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

Laura Drăgan, Markus Luczak-Rösch, Bettina Berendt, Elena Simperl, Heather Packer, Luc Moreau

Abstract

In this paper we present opportunities to leverage crowdsourcing for a-posteriori capturing dataset citation graphs. We describe a user study we carried out, which applied a possible crowdsourcing technique to collect this information from domain experts. We propose to publish the results as Linked Data, using the W3C PROV standard, and we demonstrate how to do this with the Web-based application we built for the study. Based on the results and feedback from this first experiment, we introduce a two-layered approach that combines information extraction technology and crowdsourcing in order to achieve both scalability (through the use of automatic tools) and accuracy (via human intelligence). In addition, non-experts can become involved in the process.

1. Introduction

The need to treat research datasets as “first-class citizens” in the scientific publishing process is recognised in many disciplines. Many popular citation guidelines have been enriched with templates for data publication and citation[1]. This enables a more informed review and reuse of scientific work, as readers of scholarly publications can now easily consult the relevant datasets and assess their quality. References to datasets could also become an integral part of bibliographic algorithms in order to add data-specific statistics to traditional citation graphs. Going a step further, datasets could have their own form of citation: a dataset could be composite of, derived from, a subset of, the aggregate of, or a new version of other datasets. The combination of metadata about scientific publications and the related data, citation links between these artefacts, and versioning information could be the source for rich analytics, which would offer a more complete picture of the scientific publishing process and would drastically improve reproducibility of research results. However promising, and conceptually simple, such idea might sound, exploring this integrated information space is still a thing of the future.

In this paper we propose different opportunities to leverage crowdsourcing for a-posteriori creating dataset citation graphs. By *a-posteriori* we mean that the information is captured “after publication”, as opposed to “at the time of writing or submission”. This is motivated by the large amount of existing publications and datasets that are already published, but not interlinked. We describe a practical approach, which exploits a specific crowdsourcing technique to elicit these graphs from domain experts. The results cover both types of information mentioned earlier: the relationship between publications and data sources, as well as between different dataset

[1] For example, Nature’s Scientific Data, “an open-access, online-only publication for descriptions of scientifically valuable datasets”, which collaborates with several existing data repositories.

versions or derivatives. For the representation of these augmented citation graphs we apply provenance modelling as recommended by the W3C provenance working group, as well as the Linked Data principles (Berners-Lee, 2006) to facilitate online access and data integration.

In the following we will refer to two examples to illustrate the main idea of our approach: the DBpedia Linked Open Data dataset, and the USEWOD log file datasets. We report on a small user study which was run during the USEWOD2014 workshop with a group of experts as participants in the crowdsourcing process. Following up the findings of this study, we redesigned the approach to balance accuracy and scalability; we combined information extraction technology (automatic, hence fast) with crowd intelligence (manual, hence accurate). This hybrid workflow opened up the possibility to use multiple forms of crowdsourcing for different tasks, most importantly enabling us to involve non-experts (hence, a larger crowd than the research community) in the information collection and analysis process.

We define two types of relationships between publications and the datasets they refer to: a generic, high-level relationship which merely captures the fact that a dataset (possible with some versioning information) is "used" in a paper; and a more specialized set of relationships which provide details about the role of the data artefact in that line of scientific inquiry. This distinction makes it easier to collect information; some contributors to our crowdsourcing endeavour might not have the time or knowledge to identify very specific data citation information from a publication. In those cases in which this information can be elicited, we offer the conceptual gestalt to represent it, hence enabling more advanced analytics and giving a more complete picture of the scientific process.

In Section 2 we present some of the current developments around data citation, and briefly introduce the fundamentals of crowdsourcing. In Section 3 we describe our two use cases, while in Section 4 we discuss a particular instantiation of the framework and its outcomes. In Section 5 we present the enhanced design of our system and conclude with an outlook on future work.

2. Background and related work

2.1 Capturing data citations

Many organisations have identified the need for data citation and have developed principles and rules to support it. The Force11 community (Bourne et al., 2012) has created a list of eight data citation principles, which cover purpose, function, and attributes of citations. The principles describe data citation importance, credit and attribution, evidence, unique identification, access, persistence, specificity and verifiability, and their interoperability and flexibility.

> "*These principles recognise the dual necessity of creating citation practices that are both human understandable and machine-actionable.*" (Force11, 2014)

Michigan State University provide guidelines for citing data using established bibliography styles such as APA, MLA or Chicago and Harvard (Michigan State University Libraries, 2014).

Attributes recommended to describe a dataset include author or creator, date of publication, title and publisher, and additional information such as edition or version, date accessed online, and a format description can be included. They recommend using an identifier that is persistent, such as a URL or DOI.

Some e-Science initiatives such as OpenML (a repository of machine-learning experiments, in which datasets play a pivotal role) supply exportable metadata records for datasets; they also measure citation counts and usage in experiments and support links to dataset description publications, so that scientists who publish a dataset can indicate which paper they want other people to cite when reusing the data. This citation of a publication - rather than of the data artefact itself - is customary in some disciplines such as Computer Science. It ensures that credit is given to the creators of the dataset. However, as we show in the following sections, it creates a disconnect between the data and the results, which can be detrimental to the scientific workflow. We therefore propose to go beyond citing publications only.

