Digital Preservation
Pamela Vizner Oyarce
December 13, 2013

Improving Crowdsourced Metadata Through Linked Open Data, Case Study: Museum of the City of New York and Institute for Sound and Vision Netherlands

1. Introduction

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 some for-profits as well) realized its potential and implemented several projects based on this now popular model.

The advantage of crowdsourcing applied to cultural institutions is primarily the possibility to identify and describe content that otherwise would be impossible to catalog because of time constraints, lack of resources or lack of staff. Furthermore, in the digital era, creation of content has increased exponentially, enlarging the already huge backlog that collective 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 to the collections.

However, crowdsourced metadata - regardless of the type of media it describes - also 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, locally and online. In the same vein, retrieval of crowdsourced information is clearly an unsolved issue.

Many institutions and other collaborative projects, such as the Library of Congress, Europeana and GeoNames [1], have been using Linked Open Data, 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 paper, 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 and quite unexplored match. For collecting institutions it offers the opportunity to solve the main problems with multiple tags, controlled vocabularies and making objects searchable on catalogs.

This paper would focus on the relationship of crowdsourcing and Linked Open Data through the case study of two institutions that have already applied the hybrid model to gather metadata for audiovisual collections: the Museum of the City of New York in collaboration with Tagasauris, and Waisda?, the project of the Netherlands Institute for Sound and Vision. Both projects represent and interesting contrast; the first one tackled crowdsourcing and Linked Open Data together. The second one first started with a crowdsourcing project and Linked Open Data was introduced to solve a very specific problem. The approaches, though the methodologies where different due

to the initial starting points, were quite similar. Both projects are in the stage of a finalized first prototype.

This paper will provide a very general overview of both models in order to understand the methodologies of the studied projects. I will then describe both of them to finally provide a comparative analysis of the results. Since the conjunction of both models is still a very new idea, the aim of this paper is to provide a broad visualization of the effectiveness of the model, if it can be improved and under what circumstances can be used, specifically addressing the issues with audiovisual collections.

2. Crowdsourcing and Linked Data: Basics

2.1 What is crowdsourcing?

The highly interactive and participative web has become a hub for the development of open projects with the participation of the community since the beginning of the 21st century. The exponential growth, broad scope and variety of collaborative projects online have made it difficult to stop, observe and evaluate the phenomenon. Defining and even classifying these models is not easy. Crowdsourcing was not the exception.

The first definition of crowdsourcing appeared in 2006 on Wire Magazine in an article written by Jeff Howe titled "The Rise of Crowdsourcing". [2] According to Howe, crowdsourcing is "*…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."* A second definition, which I personally like better because it encompasses the broadness of crowdsourcing, was given by Daren Brabham also in 2006: *"Crowdsourcing is an online, distributed problem-solving and production model."* [3].

The truth is, the model already existed and some companies, and even cultural institutions were already experimenting with it [4] having different results that have helped improving the model by providing information to study implementation issues and evaluate its effectiveness.

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."* [5]

Furthermore, non-profit cultural organizations can benefit from online mass behavior, which rewards the openness and transparency of these projects, the sense of community they create, the feeling of contributing to a higher endeavor, the trust people have on cultural organizations and the possibility of using and enjoying the information in the future. For the institutions it is a great opportunity to engage people with the institution, to introduce collections to the community and to tackle the never-ending cataloging backlog to finally make collections available for use.

However, the use of this model is not exempt of problems, which are mostly related to its implementation. Focusing primarily in the implementation of crowdsourcing projects to describe audiovisual collections many will agree that there are several practical issues to solve, especially regarding the storage of data and how we can make it useful in the future. 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. In their defense, we can all agree that these results were somewhat expected since many of these projects were exploratory. First of all, lessons learned tell us that defining very clear and discrete tasks for volunteers is key in the success of the project. Other solution is to implement games as a way of validating information. 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. If I see a fruit dish with apples I can type "apple", "apples", "red", "fruit" and even "healthy" or "crunchy". All of them are right, but are they all useful?

Once again the interactive web is giving us a chance to tackle this problem with a crowdsourced resource: Linked Open Data.

