A-posteriori provenance-enabled linking of publications and datasets via crowdsourcing

Laura Dragan[1], Markus Luczak-Roesch[1], Bettina Berendt[2],

Elena Simperl[1], Heather Packer[1], Luc Moreau[1]

[1] University of Southampton, UK, [2] KU Leuven, BE

Abstract

This paper aims to share with the digital library community different opportunities to leverage crowdsourcing for a-posteriori capturing of dataset citation graphs. We describe a practical approach, which exploits one possible crowdsourcing technique to collect these graphs from domain experts and proposes their publication as Linked Data using the W3C PROV standard. Based on our findings from a study we ran during the USEWOD 2014 workshop, we propose a semi-automatic approach that generates metadata by leveraging information extraction as an additional step to crowdsourcing, to generate high-quality data citation graphs. Furthermore, we consider the design implications on our crowdsourcing approach when non-expert participants are involved in the process.

1. Introduction

The need to treat datasets used in research as "first-class citizens" of the scientific process is by now recognised by many disciplines. Many standard citation guidelines have been enriched by templates for data publication and citation. For example, Nature's Scientific Data[1] now has platform for the publication of metadata about datasets. Thus, in principle, readers of scientific publications can consult cited datasets and form reasoned opinions on the quality of the use of data, similar to the opinions that they can form on the quality of the use of cited publications. In addition, bibliometric statistics and algorithms can be applied to trace and evaluate data citation and to derive useful analytics.

Datasets can be composites of other datasets, thus requiring them to have their own form of "citing". A dataset can be derived from, a subset, aggregate, or a new version, of other datasets. The combination of the metadata of scientific publications, the metadata of datasets, citation links from publications to datasets, and version information, can provide a rich dataset for analytics. However, while conceptually simple, such information spaces have remained a remote vision so far.

[1] "Scientific Data is an open-access, online-only publication for descriptions of scientifically valuable datasets, and exists to help you publish, discover and reuse research data." http://blogs.nature.com/scientificdata/2013/07/23/scientific-data-to-complement-and-promote-public-data-repositories/

This paper aims to share with the digital library community different opportunities to leverage crowdsourcing for a-posteriori[2] capturing of dataset citation graphs. We describe a practical approach, which exploits one possible crowdsourcing technique to collect these graphs from domain experts. For the representation of the citation graphs we apply provenance modeling as currently investigated and recommended by the W3C PROV working group, and they are published conforming to the Linked Data principles. We support both types of information described above: the relationship between publications and datasets, as well as between different datasets or different versions of one dataset.

We present an approach to link datasets used in research to their publications and describe two use cases: the DBpedia[3] Linked Open Data dataset, and the USEWOD[4] log file dataset. Then we report on a user study which was run during the USEWOD workshop simulating crowdsourcing with a group of experts. Based on our findings, we propose a semi-automatic approach that produces metadata by leveraging information extraction as an additional step to crowdsourcing, to generate high-quality data citation graphs. Furthermore, we consider the design implications on our crowdsourcing approach when non-expert are used to provide information.

Our approach defines two types of relationships between dataset references from publications: a general relationship of the form “this dataset (version) is used by this publication”; and a specialised set of relationships which provide more information about how the dataset or its version actually contributes to the publication. We argue that the enriched set of metadata, compared to the general case, can provide insight about the role of data used in the scientific process. Therefore, it can be used to better analyse a publications scientific output. However, simple links between the publications and the accurate version of the datasets used is critical in evaluating and reproducing research.

2. Background and related work

Many organisations have identified principles and rules for data citation. Force11 (Bourne, 1) declared a Joint Declaration of Data Citation Principles, which cover purpose, function and attributes of citations. “These principles recognise the dual necessity of creating citation practices that are both human understandable and machine-actionable”[5]. They identify eight principles declaring data citation importance, credit and attribution, evidence, unique identification, access, persistence, specificity and verifiability, and their interoperability and flexibility. Michigan State University outline citation elements that need to be considered when publishing data citations[6].

[2] A-posteriori, because the capture of the information is done after the publications were written and published, as opposed to being made explicit during the writing process.
[3] http://dbpedia.org/About
[4] http://usewod.org/data-sets.html
[5] Force11 Joint Declaration of Data Citation Principles: https://www.force11.org/datacitation
[6] Michigan State Univeristy, How to Cite Data: http://libguides.lib.msu.edu/citedata

Established citation styles such as APA, MLA or Chicago[7] and Harvard[8] now all contain guidelines for citing datasets, all agreeing on the necessity of a core set of metadata to describe a dataset: author or creator, date of publication, title or description, publisher (the entity, an organization, database, archive, or journal, that is responsible for hosting the data), and URL or DOI, i.e. an identifier and locator that are, if possible persistent. Certain styles also ask for additional information such as edition or version, date accessed online, and a format description such as: data file, database, CD-ROM, computer software. Some e-Science initiatives such as OpenML[9] (a database of machine-learning experiments, for which datasets play a pivotal role) go beyond this by supplying exportable metadata records for datasets and additional options. For example, scientists who publish/upload a dataset can indicate that they want people to cite a specific paper if they reuse the data, e.g. with a CC-BY attribution licence.