The data citation approach recommended by most publishers is to use the well-established Digital Object Identifier (DOI) system. A DOI is a persistent identifier which is dereferenceable and provides metadata describing the object, in our case the dataset. DataCite and Cite My Data create DOIs for the datasets published through their platform.

The use of standardised vocabularies to describe data citation supports many of the guidelines and principles mentioned above. Some vocabularies are specifically designed for this very type of references, while others could be re-purposed to cover data citation as well. These include:

- Semantic Publishing And Referencing (SPAR) (Shotton & Peroni, 2010), which consists of eight core ontologies that describe bibliographic records, their citations, and the relationships between records and citations. In particular, the Citation Typing Ontology (CiTO) (Peroni & Shotton, 2012) defines object properties for citations, and includes properties that describe how data is used in publications, for example *is discussed by*, *cites as evidence*, and *uses data from*. This ontology can be used independently of the other SPAR ontologies because it does not use restrictions of domains and ranges.
- The W3C PROV recommendations can be taken into account to record provenance information about the entities, activities, and people involved in producing an artefact (data or any other type of digital object), which can, in turn, be used to form assessments about its quality, reliability, and trustworthiness. The PROV data model (Moreau & Missier, 2013) is amenable to data citations, and can be extended to describe more specific connections between datasets and publications.
- schema.org proposes a collection of schemas to markup Web pages with structured information. The widely used vocabulary is designed to be generic and make the creation of structured markup for Web content as straightforward as possible. It contains classes for *Dataset*s and publications (*Scholarly Article*, *Book*, etc.), all subclasses of *Creative work*, which defines *citation* links and other metadata. The citation property was only recently added to the *Creative work* class, thanks to the work undertaken in the W3C Schema Bib Extend Community Group and W3C WebSchemas Group.
- VoID is an RDF Schema vocabulary for expressing metadata about RDF datasets. Its main purpose is to allow publishers to express metadata about RDF datasets for applications ranging from data discovery to cataloguing and archiving. It is based on the

Dublin Core vocabulary and describes access metadata, structural metadata, and links between datasets. It is not specific to datasets used in research.

- The RDF Data Cube vocabulary is concerned with statistical datasets, observations about them, and their organisational, structural, and reference metadata. The datasets can be linked to other resources, such as publications.

Research into data citation includes several domain-specific projects, including the Advanced Climate Research Infrastructure for Data (ACRIF) project, which developed a Linked Data approach to citation and publication of climate research data along with full provenance information, including the workflows and software that was used (Ball & Duke, 2012).

2.2. Crowdsourcing

Crowdsourcing was defined by Howe as:

> "*the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. 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.*" (Howe, 2006)

There are various ways in which data citation links could be created through crowdsourcing. Putting aside the various forms to deploy the original notion defined by Howe, contributions could be sought from multiple audiences, or *crowds*, from the researchers authoring or reviewing a publication to publishers, dataset owners and users, and the general public. Going a step further, we could imagine various types of contributions and crowdsourcing workflows, ranging from the identification of links between papers and datasets to validating existing citations or eliciting further information about the dataset itself, including versioning. Automatic techniques could be exploited to identify potential dataset references, or to discover potential inconsistencies in the responses submitted by the crowd. All this could happen either at publication time (e.g., when the camera-ready version of an article is prepared for submission) or independently of the publication life-cycle.

In this section, we focus on settings where such information has *not* been collected at the time of publication and data citation tasks are outsourced to an open crowd of contributors using 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 choose from in terms of specific contributions, their use as part of the solution to the problem, and the ways in which participants will be incentivised. Each of these is a dimension of the crowdsourcing design space. 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 remainder of this section we introduce these dimensions and discuss the implications of choosing one alternative over the other.

A first dimension of crowdsourcing refers to the task that is assigned. The literature distinguishes between two types of tasks: macro- and micro-tasks (Dawson & Bynghall, 2012). Macro-tasks are outsourced via an open call without specifying how they are to completed. This is the case, most importantly, when the task is of a creative nature, and as such difficult to define one structured workflow that will achieve the goal (e.g., scientific challenges à la InnoCentive, or eParticipation approaches to policy making (Aitamurto, 2012)). A second category of crowdsourcing deals with micro-tasks: these are much more constrained and at a level of granularity that allows the contributors to solve them rapidly and without much effort. A typical project contains a number of such micro-tasks, which are outsourced to different contributors that approach them in parallel and independently of each other. This makes the overall project very efficient, though it adds overhead in consolidating the individual inputs into the final result.

Given the nature of the data citation problem, we expect a micro-task approach to be beneficial. For any collection of papers and datasets, one can easily define micro-tasks referring to pairs of papers and datasets, or one specific paper and all datasets that are relevant to it. No matter what the actual micro-task looks like, the requester has to carefully craft the description of the task and the instructions given to the contributors. Assuming the task 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 also depends on the affordances given by the crowdsourcing platform used.

A second dimension of crowdsourcing describes the targeted crowd of participants. The preconditions of the task may determine who can be in the targeted group. This group may be a restricted group of experts who qualify by fulfilling a given condition (e.g., having a skill, being of a certain height, or living in a given area). Alternatively, an open call for participation can be made, where anyone can take part. However, even when no explicit prerequisite apply, it is still worthwhile to consider what would drive contributors to engage with the task at all and to design incentives that would encourage them to do so (Simperl, Cuel & Stein, 2013). For data citation tasks, different types of participants can provide different types of useful information:

- the publication authors are in the best position to offer precise information about how the datasets were used in their research;
- the dataset creators can describe the dataset-to-dataset relations, and various versions;
- domain experts can provide information on which datasets are used and how in publications based on their domain of expertise;
- a non-expert crowd can extract basic metadata information from text snippets from publications or dataset documentation.