2.2 What is Linked Data?

To explain what Linked Data is we have to go back to the definition of the semantic web. The semantic web is an idea created by Tim Berners-Lee and started by the World Wide Web Consortium to transform the "web of content and documents" in the "web of data". Up until now, websites are a set of documents stored on a server and displayed in a certain way. 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. Now, what is the problem with it? Well, the web of content doesn't allow machines to understand the language in order to improve the use of the web. The Semantic Web allows *"large scale integration of, and reasoning on, data on the Web"* [6] 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, in a way that 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 data must be structured and linked in a very particular way. Linked Data's basic structures are called 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 (see Figure 1). This is the fundamental principle of Linked Data, something like a huge relational database. [7]

Triples are stored in data sets, most frequently using the Resource Description Framework (RDF) data model [8], which is a *"family of international standards for*

data interchange", although this is not the only standard used. [9] Many different data sets and the relationship between them can be represented using multi-graphs (see Figure 2) and data sets can be shared and exchanged using the data query protocol SPARQL.

Now, how are things linked? Each element in a triple ideally has a Uniform Resource Identifier (URI) that may or may not link to a website. [10] Those elements can also be plain text values, called literal values. 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.

Each Linked Data project, although based in the triplestore formula explained above, has its own organizational structure, its ontology. As defined by the W3C an ontology describes *"the types of things that exist (classes), the relationships between them (properties) and the logical ways those classes and can be used together (axioms)."* [11] I will explain it further on the next section.

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 data sets. 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 data sets available for free use, however, true Linked Open Data must be linked to other people's data to provide context. [12] This aspect is key for the success of the model; everyone uses everyone else's data, avoiding redundancy.

Institutions can make their data sets available online or they can include their data sets in other bigger Linked Open Data projects, such as Freebase, DBpedia or Europeana, which are explained below.

2.3 LOD projects: Freebase, DBpedia and Europeana

Freebase, DBpedia and Europeana are not the only Linked Open Data projects online [13]. However, I will focus on them for two reasons. First, these are the biggest ones and second, because both projects studied for this paper either used or considered using their data sets. The goal of this section is to determine the main differences between them to understand why and how they were used in both cases.

Freebase, as defined in their website is a *"community-curated database of well-known people, places and things"* [14]. 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. [15]

The data available on Freebase was originally gathered by the Freebase team from open data sources online. Today, the database can be corrected by anyone and data can be provided by anyone as long as they follow their Contribution Guidelines. [16]

Freebase, as a semantic web project is based on triples. However, the organization of the information is a little more complex than just storing triples on RDF files; this is its ontology. Roughly explained, Freebase stores data using nodes (explained above as subject/object) and edges (predicate). Nodes represent people, places and 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 for RDF - using the SPARQL protocol - ,MQL [17] 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."* However, DBpedia is also linked to other data sets online. [18] DBpedia also has a special ontology, which includes classes, statements and properties (http://dbpedia.org/Ontology). As opposed to Freebase, contributing or editing data to DBpedia appears to be more restricted and controlled. Contributions are more directed to improving the service rather than providing content. In terms of using DBpedia, data sets can be accessed through semantic web browsers [19], using the SPARQL protocol or downloadable RDF dumps.

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 [20]. Data sets are accessible through a SPARQL endpoint and also as downloadable data dumps. This project, however, is in a pilot stage.

Another fundamental service provided by Linked Data projects is the one related to establishing a relationship between already existing data and public data sets. This process is called reconciliation. Reconciliation services basically *"allow third-party publishers to link their data to LOD hubs as part of the data publishing process."* [21] Reconciliation is a semi-automated process in which the application provides a list of suggested terms from a particular data set that can potentially be matched with a particular tag. Some of them allow selecting the data set used for the process, which can be very useful to narrow the number of terms presented for reconciliation. Freebase provides a reconciliation service based on an API, DBpedia

only provides this service through SPARQL and Europeana doesn't provide this service.

Finally, using Freebase or DBpedia will only depend on the type of project and the technical capabilities or constraints of the project. It is worth noting that DBpedia, because of its contribution restrictions, could be more reliable in terms of quality of content. In that sense, Europeana can also be considered more reliable since only cultural institutions can contribute with metadata. In addition, DBpedia is interliked to many other data sets online - including Freebase - being a more comprehensive project. Another advantage of DBpedia is that it has data sets published in many languages, therefore expanding the possible users.

