A Lightweight Provenance Pingback and Query Service for Web Publications

Tom De Nies[1(✉)], Robert Meusel[2], Dominique Ritze[2], Kai Eckert[2], Anastasia Dimou[1], Laurens De Vocht[1], Ruben Verborgh[1], Erik Mannens[1], and Rik Van de Walle[1]

[1] Ghent University - iMinds - Multimedia Lab, Ghent, Belgium
{tom.denies,anastasia.dimou,laurens.devocht,
ruben.verborgh,erik.mannens,rik.vandewalle}@ugent.be
[2] Research Group Data and Web Science,
University of Mannheim, Mannheim, Germany
{robert,dominique,kai}@informatik.uni-mannheim.de

Abstract. Web resources, such as publications, datasets, pictures and others can be directly linked to their provenance data, as described in the specification about Provenance Access and Query (PROV-AQ) by the W3C. On its own, this approach places all responsibility with the publisher of the resource, who hopefully maintains and publishes provenance information. In reality, however, most publishers lack incentives to publish the provenance of resources, even if the owner would like such information to be published. Currently, it is very intricate to link existing resources to new provenance information, either provided by the owner or a third party. In this paper, we present a solution for this problem by implementing a lightweight, read/write provenance query service, integrated with a pingback mechanism, following the PROV-AQ recommendation.

1 Introduction

Provenance is an essential part of trust and value assessment of web content, as it describes everything involved in producing this content. The PROV-AQ document [KGM+13] describes several options to access provenance:

- providing a *link header* in the HTTP response of the resource
- providing a *link element* in its HTML representation
- providing a `prov:has_provenance` *relation* in its RDF representation

In all these cases, however, the representation of the resource is directly linked to its corresponding provenance, so that only the publisher of the resource is in control of which provenance information is provided. This type of *"packaged"* solution gives rise to multiple issues, particularly when the owner of the resource is not in control of the publication process. In this paper, we will focus on the domain of *scientific publishing* since it is a striking example showing this

B. Ludäscher and B. Plale (Eds.): IPAW 2014, LNCS 8628, pp. 203–208, 2015.
DOI: 10.1007/978-3-319-16462-5_16

characteristic. Furthermore, the need of providing additional provenance information in this domain has long been identified [DF08,ZGSB04].

In the domain of scientific publishing, the resource (usually a PDF document) is published by the publisher, whereas its provenance (e.g. datasets, processes, and/or software used) is generally controlled by the author. Besides provenance information created at publication time, additional information such as pointers to corrections or derivations – forward-links in the provenance chain – should be added to enhance the value and the trustworthiness of the resource. The process of most publishers is currently not designed for this kind of updates, as they do not include information about the creation process at all. For example, an empirical study for economics journals shows that of all 141 considered journals, over 70 % do not have any policy dealing with the data used in the journal publications [Vla13].

While general approaches to store and query workflow provenance have been introduced, c.f. [DWW+11,DMMM11,GJM+06], these solutions date from before the publication of the W3C PROV standard, and/or constitute highly customized architectures. Additionally, in these solutions, the responsibility for publishing the provenance still lies either with the author or publisher, with no method to establish a *pingback* or *backlink* to the other party. Despite the PROV-AQ description [KGM+13] and the possibility to apply basic technologies, to the best of our knowledge a lightweight, distributed solution has not been implemented yet.

A possible, fully distributed solution to this problem is the concept of *provenance pingback*, as introduced in PROV-AQ. Provenance pingback enables the establishment of forward-links, e.g. to get to know which resources are based on a certain resource or who makes use of the resource. This solution, however, also highly relies on the goodwill and technological know-how of publishers to provide a pingback URI. Additionally, this would require the publishers to implement a management system aiding in the decision of which provenance is accepted to be published with the associated resource(s). These facts justify the clear need for a *lightweight* and *flexible* solution, in the form of an independent service. An independent service has the advantage that it does not rely on the cooperation of the publishers and enables all authors to use this service. The distributed nature of the Semantic Web makes this technically possible. Such a service needs to allow the storage and retrieval of provenance links for published resources, thereby enriching them with information that is otherwise hard to expose. PROV-AQ defines a mechanism for this concept, named *provenance query services*.

In the following, we introduce our implementation of such a service targeted at the domain of scientific publishing (Sect. 2). Further, we show the advantages of our solution in this application domain (Sect. 3). We discuss the presented approach within Sect. 4 and finish the paper with the conclusion.