Targeting a specific crowd can have advantages and drawbacks. For example, experts are few and usually constrained by time, whereas non-experts may provide less precise information that must be validated.

A third dimension of crowdsourcing is how results are managed. Depending on the task design and chosen crowd, the results may need to be validated, aggregated, integrated, etc. The solution space of some problems can be open, meaning that the number of possible correct solutions is unlimited. This is mostly the case with macro-tasks and open calls which require creative work and innovation. In the case of crowdsourcing data citation links however, the solution space is rather constrained, since there is a finite number of correct links that can be created between given publications and datasets. The solutions submitted by participants can be assessed automatically by comparing results against one another, using simple or weighted averages, or majority voting (Kittur, Chi & Suh, 2008). It is also possible to crowdsource the evaluation of results by assigning participants evaluation tasks, in which they verify the accuracy of previously submitted answers. This option is more costly, as it require additional participants, but also likely to produce more reliable results, especially in cases in which answers are not straightforward to obtain or require very specific insight. A hybrid approach can employ both methods, by either randomly selecting results to be evaluated from the total set of submitted results, or by only evaluating the ones where agreement between participants is low. The performance of a crowd member can be evaluated by assessing their contributions' divergence from the "ground truth" approximated from aggregating over all contributions. When the ground truth is known partially (e.g. for a limited number of cases out of the total, for which we have expert-made annotations) we can test the participants' contributions by randomly or selectively assigning them tasks to which we know the true solution. Social mechanisms like participant reputation, and rating and voting (Packer, Dragan & Moreau, 2014), can be also used to infer the quality of work.

We now turn to an analysis of these considerations.

3. Use cases: two datasets, two types of links, two crowds

In this section, we present two datasets that initially motivated our research, USEWOD and DBpedia. Our resulting approach, however, can be applied to any dataset and any domain. Based on these datasets and their characteristics we describe the relationships that exist - between datasets, or different versions of a dataset, and between datasets and the publications using them. We then show how these relations can be crowdsourced to obtain data citation graphs, and how the process varies with the level of expertise of the participants.

DBpedia is the most prominent Linked Open Data source containing structured data that is automatically extracted from Wikipedia. Hence, DBpedia is a cross-domain open dataset, and a fantastic example of the problem we are attempting to solve. It has well-established creation and publication processes, which generate dataset versions with complex relationships between them. The DBpedia project wiki is well maintained, and it provides a comprehensive version history and detailed information. This makes it easy to set up mirrors of any particular DBpedia version and granularity (e.g. only a specific language or excluding particular link sets). In a change log the DBpedia team documents changes on the DBpedia ontology as well as changes of the extraction and interlinkage framework. 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. The dataset, its versions, and the protocol for their generation evolve dynamically, based on community input and collaboration. It is however provider-dependent[2], and neither sustainable availability nor reliable long term archiving can be assured.

Additionally, every DBpedia version originates from a particular Wikipedia dump. 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 dumps contain data created and altered by millions of people, and thus the relationship between DBpedia versions inherits the complexity of this provenance chain. 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 specific to DBpedia.

DBpedia is also associated with a large number of research publications that claim to use it in some way - at the time of writing, approximately 10,000 articles found in Google Scholar with the "dbpedia" keyword[3]. However, the majority of the publications do not explicitly reference the particular DBpedia version they use, and those that do reference it, do not do so in a consistent way. As described in Section 2, the key papers of the DBpedia publishers are cited rather than the actual DBpedia dataset version that was used in a particular study. This limits others' ability to reproduce or evaluate the published results, and it 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, and BioPortal 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 individual USEWOD dataset versions are available upon request from http://usewod.org, and a description of the contents is included in the respective compressed 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 was changed in 2014 to release additional data only. As a lightweight citation policy, the workshop chairs asked users of the USEWOD dataset to cite one of the initial papers describing the workshop and the research dataset (Berendt et al., 2011).

The provision of the USEWOD dataset is representative of research datasets that are hosted by an academic unit or an individual researcher, such as the UCI Machine Learning Repository or the Stanford SNAP dataset collection. They all employ a non-standard way of hosting and maintaining research data without any guarantee of long-term availability of the service, and they do not provide their users with an option to refer to a persistent identifier controlled by an official entity managing research data.

[2] 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.

[3] And the count is growing: Google scholar returned 9650 results on 04/07/2014, and 10700 results on 24/10/2014 for the search term "dbpedia".

We detailed above through the DBpedia example how the relationships between various versions of the same dataset are relevant for 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 relationships between datasets are just as important: 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 by the creators of each dataset or version thereof. As such, the capture of these relations can be done in two ways: by extracting the information automatically, where possible, from the documentation, or 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 publication, *uses* a dataset (or more than one). Our first user study, described in the next section, suggests that this simple metadata is enough to gather a clear data citation graph. This general way of establishing the link between a publication and the precise dataset and version used for the research has the added advantage of being easy to elicit from non-experts, as no further information is 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 *uses* relationship however does not cover all cases. Some publications do more than just use a dataset, they describe how a new one was generated, or they analyse, compare, and evaluate existing datasets. This more detailed metadata provides richer information on *how* the data is used in research, but it is more difficult to extract automatically with high accuracy, and also more difficult to elicit from the crowd, as it requires expert users.