3. *Waisda?*: Cleaning Tags using LOD

Implementing tag systems to gather descriptive metadata for video elements is a particular daunting task. 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.

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. In a joined effort they developed an online platform called Waisda?, a crowdsourcing project for audiovisual tagging based on a game [22]. This project is not only interesting for the novelty of tagging videos but also because it involved a lot of research about the validation of the information 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.

Here is how it works: two participants are presented with the same video at the same time. Every time they use the same tag, in a time frame of ten seconds, for describing a part of the video they receive points. 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) and tags only related to image are rewarded with extra points. This last rule was implemented after discovering that people realized that using tags related to the audio content would increase the chances of coinciding with other participants, thus losing the balance between tags related to image and audio.

However, all their efforts to improve the platform only using the web-based application and with only the help of volunteers was not enough to collect good quality tags. Content was still vague, subject to multiple interpretations or even incomplete. Issues such as folksonomies [23] were still a problem. Furthermore, in terms of search, tags were not very useful since they didn't follow a controlled

vocabulary. On the bright side, tags were good because they were time-based as opposed to the professional generated metadata which was at the item level, giving an in depth description of the content. The potential was there, but tags had to be cleaned. This is when they thought of Linked Open Data.

The main goal of the project 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 providing 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.

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 and Sindice [24] as communication standards to link with Open Refine's reconciliation API. They also decided to include two data sets from Europeana: the GTAA and Cornetto. GTAA is the data set of the Institute for Sound and Vision's Thesaurus. Cornetto is a semantic database in Dutch, which 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. [25]

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 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) threw better results, 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 data set was best for each term, saving time during reconciliation.

4. Museum of the City of New York and Tagasauris: Collecting Tags Using LOD

The Museum of the City of New York, 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 both models: crowdsourcing and Linked Open Data. [26] 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 needed more basic information about each photograph, such as

number of people, horizontal or vertical, nigh or day, etc. in order to provide straightforward sorting information to patrons. Unfortunately, the institution was unable to hire more catalogers to do this because of space and budget constraints.

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. This would allow avoiding the repetition of some entities as well as to contribute 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 digital asset management system. [27]

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. This initial sorting was key to the results of the project, since every task was associated with a determined data set on Freebase, decreasing the chance of error by the workers.

Each worker, then, would choose from all this 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.

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. This resulted to be a very good way to ensure the effectiveness of the model.

In order 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 workers as well as compare crowdsourced metadata with professional metadata previously recorded by the museum´s staff.

The results, in terms of quantity, were somehow expectable: online workers provided more tags per photograph than professional catalogers. To assess the performance of the project in term of quality the museum developed a model based on Panofky-Shatform matrices [28], a hybrid model that basically divides the tags in two levels of complexity: generic/specific/abstract and who/what,/when/where. In this sense, 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. This evaluation also showed that professional catalogers can sometimes provided more complete information, because they have knowledge of the background of the collections.

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. 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. This issue has been considered by the institutions involved in the project, but there has been a lot of discussion regarding the way in which this information is displayed, since for the Museum validation of the crowdsourced information is vital before making it public.

5. Conclusions

After this very general study of the use of crowdsourcing and Linked Open Data for the description of audiovisual collections by two institutions, I have the following comments. I called them reflections because I believe that it is not possible to draw any final conclusions after only studying a couple of examples and also because the implementation of this hybrid projects is very new. Taking that in consideration, I can say that the major problems are still related to the quality of the tags. When information is gathered without much control the resulting tags tend to be messy. Problems such as folksonomies arise making very difficult for museum professionals to use tags in their catalogs or event available for end users. Using Linked Open Data presents a possible solution for the problem of uniformity, controlled vocabulary and ambiguity. However, the examples examined in this study tell us that there are many things to consider before implementing these projects.

First, providing a structure and clear instructions for workers or volunteers seems to be key for the success of the project and the quality of tags. But also giving them enough freedom to enjoy their job and let them decide where to focus. Providing a game environment can be also a solution. This offers the opportunity to apply Linked Open Data more effectively, since tasks can be separated according to different data sets, making the process easier for the user and more effective in terms of the final results. In the MCNY's project, although the separation of the micro-tasks were very visually-based, they ended 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.