2 Lightweight Distributed Provenance Service

We propose a lightweight, RESTful web service for linking resources published on the Web with their provenance information. The solution allows pushing

and querying of provenance information. This way, a seamless integration with existing publication management systems, such as *Research Gate*, *Mendeley*, *Google Scholar*, etc., is achieved. Figure 1 shows the process diagram of our service.[1] If possible, the publisher should support a provenance service by linking to it using a *pingback URI* and *provenance query service URI* as specified in PROV-AQ, but this is not a strict prerequisite. Note that in Fig. 1, both these URIs are represented by the `prov_service_uri`.

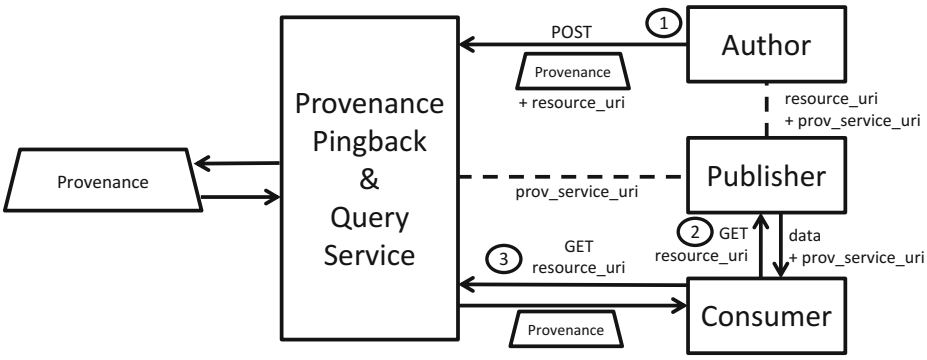

Fig. 1. The process diagram of our provenance pingback and query service.

1. An author **POST**s provenance about a published resource, identified by the `resource_uri`, to a service, identified by the `prov_service_uri`. Both, the `resource_uri` and the `prov_service_uri` are forwarded to the publisher.
2. A consumer requests (**GET**) the publication with the `resource_uri` at the publisher and gets the data about the publication resp. the publication itself. Ideally (but not necessarily), the publisher of the resource provides the whole service as a *pingback URI*. This way, whenever consumers access the resource through the publisher, they are provided with the proper `prov_service_uri`, at which the provenance can be found. Note that if the publisher does not provide a `prov_service_uri`, this does not prevent the author from posting his/her provenance to a service of his/her choice (e.g., where provenance of the same domain is collected). We briefly elaborate on alternatives in Sect. 4.
3. With both, the `prov_service_uri` and the `resource_uri`, the consumer `GET`s all additional provenance information of the resource provided by the author. Using the PROV Data Model allows users to provide and retrieve provenance of the resource as a whole, as well as the provenance of certain sub-parts of the publication, such as data, code, etc.

[1] A live demonstration of this service can be accessed at http://git2prov.org/prov-pings.

3 Application Domain

Application domains that illustrate the merit of provenance query services include, but are not limited to: online news, blogs, digital books, code repositories, and data sets. In the following, we describe use cases that illustrate the different benefits provided by such a query service in our chosen domain of *scientific publications*:

Increase the trust in published results: In the area of scientific publications the typical metadata provided by the publishers are information about the authors, the proceedings or book where the publication can be found and temporal information as the year and month of the publication. It is metadata about the finalized publication, not metadata about the creation process. The metadata describing the process – the provenance data, as provided by a provenance query service – is much richer, revealing not only publications that the author has used to compile the text, i.e., the references, but also additional information about the original research data used, the methodology and the configurations of experiments to derive the results. The availability and verifyability of this information contributes to the trust in the published results.

Find related work: Beyond building trust in a specific publication, the provenance data also helps to identify *related work*, in this case work that uses the same original data or the same method. Results obtained on the same data are much more comparable. Applications of the same method on different data can demonstrate the general applicability of an approach. Contradicting interpretations of data can be found simply by the fact that both interpret the same data. Currently, information about original data can only be derived by reading the publications, which make it very time consuming or even practically impossible to find all relevant publications. With proper provenance data, this becomes trivial. To support this use case, our service specifically supports the submission of links between publications and used datasets by third parties, e.g. by an (semi-) automated process as described by Boland et al. [BREM12].

Update and link to future work: Although the authors as well as the publishers are making huge efforts to create a final, perfect and error-free version of a publication, it happens that published results are superseeded by future work, not to mention actual corrections in the case of errors identified after the publication. Minor updates of applied methods, adoptions to newer datasets or application versions, as well as errors in the code, dataset and process happen more often than not. Even when the additions to existing work lead to a new publication, it is not trivial to find this newer publication. Smaller corrections, however, often do not even result in a proper new publication and an author has no reasonable way to add something to already published work. A provenance query service including the capacity of a pingback overcomes these problems, as the author is able to point to a newer, updated version of a publication. Such forward links in the provenance chain are not limited to the original author, in fact everyone can indicate that a later work builds on top of the publication.