We investigate how to utilize the power of the two types of crowds, that of experts and that or non-experts, in the way most suitable to each type. For example we target the authors of publications and other domain experts for the crowdsourcing of detailed usage metadata. For general usage metadata we propose to engage a wider set of 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 whether any dataset reference can be found automatically. This is straightforward when the paper contains the version of the dataset in plain text, as for example (Morsey, Lehmann, Auer & Ngonga Ngomo, 2011) 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 DBpedia Spotlight paper (Mendes, Jakob, Garcia-Silva & Bizer, 2011) uses the DBpedia dataset for evaluation of a tool, but does not specify in the text which version was 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, according to the changelog). Taking into account the fact that the submission deadline for the conference was in April 2011 (according to the call for papers), 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 USEWOD2014 study

During the USEWOD2014 edition of the workshop we ran a small user study in which we asked workshop 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 study. It is available online at http://prov.usewod.org/. The functionality of the tool, the data modelling, and the vocabularies used are described in detail in (Dragan, Luczak-Roesch, Simperl, Berendt & Moreau, 2014). The vocabulary provided five possible relationships between publications and datasets: ‘mentions’, ‘describes’, ‘evaluates’, ‘analyses’, ‘compares’; and five possible relationships between datasets: ‘extends’, ‘includes’, ‘overlaps’, is transformation of’, ‘is generalisation of’. The simplest relation between a publication and a dataset was ‘mentions’, which was intended to be used when none of the others were suitable, but *a mention of the given dataset appears in the given paper*.

We built on the W3C PROV vocabulary, which we extended and used to capture provenance metadata about the links between the datasets and publications, and also to capture 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 about the solving of the task itself - the author and creation time - is stored.

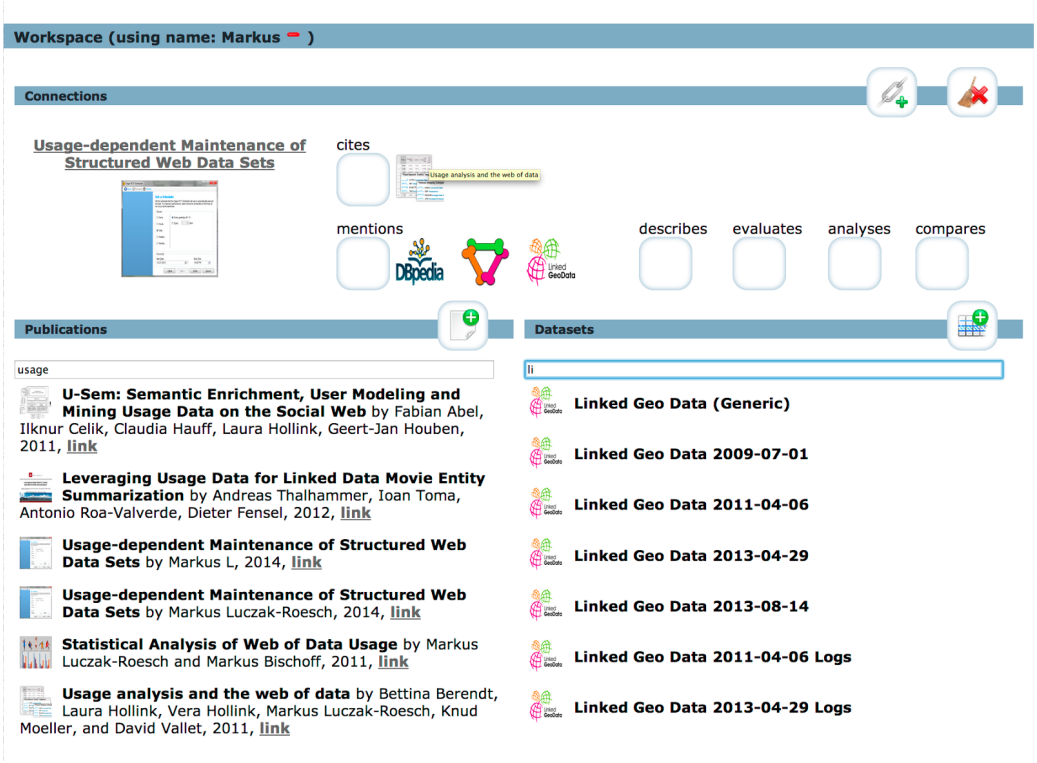

Figure 1: Screenshot of the crowdsourcing tool developed for the USEWOD study.

