tagging, comments, reviews, images, videos, ratings, recommendations —that helps people find, understand, or evaluate the content.

Resource Description Framework (RDF): RDF is both a Model and Syntax Specification (RDFMS), and a Schema Specification (also known as RDF-S). The first one is a syntax model used as the foundation for processing metadata. Its basic data model includes three object types: Resources, Properties and Statements. RDF-S is a mechanism to declare and define the model's properties and resources. In other words, RDFMS is the conceptual framework in which RDF-S is based on. RDF can be expressed in several serialization formats, including Turtle, N-Triples, RDF/XML among others.

Rich Site Summary (RSS): xml-based format used mainly to distribute news headlines online, in other words, provide more information about the content of the website.

SPARQL: RDF query language, that is, a query language for databases, able to retrieve and manipulate data stored in Resource Description Framework format. It was made a standard by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium (W3C), and is recognized as one of the key technologies of the semantic web. On 15 January 2008, SPARQL 1.0 became an official W3C Recommendation and SPARQL 1.1 in March, 2013

Tag: A keyword annotation that acts like a subject or category for the associated content. Tags are normally *free-form keywords,* ideally of only one word.

Tag cloud: The schematic visualization of a group of tags, for example in one particular website. Tag clouds not only allows to see all the tags used, but it also highlights the most popular terms by making them bigger or bolder.

Tag Validation: process in which tags assigned to a resource are verified to determine if they are accurate or useful. It can be done via an algorithm, game design or manual selection.

Triple: basic data entity for semantic web applications, composed by subject, predicate and object.

Triplestore: purpose-built database for the storage and retrieval of triples.

Uniform Resource Identifier (URI): is a unique string of characters used to identify the name of a resource. Such identification enables interaction with representations of the web resource over a network (typically but not necessarily the World Wide Web) using specific protocols. Schemes specifying a concrete syntax and associated protocols define each URI.

Uniform Resource Locator (URL): is a specialization of URI that defines the network location of a specific representation for a given resource.

Reconciliation: semi-automated process in which an application (usually an API) provides a list of suggested terms from a particular Linked Data dataset, which are matched with a particular word in the institution's dataset. Doing this allows linking your data with the data cloud of that particular Linked Data project.

WordNet: is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

BIBLIOGRAPHY

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehman, Richard Cyganiak, Zachary Ives, "DBpedia: A Nucleus for a Web of Open Data" in *6th International Semantic Web Conference 2007* http://www.cis.upenn.edu/~zives/research/dbpedia.pdf

Howard Besser, "Use of Non-Broadcast Channels to Communicate Information In Social Change Situations:Berkeley Anti-Apartheid and Solidarity Poland", Jan. 21, 1986, http://besser.tsoa.nyu.edu/howard/Papers/Poland-berkeley/

Christian Bizer, Tom Heath, Tim Berners-Lee, "Linked Data – The Story so Far", http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

Daren C. Brabham, "Crowdsourcing" (Cambridge, Massachusetts ; London, England : The MIT Press, 2013)

Daren C. Brabham "Crowdsourcing as a Model for Problem-Solving: An Introduction and Cases", 2006, http://www.clickadvisor.com/downloads/Brabham_Crowdsourcing_Problem_Solving.pdf

John G. Breslin, Alexandre Passant, Stefan Decker, *The Social Semantic Web* (Heidelberg ; New York : Springer, 2009)

Maarten Brinkerink, "MS40: Report on Innovative Applications, Second Iteration" http://pro.europeana.eu:9580/documents/866067/983534/MS40+Report+on+innovative+applications,+second+iteration

Nicole J. Caruth and Shelley Bernstein, "Building an Online Community at the Brooklyn Museum: A Timeline", 2007 http://www.archimuse.com/mw2007/papers/caruth/caruth.html

Maria Chatzigiorgaki and Athanassios N. Skodras, "Real-Time Keyframe Extraction Towards Video Content Identification" in *16th International Conference on Digital Signal Processing, 2009*, http://dl.acm.org/citation.cfm?id=1700462