Now, when it comes to the evaluation of the use of public data sets, the question that lingers in my mind is, what are the benefits of using Linked Open Data over other controlled vocabularies? Both projects studied used their own data sets in addition to other publicly available data sets, finding both more useful their own metadata. My personal concerns in relation to that issue are related to the exploratory stage of Linked Data projects among cultural institutions. Are publicly available data sets mature enough to provide the required information to describe cultural elements? Is that why these institutions are ingesting their own? Maybe in the future,

when public data sets are big enough institutions will not need to contribute with their own and they'll be able to reconcile their terms with existing data sets. In the mean time, and for the sake of contributing with these projects, institutions should make their data sets available for free use.

Another unanswered question is the one related to the openness of projects such as Freebase and DBpedia, which brings the old question of trusted information, a concern already raised by Wikipedia years ago. 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.

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 again limiting the number of institutions that can afford to have IT 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.

Although discussing the impact of these project on the communities and the evaluation of the participation of volunteers versus paid workers was out of the scope of this paper, I would still like to dedicate some lines to that issue. After all the projects I studied were different in that aspect. 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.

Finally, it is great to see how these projects keep evolving, how obstacles are being circumvented and how technology brings answers to problems we thought had no solution. I hope institutions keep experimenting, there's no question that

crowdsourcing and Linked Open Data have a huge potential for cultural institutions, which is evident even in this very early stage of development.

Footnotes

[1] More information about these projects on their respective websites: Library of Congress http://id.loc.gov/, Europeana http://pro.europeana.eu/linked-open-data and GeoNames http://www.geonames.org/.

[2] The original article can be accessed on the magazine´s website. Subscription required. A later article by Howe about crowdsourcing can be found on his website: http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html

[3] Brabham, Daren C.

[4] Some examples: the first online implementation of Wikipedia was in January of 2001. In the cultural world one of the first examples of crowdsourcing was the project *The Commons* created by the Library of Congress using Flickr, which started in 2006.

[5] Holley, Rose.

[6] *What is Linked Data?* on the website of the World Wide Web Consortium, available at http://www.w3.org/standards/semanticweb/data

[7] The names of the elements of a triple can change according to different ontologies. For example the element *resources* can also be fund as *names* or *entities*.

[8] More information can be found on *Quick Intro to RDF* on this website: http://www.rdfabout.com/quickintro.xpd

[9] Definition of RDF, found here http://www.w3.org/TR/ld-glossary/#resource-description-framework-rdf. Additionally, these data sets can be expressed in XML format or simply .rdf or N3.

[10] A URI that uses the HTTP protocol to retrieve the description of the resource id known as HTTP URI or Dereferenceable URI. Definition available on the W3C Linked Data Glossary available at http://www.w3.org/TR/ld-glossary/#dereferenceable-uris

[11] http://www.w3.org/TR/ld-glossary/

[12] *Is your Linked Open Data 5 Star?*, Tim-Berners Lee, available here http://www.w3.org/DesignIssues/LinkedData.html

[13] More information on the projects' websites: DBpedia http://dbpedia.org/About Europeana http://pro.europeana.eu/linked-open-data

[14] Definition available on the project's website http://www.freebase.com/

[15] Creative Commons Attribution License
http://creativecommons.org/licenses/by/2.5/

[16] Freebase, Contribution Guidelines, available here
http://wiki.freebase.com/wiki/Contribution_guidelines

[17] Metaweb Query Language (MQL is an API developed by Freebase which uses the Java Script Object Notation, Java Script Object Notation (JSON) protocol.

[18] All data sets interlinked with DBpedia and its multi-graph can be seen here
http://wiki.dbpedia.org/Interlinking

[19] A semantic browser is a tools that usually works without the necessity of a local browser, they retrieve the data on the Web directly, dereferencing the URIs.
http://www.w3.org/2001/sw/wiki/Category:Semantic_Web_Browser

[20] License and Exchange Agreement available here:
http://creativecommons.org/publicdomain/zero/1.0/
http://pro.europeana.eu/support-for-open-data

[21] Maali, Cyganiak, Peristeras.

[22] Waisda? Website http://woordentikkertje.manbijthond.nl/

[23] Folksonomy is a system of classification created by collectively assigning tags to annotate content. They can represent individuals and communities.