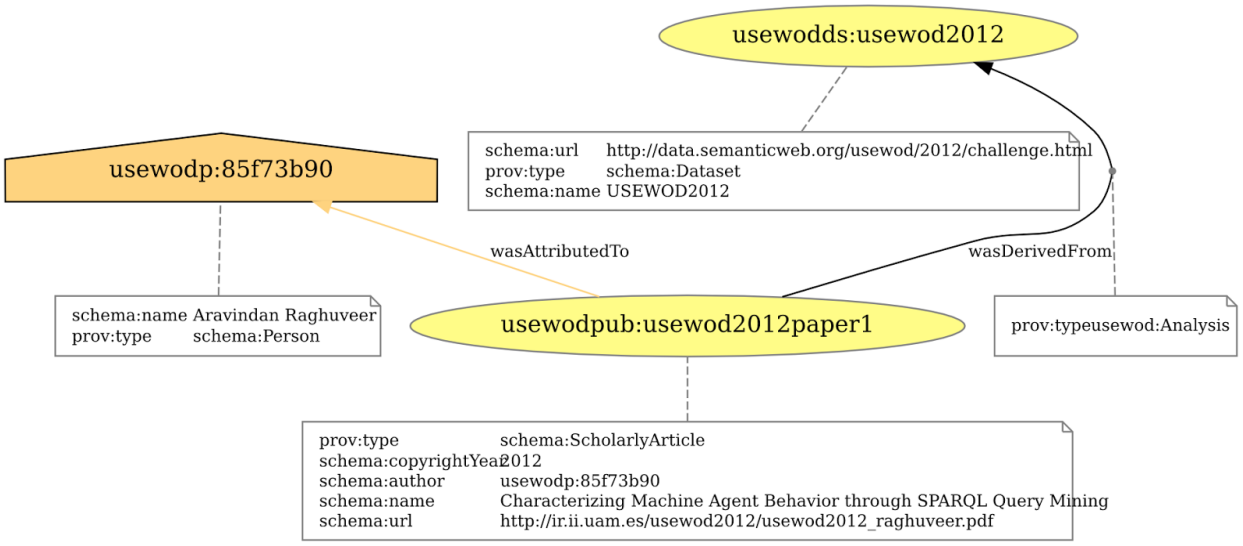

Figure 2: Citation graph of USEWOD2012 dataset from an analysis paper.

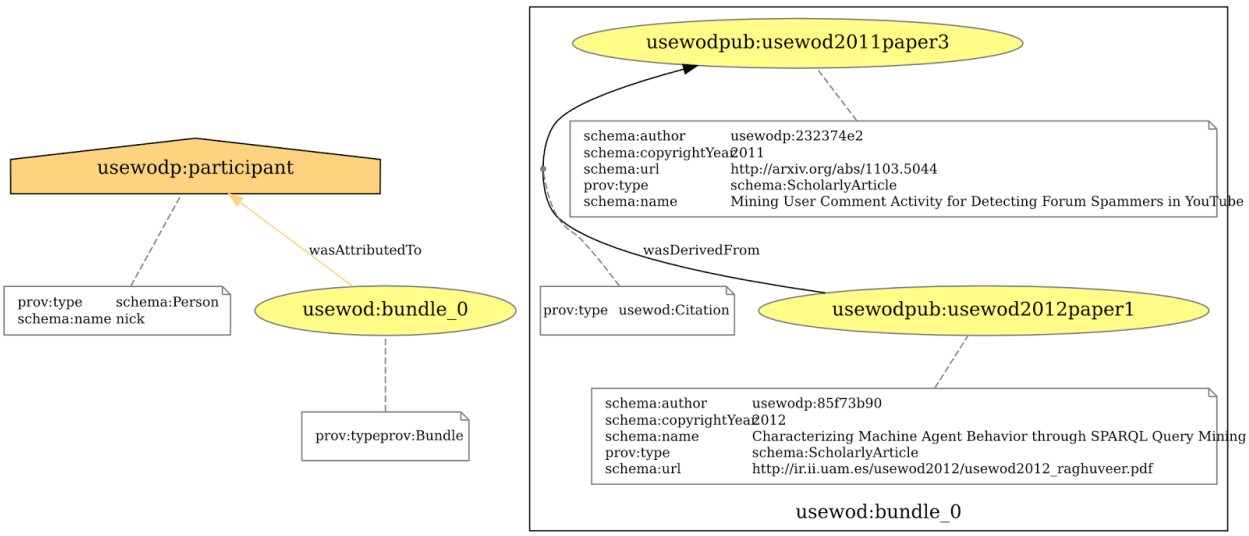

Figure 3: Provenance bundle captured for a crowdsourcing task.

The resulting information is stored and published as Linked Data at usewod.org. During the one hour run of the study, the six participants solved 81 tasks adding in the system 19 new publications and 2 new datasets. They created 95 new relations, 27 of which linked datasets to publications.

Besides the actual data collection, the study was designed to test 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 the overall suitability of using crowdsourcing for such data collection.

We learned that the participants considered it a good idea to collect detailed information about how the datasets are used in publications. However, the limited vocabulary we provided was not descriptive enough for some of the complex usage scenarios the participants wanted to capture.

At the same time it was confusing to describe the simple usage links, by choosing one of the given relations. Of the 27 dataset to publication links created, 21 were of type “analysis” and 6 were of type ‘mention’. Our intent was that analyses’ refers to in depth analysis of a dataset in a publication, however, from the feedback received we concludes that both types of links were applied in the very generic sense of ‘*publication - uses - dataset*’, and the participants commented that they would have preferred to have available a ‘uses’ relation, which was not provided by the vocabulary (‘mentions’ was the most generic one offered). The study thus showed, that while the crowd consisted of experts, they in fact only collected simple usage metadata. We attribute this outcome to the predominance of papers that describe processes in which data is used, which outnumber papers that 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 conclude that crowdsourcing can indeed be suitable for the a-posteriori collection of links between publications and datasets, albeit with a few modifications to the way we use it in our system. Tasks should be simplified, and data input required from participants should be minimised in order to maintain the focus on the link creation.