4 Discussion and Future Work

To realize the full potential of our approach, there are a number of considerations to be made for its integration.

The first issue to be considered is the *author verification & curation*. When a third party provides provenance information of a resource, this provenance might be inaccurate or even harmful when used to assess the trustworthiness of the resource. In order to prevent this, a form of verification should be deployed by the author upon the submission of provenance information. An already practiced solution, which is also applicable for scientific publications, is the approval of the email address which is usually associated with the publications of an author. This mechanism is exemplarily used by *Google Scholar*. Alternatively, an authorship claiming mechanism similar to http://authorclaim.org could be implemented. Here, authors of information linked to provenance can claim ownership of the published provenance as well.

Another issue is the tracking of *provenance of the provenance*. Within a system where anyone can make claims about any resource, keeping track of the origin of submitted information and the evolution is crucial. Possible mechanisms to overcome this, can be found in version control systems, from which the provenance information can then be extracted using a mapping service such as Git2PROV [DNMV+13]. A similar mapping could also support the resolution of the *provenance authoring* issues. Needles to say, that such a service needs an user-friendly way to specify provenance information, otherwise the obstacle of getting started will prevent authors and publishers to adapt the service.

At last, the question remains what happens when the publisher does not play along and refuses to publish the link to a provenance service. A single, global provenance service is neither realistic nor desirable. Whereas a peer-to-peer communication between provenance services could be a possibility, a more straight-forward solution would be a registry for provenance services or a dedicated search engine functioning as main entry point to provenance information. The investigation of all these issues remains future work.

5 Conclusion

We have shown that the wide-spread provision of provenance query services will be a useful addition to the Web. We illustrated this by implementing such a service for the domain of online (scientific) publications, where it has important implications regarding discoverability and reproducibility. Provenance information can not only increase the trust in the published results, it also allows the retrieval of publications that share parts of their provenance, most importantly publications that use the same research data. The same holds for future publications that build on current ones.

We believe these services will form an essential step towards a distributed Web of publications, where the provenance provides the silk to make it sustainable and trustable.

Acknowledgments. The research activities in this paper were funded by Ghent University, iMinds (by the Flemish Government), the IWT Flanders, the FWO-Flanders, and the European Union.

References

[BREM12] Boland, K., Ritze, D., Eckert, K., Mathiak, B.: Identifying references to datasets in publications. In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) TPDL 2012. LNCS, vol. 7489, pp. 150–161. Springer, Heidelberg (2012)

[DF08] Davidson, S.B., Freire, J.: Provenance and scientific workflows: challenges and opportunities. In: Proceedings of the International Conference on Management of Data (SIGMOD), pp. 1345–1350. ACM, New York (2008)

[DMMM11] Ding, L., Michaelis, J., McCusker, J., McGuinness, D.L.: Linked provenance data: a semantic Web-based approach to interoperable workflow traces. Future Gener. Comput. Syst. **27**(6), 797–805 (2011)

[DNMV+13] De Nies, T., Magliacane, S., Verborgh, R., Coppens, S., Groth, P., Mannens, E., Van de Walle, R.: Git2PROV: exposing version control system content as W3C PROV. In: Proceedings of the Posters & Demonstrations Track within the12th International Semantic Web Conference (ISWC), pp. 125–128. CEUR-WS, Aachen (2013)

[DWW+11] Dalman, T., Weitzel, M., Wiechert, W., Freisleben, B., Noh, K.: An online provenance service for distributed metabolic flux analysis workflows. In: Proceedings of the 9th European Conference on Web Services (ECOWS), pp. 91–98. IEEE Computer Society, Washington, DC (2011)

[GJM+06] Groth, P., Jiang, S., Miles, S., Munroe, S., Tan, V., Tsasakou, S., Moreau, L.: An Architecture for Provenance Systems. Technical report, University of Southampton, February 2006

[KGM+13] Klyne, G., Groth, P., Moreau, L., Hartig, O., Simmhan, Y., Myers, J., Lebo, T., Belhajjame, K., Miles, S.: PROV-AQ: Provenance Access and Query, W3C (2013)

[Vla13] Vlaeminck, S.: Data management in scholarly journals and possible roles for libraries-some insights from edawax. Liber Quart. J. Assoc. Eur. Res. Libr. 23(1), 48–79 (2013)

[ZGSB04] Zhao, J., Goble, C.A., Stevens, R., Bechhofer, S.: Semantically linking and browsing provenance logs for e-science. In: Bouzeghoub, M., Goble, C.A., Kashyap, V., Spaccapietra, S. (eds.) ICSNW 2004. LNCS, vol. 3226, pp. 158–176. Springer, Heidelberg (2004)