[24] Sindice is an index for semantic web documents, definition and features available at http://www.w3.org/2001/sw/wiki/Sindice

[25] GTAA translates to English as Common Thesaurus for Audiovisual Archives. Cornetto stands for Combinatorial and Relational Network as Toolkit for Dutch Language Technology and it is a lexical semantic database.

[26] More information about the institutions on their websites
http://www.mcny.org/ and http://www.tagasauris.com/

[27] More information about the digital asset management system here
http://www.orangelogic.com/

[29] Schutz, Lacy.

Figures

Figure 1: RDF as a graph. It shows the relationships (predicates) between objects and subjects. Image by Joshua Taubere, available at http://www.rdfabout.com/intro/

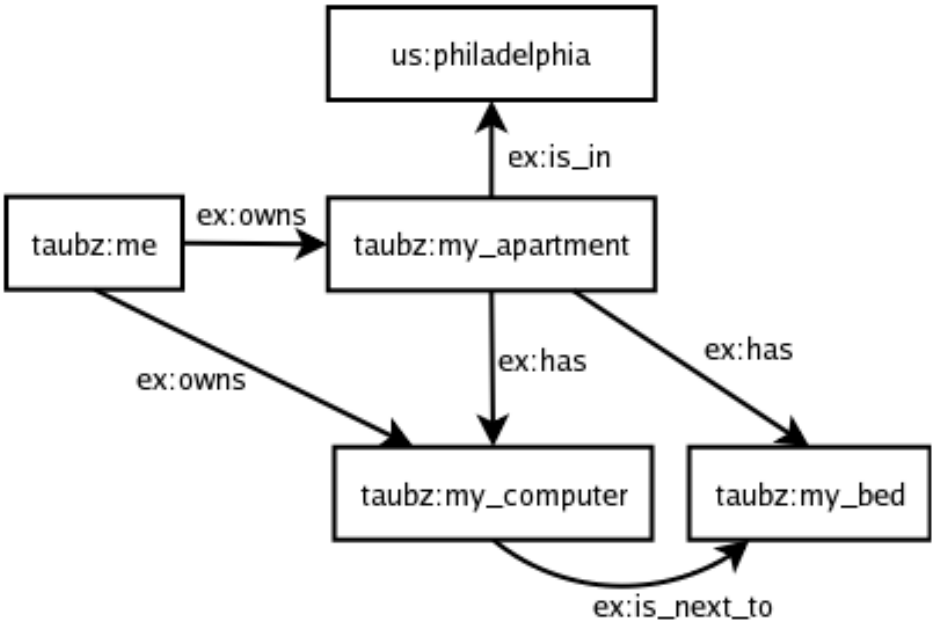

Figure 2: RDF graph using URIs to describe Eric Miller, available at http://www.w3.org/TR/rdf-primer/

Bibliography/Webography

All sites last accessed on December 13th, 2013.

Howe, Jeff, *Crowdsourcing: A Definition*, 2006, available at http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html

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

Holley, Rose, *Crowdsourcing: How and Why Should Libraries Do It?*, 2010, available at http://www.dlib.org/dlib/march10/holley/03holley.html

World Wide Web Consortium (W3C), *What is Linked Data?*, available at http://www.w3.org/standards/semanticweb/data

Tauberer, Joshua, *rdf:about,* available at http://www.rdfabout.com

Maali Fadi, Cyganiak, Richard, Peristeras, Vassilios, *Re-using Cool URIs: Entity Reconciliation Against LOD Hubs*, 2011, available at http://events.linkeddata.org/ldow2011/papers/ldow2011-paper11-maali.pdf

Tomlinson, John, *Crowdsourcing & Linked Open Data: New ways to make collections visible*, 2011, available at http://mysite.pratt.edu/~sla/events/2011crowdsourcingandlinkeddata.html

Hildebrand, Michiel, van Ossenbruggen, Jacco, *Linking user-generated video annotations to the web of data*, 2012, available at http://www.few.vu.nl/~michielh/resources/mmm2012.pdf

Schutz, Lacy, *White Paper Report, NEH Grant, Improving Digital Record Annotation Capabilities with Open-sourced Ontologies and Crowd-sourced Workers*, 2013, available at https://securegrants.neh.gov/PublicQuery/main.aspx?q=1&d=0&f=0&p=1&pv=247&s=0&y=0&n=0&o=0&t=0&ob=year&or=DESC