5. A generic architecture for crowdsourcing data citation

The user study we described in the previous section was small scale – around a hundred possible publications, and also tens of possible datasets and versions thereof. However, it served to test our hypotheses and set-up, and it let us gain insights into how we can improve the approach before deploying it on a larger scale. In this section we propose an improved process for crowdsourcing data citation links between publications and datasets.

The experiment showed that crowdsourcing was a suitable tool for a-posteriori collection of links between publications and datasets. However, we presented a rather complex and underspecified task to the group of experts, asking them for multiple contributions at once. To improve this aspect, the tasks have to be simplified to keep the participant's focus and avoid confusion. Consequently, crowdsourcing the links between publications and datasets needs to be decomposed into a pipeline of multiple small tasks targeted to different crowds depending on the domain knowledge they require.

The differentiation between the two types of possible links - simple 'use' and the more detailed relations should be considered in the task design, as well as in the possible automation of parts of the tasks. Automatic extraction of mentions of dataset names in publications can serve as a pre-processing step before the task is assigned to a human participant, and can be realised with existing text processing tools.

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 micro-tasks on crowdsourcing platforms such as Mechanical Turk. The overall process is depicted in Figure 4.

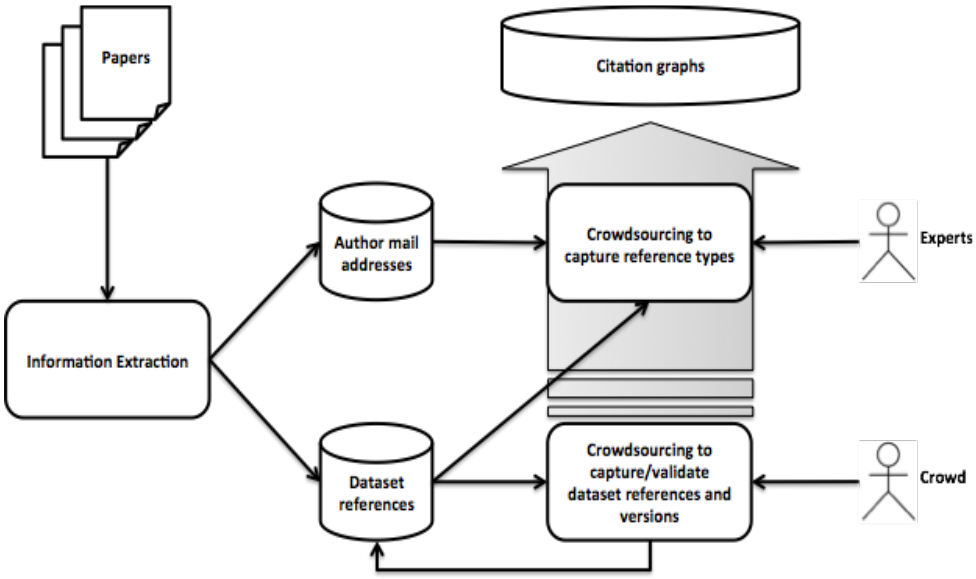

Figure 4: The proposed general architecture.

In the 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. We expect that the precision of the email addresses list will be relatively high, although the age of a publication can influence the likelihood that an email address is no longer valid. Nonetheless, it is 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 text. An example of a real, albeit simple example of such a match is given in Section 3.

Since we cannot guarantee the accuracy of the automatic extraction of dataset references, we employ crowdsourcing to verify and improve the quality of results. The micro-tasks can be designed in two complementary modes, both can be used in the system at the same time: validation mode, when participants are asked to verify information which was automatically extracted; or generation mode, when participants are given access to the full or partial text of a publication and asked to input names of datasets which appear in it. Each task mode can then be further defined, for example in a validation task participants can be asked to answer with yes or no if a dataset name is used in a publication, or they can be asked to choose which of several datasets are mentioned in the publication. The results from the generation tasks can be used as input for further validation tasks until a desired level of certainty is achieved. Another aspect of the task design include the size of the text given to participants – the full text of a publication, a page, a paragraph, etc,

Once we have validated dataset references, they can be then used to elicit additional, detailed usage links from experts. The potential types of detailed relations between research papers and datasets are manifold. Based on the informal feedback we received following our user study, we

found that while the expert participants considered it desirable to capture as much detail as possible, the types of relations we offered need 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 provided to them from the non-expert crowdsourcing effort. They will be thus allowed to focus on the task of describing the details of data 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, for example, the ACM Computing Classification System. In the latter case, inter-rater agreement and clustering of the free text annotations can be used to build and refine the relations codebook.

A more careful consideration of motivation and incentives could make other groups of participants more willing to make the cognitive effort required create the more specific types of links. Researchers other than the authors of the publications can be incentivised to participate, or even citizen scientists interested in the domain. However, we would have to be prepared to deal with subjective views and diverging answers to each paper-dataset pair. Frameworks such as CrowdTruth offer inspiration on dealing with such cases.

The results of the references that were validated via non-expert crowdsourcing and subsequent enrichment via expert crowdsourcing are made persistent in a data citation repository. As in our first user study, provenance information should be saved to keep track of how a particular reference was derived.