The most common approach for data citation is to use the well-established Digital Object Identifier (DOI), which is a URI that contains metadata about the dataset or the dataset itself. A DOI is a type of persistent identifier that indicates a dataset will be well managed and accessible for long term use. It is now routine practice for publishers to assign DOIs to journal articles and for authors to include them in article citations. DataCite[10] and Cite My Data Service[11] provide DOIs for researchers to reference in their work. Their aim is to provide and support data citations in publications by providing datasets and their metadata to researchers.

The use of vocabularies to describe data citation supports many of the guidelines and principles mentioned above. Some vocabularies are specifically designed for data citations, while there are others that are powerful enough and extensible for this purpose. These include:

- Semantic Publishing And Referencing (SPAR) (Shotton,1), which consists of eight core ontologies that describe bibliographic records, their citations and the relationships between the records and citations. In particular, the CiTO (Peroni, 33) defines object properties for citations, and includes some properties which describe how data is used in publications such as the properties: is discussed by, cites as evidence, and uses data from. This ontology can be used independently to the other SPAR ontologies because it doesn't not use restrictions of domains and ranges.
- schema.org[12] provides a collection of schemas that can be used to mark up structured data. These schemas can be reused to mark up data citations, the Creative Work vocabulary defines markup for citations and datasets, and can be used to describe links to data. The vocabulary is broad and there is limited number of concepts that can be reused. The vocabularies main purpose is to mark up html so that popular online search engines can recognise them when providing users with results.

[7] e.g. http://libguides.lib.msu.edu/content.php?pid=120322
[8] e.g. http://guides.is.uwa.edu.au/content.php?pid=43218
[9] http://openml.liacs.nl/
[10] DataCite: http://www.datacite.org/
[11] Cite My Data Service: http://www.ands.org.au/services/cite-my-data.html
[12] schema.org: https://schema.org/

- The PROV-DM[13] can be used to record provenance information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. The PROV-DM can be used to capture data citations, and can be extended to describe more specific connections between datasets and publications.
- VoID[14] is an RDF Schema vocabulary for expressing metadata about RDF datasets. Its main purpose is to allow publishers and users of RDF data to express metadata about applications ranging from data discovery to cataloging and archiving of datasets. It is based on the Dublin Core vocabulary and describes access metadata, structural metadata, and links between datasets.
- The RDF Data Cube vocabulary[15] describes statistical datasets, observations about them, and their organisational structure, structural metadata or reference metadata. The datasets can be linked to a SKOS concept which is an RDF resource, such as a publication.

Research into data citation includes several domain-specific projects. The Advanced Climate Research Infrastructure for Data (ACRIF) project[16] developed a Linked Data approach to citing and publishing climate research data along with full provenance information, including the workflows and what software was used (Ball, 1). The FISH Link project[17] offers tools for converting and mapping data from the domain of freshwater biology to Linked Data. It supports semantic markup, attribution and provenance.

2.1. Crowdsourcing considerations

There are various ways in which the creation of data citation links can be envisioned. Automatic techniques that identify potential dataset references are a first step in the right direction. Eliciting more detailed provenance accounts is likely to require human intervention, which can be acquired via crowdsourcing after the paper was published, or during the publication lifecycle by any of the parties involved (authors, reviewers, editors etc.) In this section, we will focus on settings where such information was not collected at the time of publication and data citation tasks are outsourced to an open crowd of contributors following one or a combination of crowdsourcing mechanisms.

When embarking on a crowdsourcing enterprise the ‘requester’ (that is, the party which resorts to the wisdom of the crowds to solve a given problem) has a variety of options to chose from in terms of specific contributions, their use as part of the final solution to the problem, and the

[13] PROV-DM http://www.w3.org/TR/prov-dm/
[14] Describing Linked Datasets, VoID: http://www.w3.org/TR/void/
[15] The RDF Data Cube vocabulary: http://www.w3.org/TR/vocab-data-cube/
[16] Advanced Climate Research Infrastructure for Data (ACRIF) Project: http://www.jisc.ac.uk/whatwedo/programmes/mrd/clip/acrid.aspx
[17] Fish Link Project: http://www.jisc.ac.uk/whatwedo/programmes/mrd/clip/fishlink.aspx

ways in which participants will be incentivised. Each of these dimensions stand for parameters in the crowdsourcing design space, while case studies and experience reports in the field provide theoretical and empirical evidence for the extent to which certain regions in that space are likely to be more successful than others. In the following we will introduce these dimensions and discuss their implications. In Section 4, we discuss a particular instantiation of the framework and its outcomes, and in Section 5 we present alternative designs.

