Provenance-Aware Faceted Search in Drupal

Zhenning Shangguan, Jinguang Zheng, and Deborah L. McGuinness

Tetherless World Constellation,
Computer Science Department, Rensselaer Polytechnic Institute,
110 8th Street, Troy, NY 12180, U.S.A.
{shangz,zhengj3,dlm}@cs.rpi.edu

Abstract. As the web content is increasingly generated in more diverse situations, provenance is becoming more and more critical. While a variety of approaches have been investigated for capturing and making use of provenance metadata, arguably no single best-practice approach has emerged. In this paper, we investigate an approach that leverages one of the most popular content management systems – Drupal. More specifically, we study how provenance metadata can be captured and later published as structured data on the Web using Drupal. We also demonstrate how provenance metadata can be used to facilitate faceted search in Drupal.

Keywords: Provenance, Faceted Search, PML, Drupal.

1 Introduction

Information on the Web is increasingly generated using a wide variety of diverse sources. It is also pointed out in [1] that capturing provenance both within and across systems, and publishing that provenance provides potential for many benefits, such as tracing audit trails of data, reproducing scientific experimental results, finding useful information, evaluating data quality, establishing information accountability, etc.

Although there have been numerous previous research efforts aimed at representing and tracking provenance in both closed and open systems, even at different levels of granularity, many of the approaches have limitations and inflexibilities that result from decisions made from targeting a specific system, application, or scenario. Moreover, historically provenance research has often focused on capturing provenance metadata; currently there is an increasing interest in studying how provenance can be used in different ways.

With the goal of exploring the issues mentioned above, this short paper describes some of our ongoing efforts related to provenance using a Drupal-based solution. Our proposed demonstration will highlight two areas:

- Representing and publishing provenance metadata. We are exploring the configurability and extensibility of Drupal. Our approach is to create a provenance-aware Drupal-based platform for web applications.
- Providing provenance-aware faceted search. We are designing and implementing a Drupal-based faceted search that can search over metadata and use provenance to help inform search results and help filter results. We will demonstrate

D.L. McGuinness, J.R. Michaelis, and L. Moreau (Eds.): IPAW 2010, LNCS 6378, pp. 142–147, 2010.

how provenance-aware search can be more efficient, and provide insight into ranking and presentation options. We also will expose how provenance facets, such as temporal facets related to the creation and modification time of some Drupal content, are being used in our search functionality.

The rest of the paper is organized as follows. Section 2 highlights the related work. Section 3 demonstrates our initial effort to capture, encode, and publish provenance information as structured RDF. Section 4 describes how to make use of these provenance metadata to facilitate faceted search. Finally, we conclude the paper and outline some future work in section 5.

2 Related Work

Four areas of work are considered related to the topic of our paper.

There is a diverse literature on systems and applications that are capable of capturing provenance metadata, for example, Taverna [2], VisTrails [3], REDUX [4], Pegasus [5], Karma [6]. While they successfully demonstrate different approaches of capturing provenance, they can be viewed as having an application-dependent nature and thus they can be less flexible and less extensible when applications differ greatly from those that these approaches were designed to satisfy. Our work differs from these approaches in that we are investigating mechanisms to capture and publish provenance using Drupal, which has the potential to serve as the foundation of an application-independent solution.

Another spectrum of related research is the generic provenance models and vocabularies, most notably the Open Provenance Model (OPM) [1], Provenance Markup Language (PML) [7], and the Provenance Vocabulary [8]. While providing different vocabularies for representing provenance, there are certain conceptual overlaps between them. For example, both OPM and Provenance Vocabulary have similar basic concepts, i.e., Agents (OPM) and Actors (Provenance Vocabulary) denoting people, Processes (OPM) and Executions (Provenance Vocabulary) representing executions of actions or processes, and Artifacts (both OPM and Provenance Vocabulary) representing the entity produced or manipulated. Currently, we are using PML to encode the provenance metadata. However, supporting provenance representations using different domain-independent provenance vocabularies is planned as one of our future work areas.