Our work focused on gaining a better understanding of crowdsourcing for data citation tasks. However, by choosing RDF as the format for the resulting data linking publications and datasets we also provide a practical example of how data citations can be represented and persisted technically. While the graph structure of RDF already facilitates various data citation analytics to uncover co-authorship, dataset dependencies, and impact factors for example, extending it to be Linked Data makes these analytics possible across the boundaries of a single digital library data inventory.

5.1 Further applications of crowdsourcing provenance metadata

The idea of crowdsourcing provenance data can also be generalized to settings beyond data citation. This could not only increase the amount of data in a certain domain, or improve the quality of already-existing data, but also have significant emancipatory potential in an era of “Big Data” that are increasingly being regarded as telling “the truth”, crowding out alternative accounts of the objects they describe.

For example, more and more government data are becoming available about cities via “city dashboards” and similar platforms designed to enhance transparency and improve governance and politics as well as citizen information and engagement (Kitchin, 2014). However, the origin of these data, the way they were defined as operational constructs, the way they were

measured, their uncertainties, incompleteness and error margins are often lacking and in any case not part of the data, which encourages naïve interpretations (as “facts”) that may lead to erroneous further conclusions. With some data, it may even be argued that there is no “ground truth” from which they could diverge, that instead all data are ways of creating some truth. Enriching such data with provenance metadata is a useful step; allowing citizens or other “crowds” to create their annotations to data (e.g. as provenance as described in Section 4) could create a more engaging democratic discourse about the data, encourage the critical examination of supposed “facts”, etc. Similarly, *data journalism* could allow for participatory forms of data commenting and enrichment in its displays that so far are interactive, but a one-way street in terms of knowledge transfer. *Citizen science* already uses crowdsourcing in manifold ways (Luczak-Roesch et al., 2014). In *Letters of 1916 – Creating History*, this is expressly done to elicit more, and more personal and diverse, “stories” in addition to an already-existing mainstream account of a certain historical event (Trinity College Dublin Communications Office, 2013).

Provenance information is a form of “story” of the data. We envisage crowdsourcing architectures that use provenance as a conduit to *storytelling* by citizens (Berendt, 2014). This will build on the well-established power of storytelling for engaging people, helping them understand complex configurations, identify and challenge assumptions and inconsistencies, and in general engage with “the truths” they are told in a more critical way.

Such large-scale deployments for the acquisition of metadata from different contributors will increasingly also offer a valuable testing ground for ways of addressing ethical issues of crowdsourcing (for discussions of the ethics of Mechanical Turk, see for example (Silberman, Irani & Ross, 2010; Williamson, 2014); for thoughts about the ethics of citizen science, see for example (Marcus, 2014), or (OpenScientist, 2013)). To what extent will the enthusiasm and quality of open-source software, Wikipedia and other projects continue or even scale? Even if it does, are we as scientists doing the right thing when we accept volunteers’ donations of time or willingness to work for low pay? To what extent may we just be reacting to pressures of insufficient funding of science by exerting (even if soft) pressure on others to work for free, and is that the right way of responding to pressure? If science projects get to be bigger and bigger but require more and more human input, do we need new models of funding? How do we react to requests for acknowledgement from these “crowds”, whether they are financial or content-related? Is contributing to science more like a hobby, a labour-market transaction, or a civic duty, or something altogether different? We believe that ultimately, facing these questions will lead to better crowdsourcing - and better science.

6. Conclusions and outlook

In this paper we have described an approach to leverage crowdsourcing to capture data citations. We reported on the design, execution, and results of a user study tailored to a small group of experts, and we compiled the findings into a proposal for a hybrid crowdsourcing pipeline that suits large-scale collection efforts where links between research papers and the primary research data sources used are missing. The study gave us example metadata and

feedback from the participants, which together point towards ways of improving both the schema and the process. We have described these issues and ideas for future work, including generalizations of provenance crowdsourcing beyond data citation, and a critical discussion on the ethics of crowdsourcing, in Sections 4 and 5.

There is a huge potential in experimenting with analytics on the gathered provenance data into two directions. First, exploiting provenance analytics allows one to assess the accountability of crowdsourced data based on computing reputation profiles of the 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 analysing the derived entire data citation graphs.

The user study reported was only the first step towards a thorough understanding of crowdsourcing links between scientific publications and research datasets. It is the first step towards developing the set-up for the proposed two-stage crowdsourcing process and running a large-scale deployment.

In addition, we hope that the data and analytics obtained in this way will serve as incentive in two directions: to motivate more people to take part in crowdsourcing, and, more importantly, to bootstrap a process whereby better metadata about dataset citation are created earlier in the process, at the source: by authors and publishers.