A first dimension of crowdsourcing refers to the types of tasks that are the subject of the exercise. Literature distinguishes between macro and micro tasks, where the former roughly stands for those types of problems that are outsourced via an open call without any consideration about the way in which they will be solved. This applies, most importantly, for those cases in which the task is of a creative nature, and as such difficult to specify as a structured workflow, of when the workflow is part of the solution the requester is looking for (e.g., scientific challenges a la InnoCentive or soliciting ideas from a large audience). A second category of crowdsourcing scenarios deals with microtasks; these are much more constrained in their nature, and are expected to be at a level of granularity that allows the contributors to solve them rapidly, without giving too much thought to the underlying methodology. A typical project contains a number of such microtasks, which are outsourced to different contributors that approach them in parallel and independently of each other. Given the nature of the data citation problem, we would expect a microtask approach to be beneficial. For any collection of papers and datasets, one can easily define microtasks referring to pairs of papers and datasets, or one specific paper and all datasets that are relevant to it. No matter how the actual microtask looks like, one important aspect the requester has to take into account is the description of the task and the instructions he gives to the contributors. Assuming the task at hand asks for links between papers and a pre-defined list of datasets, one needs to think about the different ways in which both the paper and the dataset will be presented to the crowd. Alternatives include:

- for the paper: bibliographic entry, abstract, some pages, full paper;
- for the dataset: name, name and version number, documentation.

Each of these alternatives has advantages and disadvantages, and the choice is also influenced by the depends on the affordances given by the crowdsourcing platform used.

3. Motivation and use cases: two datasets, two types of links, two crowds

In this section, we present two datasets that we initially target with our system, USEWOD and DBpedia, as motivating use cases for an approach that can be applied to any dataset and any domain. Based on these datasets and the issues they present, we describe the relationships that emerge - one complex set between datasets or versions thereof, and one set between datasets and the publications using them. Then, depending on the level of expertise required, we show how these relations can be crowdsourced to obtain a rich set of data citation graphs.

DBpedia is the most prominent Linked Open Data source containing structured data which is automatically extracted from a particular version dump of Wikipedia. Hence, DBpedia is a cross-domain open dataset, and a fantastic example of the problem we are attempting to solve. It has a well established creation and publication processes, which generates versions with complex relationships between them. The project wiki of DBpedia is well maintained and it provides a comprehensive version history for download. This allows for the easy set up of mirrors of any particular DBpedia version and granularity (e.g. only specific language version or excluding particular link sets). In a change log the project team documents changes on the DBpedia ontology as well as changes of the extraction and interlinkage framework that affect the populated instances. But neither DBpedia in general nor any of its versions is archived in a research data repository, which would allow for referring to a persistent identifier such as a DOI for example.

The dataset, its versions, and the protocol for their generation evolve dynamically, based on community input and collaboration. It is however provider-dependent[18] and neither sustainable availability nor reliable long term archiving can be assured. As an additional complication, every DBpedia version originates from a particular Wikipedia dump version. When a DBpedia dataset is used in research, there exists a transitive dependency which makes the respective Wikipedia dump that has been processed by the DBpedia extraction algorithms the actual source of the data used in the research, influenced certainly by the scripts used to extract it. The different Wikipedia dataset dumps contain data created and altered by millions of wikipedians, and thus the relationship between DBpedia versions inherits the complexity of this provenance as well. Such complex relationships between datasets and versions are important in tracing the lineage of the data used in research publications, and the complexity is not present just in DBpedia's case.

DBpedia also has a large number of research publications which claim to use it in some way -- close to 10,000 articles found in Google Scholar with the "dbpedia" keyword[19]. However, the majority of the papers do not explicitly reference the particular DBpedia version used, and those that do reference it, do not do so in a consistent way. The key papers of the DBpedia publishers are cited instead of the actual DBpedia dataset version that was used in a particular study. This limits others' ability to reproduce or evaluate the published results, and makes it difficult to validate the research and draw useful conclusions from validation efforts.