Davide Ceolin, Paul Groth, Willem Robert van Hage, Archana Nottamkandath, and Wan Fokkink, "Trust Evaluation Through User Reputation and Provenance Analysis" in *Proceedings of the 8th Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2012), 11-12 November 2012, Boston, Massachusetts*, (*pp. 15-26*) http://www.academia.edu/2736996/Trust_Evaluation_through_User_Reputation_and_Provenance_Analysis

Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS)", June 2012, http://public.ccsds.org/publications/archive/650x0m2.pdf

Joseph B. Dalton, "Can Structured Metadata Play Nice with Tagging Systems? Parsing New Meanings from Classification-Based Descriptions on Flickr Commons" in *Museums and the Web 2010: The International Conference for Cculture and Heritage On-line http://www.archimuse.com/mw2010/papers/dalton/dalton.html*

Enrique Estellés-Arolas and Fernando González-Ladrón-de-Guevara, "Towards and Integrated Crowdsourcing Definition" on the *Journal of Information Science (JIS)*, Volume XX, pp. 1-14 http://jis.sagepub.com/content/38/2/189.abstract

Mary Flanagan, Sukdith Punjasthitkul, Max Seidman, Geoff Kaufman, Peter Carini "Citizen Archivist at Play: Game Design for Gathering Metadata for Cultural Institutions" in *Proceedings of DiGRA 2013,* Atlanta, Georgia, August 2013 http://www.tiltfactor.org/wp-content/uploads2/tiltfactor_citizenArchivistsAtPlay_digra2013.pdf

Riste Gligorov, Lotte Belice Baltussen, Jacco van Ossenbruggen, Lora Aroyo, Maarten Brinkerink, Johan Oomen, Annelies van Ees, *Towards Integration of End-User Tags with Professional Annotations* http://journal.webscience.org/363/2/websci10_submission_65.pdf

Riste Gligorov, Michiel Hildebrand, Jacco van Ossenbruggen, Lora Aroyo, Guus Schreiber, "An Evaluation of Labelling-Game Data for Video Retrieval" in *Advances in Information Retrieval Lecture Notes in Computer Science,* Volume 7814, 2013, pp 50-61 http://link.springer.com/chapter/10.1007%2F978-3-642-36973-5_5

Riste Gligorov, Michiel Hildebrand, Jacco van Ossenbruggen, Lora Aroyo, Guus Schreiber "On the Role of User-Generated Metadata in Audio Visual Collections" in *K-CAP'11 Proceedings of the Sixth International Conference on Knowledge Capture,* Pages 145-152 http://dl.acm.org/citation.cfm?id=1999702

Scott A. Golder, Bernardo A. Huberman, *The Structure of Tagging Systems*, http://arxiv.org/pdf/cs.dl/0508082.pdf?origin=publication_detail

Michiel Hildebrand and Jacco van Ossenbruggen, "Linking User-Generated Video Annotations to the Web of Data" in Advances in *Multimedia Modeling Lecture Notes in Computer Science,* Volume 7131, 2012, pp 693-704 http://link.springer.com/chapter/10.1007%2F978-3-642-27355-1_74

Michiel Hildebrand, Maarten Brinkering, Riste Gligorov, Martijn van Steenbergen, Johan Huijman, Johan Oomen "Waisda? Video Labeling Game" in *MM'13 Proceedings of the 21st ACM International Conference on Multimedia,* Pages 823-826 http://oai.cwi.nl/oai/asset/22330/22330B.pdf

Rose Holley, "Crowdsourcing and Social Engagement: Potential, Power and Freedom for Libraries and Users", in *Pacific Rim Digital Library Alliance (PRDLA) Annual meeting and Conference: Libraries at the End of the World: Digital Content and Knowledge Creation*, Auckland, New Zealand., 18-20 November 2009 http://eprints.rclis.org/13969/

Rose Holley, “Crowdsourcing: How and Why Should Libraries Do It?” on *D-Lib Magazine*, March/April 2010, Volume 16, Number 3/4
http://www.dlib.org/dlib/march10/holley/03holley.print.html