References

Aitamurto, T. (2012). Crowdsourcing for Democracy: New Era in Policy-Making. Retrieved October 31, 2014, from http://cddrl.fsi.stanford.edu/publications/crowdsourcing_for_democracy_new_era_in_policymaking/

Ball, A., & Duke, M. (2012). Data Citation and Linking. DCC Briefing Papers. Edinburgh: Digital Curation Centre. Retrieved 31 October, 2014, from http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/data-citation-and-linking

Berendt, B., Hollink, L., Hollink, V., Luczak-Rösch, M., Möller, K., & Vallet, D. (2011). Usage Analysis and the Web of Data. ACM SIGIR Forum, 45(1), 63–69. doi:10.1145/1988852.1988864

Berendt, B. (2014, September). "Stories" in data and the roles of crowdsourcing - views of a Web miner. Talk in the Brown Bag Seminar of An Foras Feasa (Institute for Research in Irish Historical and Cultural Traditions) in conjunction with the Programmable Cities Project, National University of Ireland Maynooth. Retrieved from http://people.cs.kuleuven.be/~bettina.berendt/Talks/berendt_2014_09_26.pptx

Berners-Lee, T. (2006). Linked Data - Design Issues. Retrieved October 31, 2014, from http://www.w3.org/DesignIssues/LinkedData.html

Bourne, P. E., Clark, T., Dale, R., de Waard, A., Herman, I., Hovy, E. H., & Shotton, D. (2012). Force11 White Paper: Improving The Future of Research Communications and e-Scholarship. Retrieved from http://www.force11.org/white_paper

Dawson, R., & Bynghall, S. (2012). Getting Results from Crowds: Second Edition: The Definitive Guide to Using Crowdsourcing to Grow Your Business (2nd ed., p. 242). Advanced Human Technologies, Inc.

Dragan, L., Luczak-Roesch, M., Simperl, E., Berendt, B., & Moreau, L. (2014). Crowdsourcing data citation graphs using provenance. In Proceedings of the Workshop on Provenance Analytics (ProvAnalytics2014). Retrieved from http://provenanceweek.org/2014/analytics/papers/1-1.pdf

Force11. (2014, February). Joint Declaration of Data Citation Principles. Retrieved October 31, 2014, from https://www.force11.org/datacitation/

Howe, J. (2006, June). Crowdsourcing: A Definition. Retrieved October 31, 2014, from http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html

Kitchin, R. (2014). The real-time city? Big data and smart urbanism. GeoJournal, 79(1), 1–14. doi:10.1007/s10708-013-9516-8

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing User Studies with Mechanical Turk. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 453–456). New York, NY, USA: ACM. doi:10.1145/1357054.1357127

Luczak-Roesch, M., Tinati, R., Simperl, E., Van Kleek, M., Shadbolt, N., & Simpson, R. (2014). Why won't aliens talk to us? Content and community dynamics in online citizen science. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8092/8136

Marcus, A. D. (2014, October). The Ethics of Experimenting on Yourself. The Wall Street Journal. Retrieved October 31, 2014, from http://online.wsj.com/articles/the-ethics-of-experimenting-on-yourself-1414170041

Mendes, P., Jakob, M., Garcia-Silva, A., & Bizer, C. (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. In Proceedings of I-SEMANTICS2011.

Michigan State University Libraries (2014, October). How to Cite Data. Retrieved October 31, 2014, from http://libguides.lib.msu.edu/citedata

Moreau, L., & Missier, P. (2013, April). PROV-DM: The PROV Data Model. Retrieved October 31, 2014, from http://www.w3.org/TR/prov-dm/

Morsey, M., Lehmann, J., Auer, S., & Ngonga Ngomo, A.-C. (2011). DBpedia SPARQL Benchmark – Performance Assessment with Real Queries on Real Data. In L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, … E. Blomqvist (Eds.), The Semantic Web – ISWC 2011 (Vol. 7031, pp. 454–469). Springer Berlin Heidelberg. Doi:10.1007/978-3-642-25073-6_29

OpenScientist. (2013, April). Research with a Humble Confidence - A Key to Citizen Science Ethics. Retrieved October 31, 2014, from http://www.openscientist.org/2013/04/research-with-humble-confidence-key-to.html

Packer, H. S., Dragan, L., & Moreau, L. (2014). An Auditable Reputation Service for Collective Adaptive Systems. In D. Miorandi, V. Maltese, M. Rovatsos, A. Nijholt, & J. Stewart (Eds.), Social Collective Intelligence: Combining the Powers of Humans and Machines to Build a Smarter Society. Springer.

Peroni, S., & Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. Journal Of Web Semantics: Science, Services And Agents On The World Wide Web, 17, 33-43.

Shotton, D., & Peroni, S. (2010, October). Introducing the Semantic Publishing and Referencing (SPAR) Ontologies. Retrieved October 31, 2014, from http://opencitations.wordpress.com/2010/10/14/introducing-the-semantic-publishing-and-referencing-spar-ontologies/

Silberman, M. S., Irani, L., & Ross, J. (2010). Ethics and Tactics of Professional Crowdwork. XRDS, 17(2), 39–43. doi:10.1145/1869086.1869100

Simperl, E., Cuel, R., & Stein, M. (2013). Incentive-Centric Semantic Web Application Engineering. Synthesis Lectures on the Semantic Web: Theory and Technology, 3(1), Morgan & Claypool Publishers. ISBN: 9781608459964

Trinity College Dublin Communications Office. (2013, September). Letters of 1916 Research Project Calling on Public to Contribute Family Letters. Retrieved October 31, 2014, from https://www.tcd.ie/Communications/news/pressreleases/pressRelease.php?headerID=3248

Williamson, V. (2014). On the Ethics of Crowd-sourced Research. Retrieved October 31, 2014, from http://scholar.harvard.edu/files/williamson/files/mturk_ps_081014.pdf