The USEWOD dataset is a collection of server access logs from various well-known Linked Data datasets, most prominently DBpedia, LinkedGeoData[20], and BioPortal[21] amongst many others. As part of a data analysis challenge, the chairs of the annual USEWOD workshop released four dataset versions, one before each instance of the workshop since 2011. The four

[18] The DBpedia project started as a master thesis at University of Mannheim, and it is still hosted there, despite the uptake by the research community, and the impact DBpedia has had on the Linked Data world.
[19] 9650 results returned on 04/07/2014.
[20] http://linkedgeodata.org/
[21] http://bioportal.bioontology.org/

individual USEWOD dataset versions are available upon request from http://usewod.org and a description of the contents is included in the respective compressed dataset archive file. It is noteworthy that the 2012 and 2013 versions of the dataset each contained the entire content of the preceding year plus additional data. This practice has been changed in 2014 to release additional data only. As a very lightweight citation policy, and similarly to the DBpedia case above, the workshop chairs asked users of the USEWOD dataset to cite one of the initial papers describing the workshop and the research dataset (Berendt 305, Berendt 63).

Generally the case of the USEWOD dataset is representative for research datasets that are hosted by an academic unit or an individual researcher, such as the UCI Machine Learning Repository[22] or the Stanford SNAP dataset collection[23]. Again, a non-standard way of hosting and maintaining research data without any guarantee of sustainability of the service nor the chance to refer to a persistent identifier controlled by an official entity managing research data.

We detailed above through the DBpedia example how the relationships between various versions of the same dataset are relevant to the traceability of research results using one version or another. With the USEWOD dataset, which contains information related to other datasets, it becomes clear that the same importance have also the relationships between datasets: inclusion, dependence, transformation, aggregation, projection, etc.

The links between datasets are not always expressed in a standardised, machine readable way, but rather captured in textual documentation from the creators (or aggregators) of each dataset or version of. As such, the capture of these relations can be done in two ways, one is by extracting the information automatically where possible from the documentation, and the second is by asking the creators (the experts in this case) to manually specify them.

Moving on to the relations between publications and datasets, we find that for the majority of instances we can simply restrict the vocabulary to say that a piece of research, a paper, uses a dataset (or more than one) to obtain the results claimed. Our first experiment, described in the next section, supports the fact that this very shallow metadata is enough to gather a sufficiently rich data citation graph. This general way of establishing the link between a publication and the precise dataset and version used for the research has an added advantage of being easy to elicit from non experts, as there is no other detail required. It can also be automatically extracted in a large number of cases using text analysis and restrictions on the possible date ranges, as in the examples shown below.

The usage relationship however does not cover all the cases. Some publications do more than just use a dataset, they describe how a new one was generated, or analyze, compare, evaluate existing datasets. This more detailed metadata provides a richer information on how the data is

[22] http://archive.ics.uci.edu/ml/
[23] http://snap.stanford.edu/data/index.html

used in publications, but it is more difficult to extract automatically with high accuracy, and also more difficult to elicit from the crowd, as it requires expert (qualified) users.

We look at utilizing the power of the two types of crowds - that of experts and that or non-experts - in the way most suitable to each. For example we target the authors of publications and other domain experts for the crowdsourcing of the detailed usage metadata. For the general usage metadata we engage all available participants in the crowdsourcing tasks, possibly including Amazon's Mechanical Turk or other paid micro-task platforms.

We use simple information extraction tools to detect if any of the aforementioned relations can be detected automatically. An example is when the paper contains the version of the dataset in plain text, as does (Morsey 454) which contains "DBpedia (version 3.6)" in its introduction. Additionally, if available, we can use some of the metadata about datasets and publications to restrict the set of possible datasets linked to a paper based on the intersection of the temporal range of the creation dates. For example, the (Mendes 1) paper uses the DBpedia dataset for evaluation of a tool, but does not specify in the text the version used. The paper was published in September 2011, which means it could only have used DBpedia datasets up to version 3.7 (released in August 2011[24]). Taking into account the fact that the submission deadline for the conference was in April 2011[25], we can restrict the range by one more version, to DBpedia v. 3.6 (released in January 2011).

The automatically extracted and inferred information can be used in two ways, to validate the crowdsourced information, or to be validated by the crowd. We plan to explore both options in the future.

4. Crowdsourcing dataset references with experts: the USEWOD user study

During the USEWOD 2014 edition of the workshop we ran a small scale user study where we asked participants to annotate papers and datasets with the relations between them. The participants were experts in the field of the papers they were asked to annotate, some of them were authors of the said papers. As experts, they were asked not only to capture the simple usage relation, but also the more detailed descriptions of how a dataset is used by a given publication. Figure 1 shows a screenshot of the tool developed for the experiment. It is available online at http://prov.usewod.org/. Details of the tool, the data modeling, and the vocabularies used can be found in (Dragan 1).