Jeff Howe, “Crowdsourcing: A Definition”, 2006,
http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html

IDC, “The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East”, Dec. 2012 http://idcdocserv.com/1414

Edith L. M. Law , Luis Von Ahn , Roger B. Dannenberg , Mike Crawford, *TagATune: A Game for Music and Sound Annotation*, http://www.cs.cmu.edu/~elaw/papers/ISMIR2007.pdf

Charles Leadbeater, “The Art of With”, 2009
http://www.cornerhouse.org/wp-content/uploads/old_site//media/Learn/The%20Art%20of%20With.pdf

Tiffany Leason, and the participants in steve.museum, “Steve: The Art Museum Social Tagging Project: A Report of the Tag Contributor Experience”, April 2009,
http://www.archimuse.com/mw2009/papers/leason/leason.html

Fadi Maali, Richard Cyganiak, Vassilios Peristeras, “Re-using Cool URIs: Entity Reconciliation Against LOD Hubs”, in *Proceedings of the Linked Data on the Web Workshop 2011*
http://events.linkeddata.org/ldow2011/papers/ldow2011-paper11-maali.pdf

Guy, Marieke, Tonkin, Emma, “Folksonomies, Tidying up Tags?” in D-Lib Magazine, Volume 12, Number 1, 2006, http://www.dlib.org/dlib/january06/guy/01guy.html

Alan Marsden, Harriet Nock, Adrian Mackenzie, Adam Lindsay, John Coleman, and Greg Kochanski, “Tools for Searching, Annotation and Analysis of Speech, Music, Film and Video: A Survey” in Literary and Linguistic Computing, Vol. 22, No. 4, Oxford University Press, 2007
http://llc.oxfordjournals.org/content/22/4/469.full.pdf?keytype=ref&ijkey=DSLnn4kYw0xKq9W

Museum of the City of New York, “NEH Grant Final Report: Improving Digital Record Annotation Capabilities with Open-sourced Ontologies and Crowd-sourced Workers”, April 30, 2013 https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-51480-11

National Information Standards Organization (NISO), “Understanding Metadata”, 2004
http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Johan Oomen and Lora Aroyo, “Crowdsourcing in the cultural heritage domain: opportunities and challenges”, in *Proceedings of the 5th International Conference on Communities and Technologies*, (New York, NY, USA: ACM, 2011), 138–149.
http://www.iisi.de/fileadmin/IISI/upload/2011/p138_oomen.pdf

Trevor Owens, “The Metadata Games Crowdsourcing Toolset for Libraries & Archives: An Interview with Mary Flanagan” in *The Signal*, April 3, 2013 http://blogs.loc.gov/digitalpreservation/2013/04/the-metadata-games-crowdsourcing-toolset-for-libraries-archives-an-interview-with-mary-flanagan/

Clay Shirky, “Cognitive Surplus” (New York : Penguin Press, 2010)

Karen Smith-Yoshimura and Cindy Shein, *Social Metadata for Libraries, Archives and Museums*, Parts 1, 2 and 3 OCLC, Sept. 2011.
http://www.oclc.org/content/dam/research/publications/library/2011/2011-02.pdf?urlm=162950
http://www.oclc.org/content/dam/research/publications/library/2011/2011-03.pdf?urlm=162953
https://www.oclc.org/content/dam/research/publications/library/2012/2012-01.pdf

Society of American Archivists, *Orphan Works: Statement of Best Practices,* 2009, available at http://www.archivists.org/standards/OWBP-V4.pdf

Louise F. Spiteri, “Structure and Form of Folksonomy Tags: The road to the Public Library Catalogue” in *Information Technology and Libraries*, September 2007, http://www.ala.org/lita/ital/sites/ala.org.lita.ital/files/content/26/3/spiteri.pdf

Roelof van Zwol, Lluis Garcia Pueyo, Georgina Ramirez, Börkur Sigurbjörnsson, Marcos Labad, “Video Tag Game”, http://research.yahoo.com/files/vanzwol-vtg.pdf