[24] from http://wiki.dbpedia.org/Changelog?v=1crc
[25] from http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=15031&copyownerid=16219

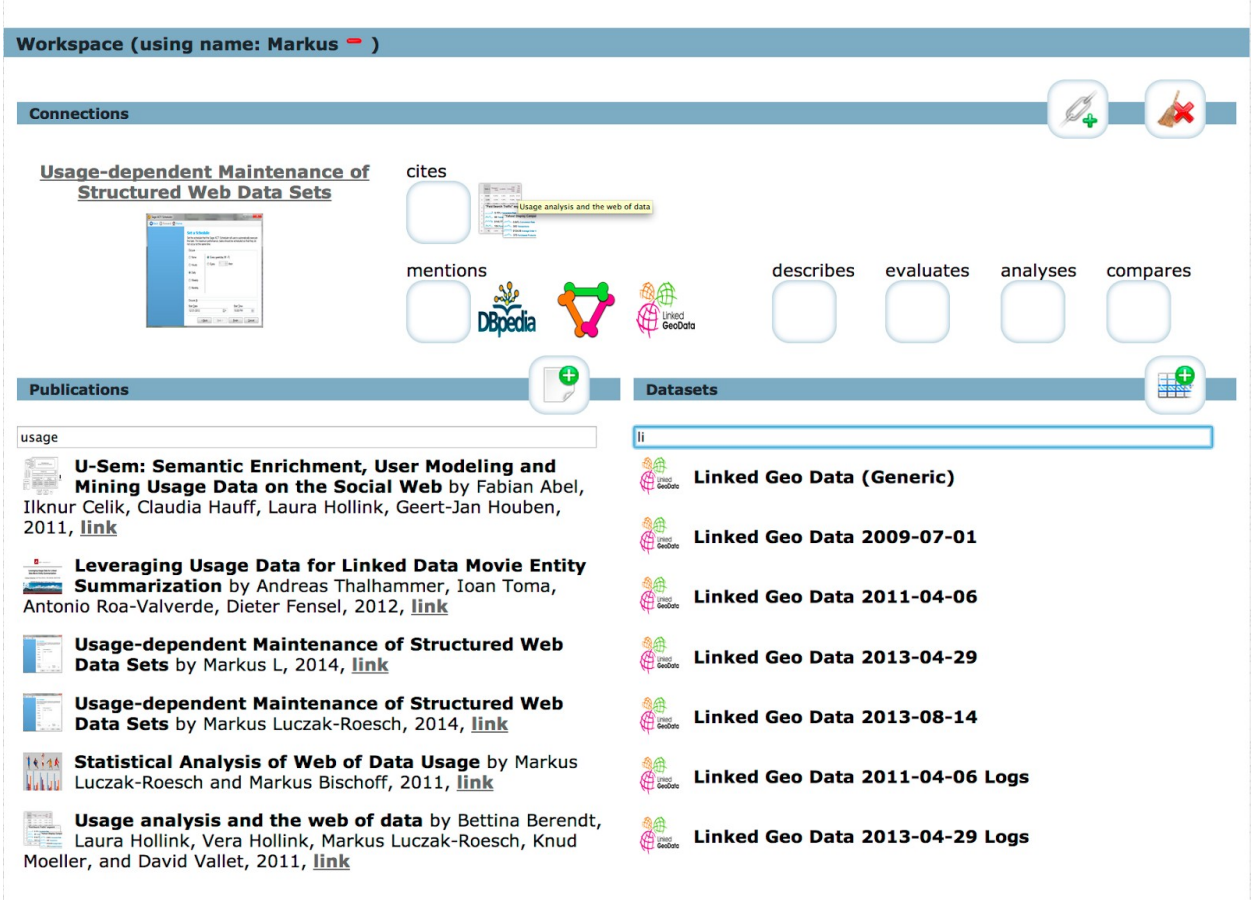

Figure 1: Screenshot of the crowdsourcing tool developed for the USEWOD experiment.

We made heavy use of the W3C PROV vocabulary, which we extended and used to capture provenance metadata about the links between the datasets and publications, but also for information about the crowdsourcing process. Figure 2 shows a citation captured as a result of a crowdsourcing task, a paper which is an analysis of a dataset. Figure 3 shows how the provenance information is captured about the task itself - the author and creation time is stored.

The resulting data is stored and published as Linked Data. During the 1 hour study, the 6 participants solved 81 tasks, adding in the system 19 new publications and 2 new datasets. They created 95 new relations, 27 of which were linking datasets to publications.

Besides of the actual data collection, the experiment had an alternative goal of testing some aspects of the system: the suitability of the vocabulary created, the perceived necessity of the detailed usage metadata in contrast with simple general usage links, and overall, the suitability of using crowdsourcing for such data collection.

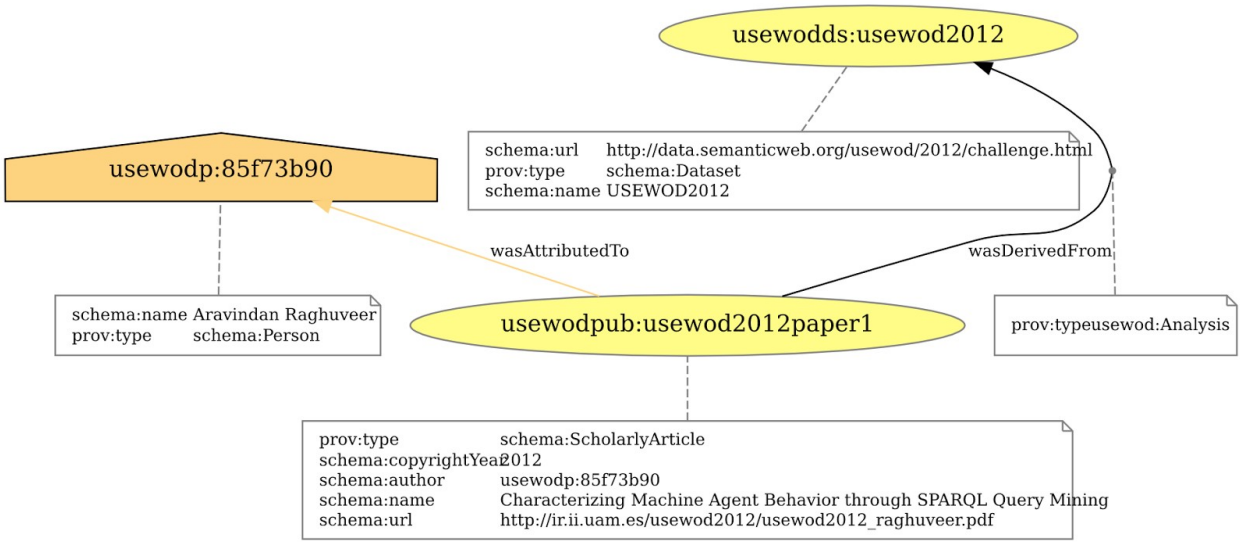

Figure 2: Citation graph of USEWOD2012 dataset from an analysis paper.

We learned that although the participants considered it was a good idea to collect detailed information about how the datasets are used in publications, the limited vocabulary we provided (5 possible relations) was not rich enough. However, it was at the same time too rich for the simple usage links, creating confusion as to which relation to use. Of the 27 dataset to publication links created, 21 were of type “analysis” and 6 were of type “mention” but participants commented that they were in fact intending to use a simple “uses” relation, which was not provided by the vocabulary.

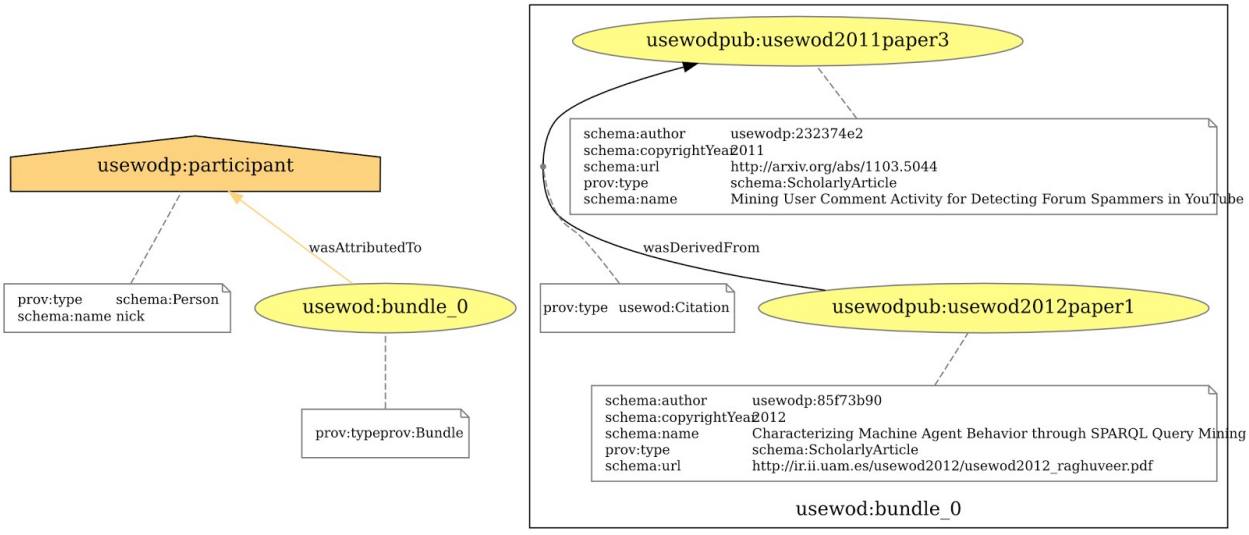

Figure 3: Provenance bundle captured for a crowdsourcing task.