Various Authors, “Emerging Practices in the Cultural Heritage Domain, Social Tagging of Audiovisual Heritage” in *Proceedings of the WebSci 10: Extending the Frontiers of Society On-Line (*Raleigh, NC: US April 26-27th, 2010)
http://journal.webscience.org/337/2/websci10_submission_23.pdf

Various Authors, *For the Common Good: The Library of Congress Flickr Pilot Project,* 2008, http://www.loc.gov/rr/print/flickr_report_final_summary.pdf

Luis von Ahn, Laura Dabbish, “Designing Games With a Purpose” in *Communications of the ACM*, Aug. 2008, Vol. 51, No. 8
https://www.cs.cmu.edu/~biglou/GWAP_CACM.pdf

Luis von Ahn, Laura Dabbish, “Labeling Images with a Computer Game” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 2004
https://www.cs.cmu.edu/~biglou/ESP.pdf

The W3C Media Fragments URI Standard is a series of syntax specifications for the construction of media fragment URIs and their use in the HTTP protocol. “Media Fragments URI 1.0 (Basic)”, W3C Recommendation, Sept. 2012
http://www.w3.org/TR/media-frags/

Patrick Wilson, *Second-hand Knowledge: An Inquiry into Cognitive Authority*, Westport, Conn. : Greenwood Press, 1983.

WEBOGRAPHY

Matt Welsh, "What life was before the Web", Feb 24, 2011, http://matt-welsh.blogspot.com/2011/02/what-life-was-like-before-web.html

Chris Lacinak, "The End of Analog Media: The Cost of Inaction and What You Can Do About It", Nov. 8, 2013 http://www.avpreserve.com/wp-content/uploads/2013/11/Lacinak_COI_AMIA_2013_dist.pdf

Trevor Owens, "The Crowd and the Library" http://www.trevorowens.org/2012/05/the-crowd-and-the-library/

Wikipedia
https://en.wikipedia.org/wiki/Wikipedia:About

Europeana Linked Open Data Project
http://pro.europeana.eu/linked-open-data

GeoNames, Geographical Database
http://www.geonames.org/

The Commons' website in Flickr
http://www.flickr.com/commons. Library of Congress,

Trove Website
http://trove.nla.gov.au/

Wikimedia
http://wikimediafoundation.org/wiki/FAQ/en

Galaxy Zoo
http://www.galaxyzoo.org/

Steve.museum Project
http://www.steve.museum/

Amazon's Mechanical Turk
https://www.mturk.com/mturk/welcome

Tagasauris
http://www.tagasauris.com/

New York Public Library "What's on the Menu?"
http://menus.nypl.org/

Brooklyn Museum “Tag! You’re It!”
http://www.brooklynmuseum.org/opencollection/tag_game/start.php

Brooklyn Museum “Click! A Crowd-Curated Exhibition”
https://www.brooklynmuseum.org/exhibitions/click/

The National Archives UK “Operation War Diary”
http://www.operationwardiary.org/

Flickr
http://www.flickr.com/about/

W3C, “What is Linked Data?”
http://www.w3.org/standards/semanticweb/data

Tim Berners-Lee, “Is Your Linked Open Data 5 Star?”, 2009
http://www.w3.org/DesignIssues/LinkedData.html

Dean Stringer’s blog post about the workshop *Sustaining Creative Communities* in New Zealand.
http://librarytechnz.natlib.govt.nz/2009/03/designing-sustaining-creative.html

Edward L. Deci’s Self-determination theory (SDT)
http://www.selfdeterminationtheory.org/

Oates, George, “Community: From Little Things, Big Things Grow”, 2008,
http://alistapart.com/article/fromlittlethings

Instructional Videos for the MCNY and Tagasauris Project
http://instructions.tagasauris.com/

Open Refine
http://openrefine.org/

Creative Commons Licenses
http://creativecommons.org/licenses/

Tiltfactor Laboratory
http://www.tiltfactor.org/

VideoTag
http://www.videotag.co.uk/

Waisda?
http://woordentikkertje.manbijthond.nl/