The study also showed us, that while the crowd was made of experts, they did in fact only collect simple usage metadata. We attribute this outcome to the predominance of papers which describe processes in which data is used, compared to the number of papers which describe how data is created, or modified. This observation holds true for the field of Computer Science, but it might not be the case for domains where data is created and published as a (or the) result of the research being conducted.

We concluded that indeed crowdsourcing is a suitable tool for a-posteriori collection of links between publications and datasets, albeit with a few modifications to the way we use it in our system. A simplification of the microtasks is required, as well as minimising data input tasks to keep the focus of the participants on the link creation.

5. A generic process for crowdsourcing data citations

We exploit the findings from our initial user study to design the first version of a generic process for crowdsourcing typed references between research datasets and publications. Our proposal is a hybrid approach applying information extraction methods chained with crowdsourcing targeted to two different crowds, namely the authors of publications as ultimate experts and typical contributors to the completion of microtasks on crowdsourcing platforms like Mechanical Turk (aka turkers). The overall pipeline is depicted in Figure 4 below.

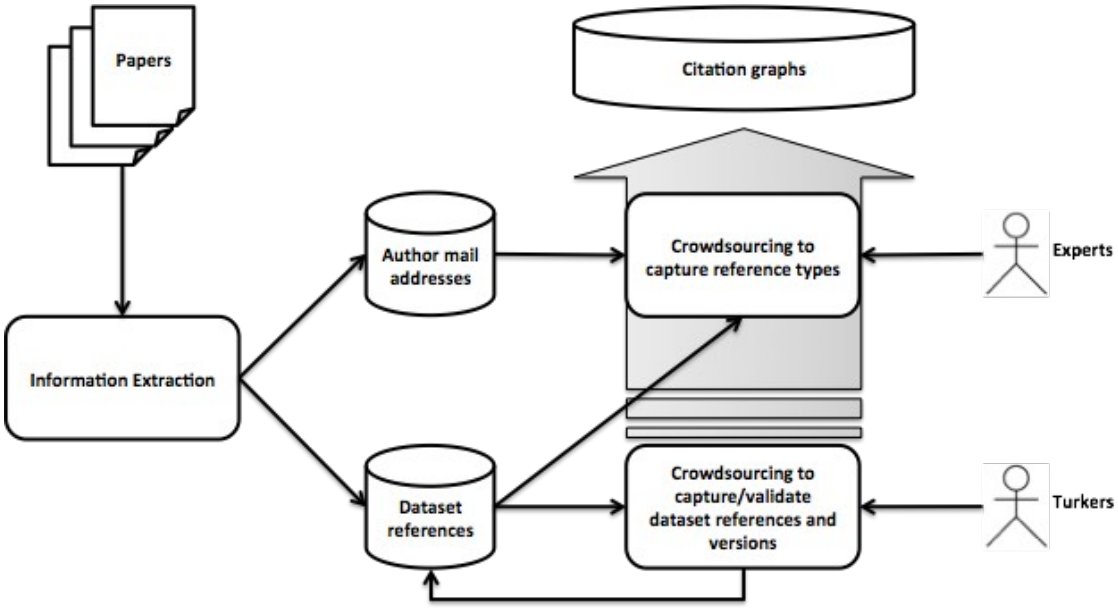

Figure 4: The proposed process for crowdsourcing typed references between research datasets and publications.

In the fully automatic information extraction step, text mining methods are applied (1) to extract email addresses of authors from papers, and (2) to extract content patterns that are potential references to research datasets. The result consists of two different sets. One contains the author email addresses with reference to the papers these authors wrote. The other contains the potential dataset references in relation to the papers these were extracted from. As a matter of fact the precision of the email addresses list will be much higher, even though it is worth to mention that the age of a publication can heavily influence the probability that an email address is no longer valid. Nonetheless, it is much more complex to match all possible content patterns of dataset references, which might be masked in regular references, footnotes or even just in the plain content. An example of a real, albeit simple example of such a match is given in Section 3 above.

Because we cannot guarantee that automatic extraction of dataset references will achieve high accuracy, we employ crowdsourcing to improve on extraction quality. The role of the microtasks performed with non-expert participants is twofold: to help validate the extraction result and thus train the automatic method, and secondly, to specify the accurate version of a dataset a paper refers to (if versions exist). Validated dataset references can be then used to elicit additional, detailed usage links from experts.

The potential types of relations between research papers and datasets are manifold. Based on the informal feedback we received following the test run of the experiment, we found that while the expert participants perceived capturing as much detail as possible as optimal, the types of relations we offered needed more discussion and specification. In the enhanced process that we propose here, the experts will no longer have to determine which version of a dataset is used in a publication, because this information will be passed on to them from the non-expert crowdsourcing effort. They will be thus allowed to focus on the task of describing the details of the usage. The purpose of the expert crowdsourcing is to capture these highly specific details, either by selecting them from an extensive list of typed relationships, or by allowing the experts to input free text. If the former option is used, the relations should be specified, and their intended application described in a codebook similar to the ACM Computing Classification System[26] for example. In the latter case, interrater agreement and clustering of the free text annotations can be used to build and refine the relations codebook.

The results of the references that underwent validation via non-expert crowdsourcing and subsequent enrichment via expert crowdsourcing are persisted in a data citation repository, which can be exposed online as Linked Data. As in our first experiment, provenance information must be saved to keep track of how a particular reference was derived.

[26] http://www.acm.org/about/class/2012

6. Conclusions and outlook

In this paper we described an approach for leveraging crowdsourcing to capture data citations represented as RDF graphs using the PROV vocabulary and schema.org primitives. The results of one user study targeted to a small group of experts has been reported and the findings were compiled into a proposal for a hybrid crowdsourcing pipeline that suits to large-scale cases where links between research papers and the primary research data sources used are missing. The study gave us example metadata and also feedback from the participants, which together point towards ways of improving both the metadata schema and the process. We have described these issues and ideas for future work in Sections 4 and 5.

Furthermore, there is a huge potential in experimenting with analytics on the gathered provenance data into two directions: First, exploiting provenance analytics allows to assess the accountability of crowdsourced data based on computing reputation profiles of the crowdsourcing participants. Second, publishing detailed data citation graphs as Linked Data opens new opportunities for developing alternative impact metrics for scholarly contributions by augmenting data from different sources and analyzing the derived entire data citation graphs.

In addition, we hope for an incentive effect of data and analytics obtained in this way. The rich sets of metadata can serve a dual purpose: to motivate more people to take part in crowdsourcing, and to bootstrap a process whereby better metadata about dataset citation are created earlier in the process, at the source: by authors and publishers.

References

Ball, Alex, and Monica Duke. "Data citation and linking." (2011).

Berendt, Bettina, Laura Hollink, Vera Hollink, Markus Luczak-Rösch, Knud Möller, and David Vallet. USEWOD2011: 1st international workshop on usage analysis and the web of data. In Proceedings of the 20th international conference companion on World wide web (WWW '11). ACM, New York, NY, USA, 305-306, 2011. DOI=10.1145/1963192.1963324 http://doi.acm.org/10.1145/1963192.1963324

Berendt, Bettina, Laura Hollink, Vera Hollink, Markus Luczak-Rösch, Knud Möller, and David Vallet. Usage analysis and the web of data. SIGIR Forum 45, 1, 63-69, 2011. DOI=10.1145/1988852.1988864 http://doi.acm.org/10.1145/1988852.1988864

Bourne, P. E., et al. "Force11 Manifiesto: Improving future research communication and e-scholarship." White paper. Retrieved online at: http://force11. org/white_paper (2012).

Dragan, Laura, Luczak-Rösch, Markus, Simperl, Elena, Berendt, Bettina and Moreau, Luc (2014) Crowdsourcing data citation graphs using provenance. In, Provenance Analytics (ProvAnalytics2014), Cologne, DE, 09 Jun 2014. 4pp.

Morsey, Mohamed, et al. "DBpedia SPARQL benchmark–performance assessment with real queries on real data." The Semantic Web–ISWC 2011. Springer Berlin Heidelberg, 2011. 454-469.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems (I-Semantics '11), Chiara Ghidini, Axel-Cyrille Ngonga Ngomo, Stefanie Lindstaedt, and Tassilo Pellegrini (Eds.). ACM, New York, NY, USA, 1-8. DOI=10.1145/2063518.2063519 http://doi.acm.org/10.1145/2063518.2063519

Peroni, Silvio, and David Shotton. "FaBiO and CiTO: ontologies for describing bibliographic resources and citations." Web Semantics: Science, Services and Agents on the World Wide Web 17 (2012): 33-43.

Shotton, David. "Introduction the Semantic Publishing and Referencing (SPAR) Ontologies. October 14, 2010." URL: http://opencitations.wordpress.com/2010/10/14/introducing-the-semantic-publishingand-referencing-spar-ontologies (2010).