Faceted search for exploration has been widely studied over the past years. Numerous research efforts [9] [10] [11] have demonstrated benefits including usability and flexibility of faceted browsers when interacting with structured data (e.g., relational databases, XML, RDF). In this paper, we are making use of the generated provenance metadata to facilitate faceted search and help locate the desired information.

Drupal, one of the most popular CMS systems, has been widely deployed. Recently, there is an emerging effort from both the Semantic Web community and the Drupal development community to enable Drupal to publish semantic metadata (e.g., RDF, RDFa) [12]. Our implementation makes use of some of the Drupal modules introduced in [12] to create provenance related node fields. We are also developing Drupal module of our own to publish provenance metadata encoded in PML.

3 Capturing and Publishing Provenance in Drupal

The information that Drupal[1] organizes and manages is called *content*. Usually, a piece of content in Drupal corresponds to a single *node* (in the form of a *page*) that has a title, an optional body text description, and perhaps several additional *fields*. Every node also belongs to a particular *content type*, such as Person, Blog and etc. The site administrator uses the Content Construction Kit (CCK)[2] to create a node by specifying its content type (e.g., Person), title (e.g., name of the person), body text (e.g., short bio of the person), and fields (e.g., first name, last name, email of the person). Furthermore, the RDF CCK[3] module extends the functionality of CCK to enable the definition of mappings between: 1) a specific content type (e.g., Person) and an RDF class (e.g., foaf:Person), and 2) a node field (e.g., field_firstname) and an RDF property (e.g., foaf:givenName).

Currently, our initial implementation of the Drupal provenance module is capable of capturing provenance at two levels of granularity.

- Node-level: This level focuses on the provenance metadata associated with a node in Drupal, such as who created the node, when the node was first created, and when the node was last modified.
- Content-level: This level keeps track of the provenance metadata associated with all the revisions of a node in Drupal, such as who modified the body text of the node and changed the values of the fields, when these modifications and changes happened, as well as the texts/values after every revision.

The benefit of having both the node-level and content-level provenance is that the former captures the basic provenance metadata about the node while the latter keeps track of the detailed edit history information.

The implementation of capturing node-level provenance is straightforward: we can use CCK and RDF CCK to define several provenance related fields and map them to the RDF properties from our chosen provenance vocabulary. Currently, we define two fields for every node in Drupal and map them to *pmlp:hasCreationDateTime* and *pmlp:hasModificationDateTime*, representing when the node was first created and last modified respectively.

In contrast to capturing node-level provenance, keeping track of the content-level provenance is not natively supported in Drupal. Thus we develop our own provenance module[4], making use of various Drupal core hook function and core API.

Besides capturing provenance metadata, our module is also capable of publishing these metadata as RDF on the Web. To access both the node-level and the content-level provenance metadata, users can follow the URL pattern *http://your_domain_name/drupal-dir/node/node-id/pml* for every node in a standard Drupal installation

[1] We are using Drupal version 6.16 and PHP 5.2 at the time of this writing. Unless explicitly stated, this holds for all the discussions throughout this paper.

[2] Content Construction Kit (CCK) module: http://drupal.org/project/cck.

[3] RDF CCK module: http://drupal.org/project/rdfcck.

[4] The source code for our provenance module can be found at http://tw2.tw.rpi.edu/drupal-dev/provenance.zip.

without clean URLs[5] turned on. Currently we are only using PML to represent the provenance metadata, with support for OPM and Provenance Vocabulary in progress. Figure 1 shows the exported provenance metadata for a node in our experimental Drupal installation.

Fig. 1. Exported Provenance Metadata for a Drupal node

4 Provenance-Aware Faceted Search in Drupal

Being able to capture both the node-level and content-level provenance immediately brings about a lot of benefits, especially for provenance-ware search in Drupal. Currently our provenance-aware faceted search is implemented using the Exhibit module[6] of Drupal. More specifically, we leverage the functionality of it to build the faceted search interface, with some facets generated from the node-level provenance metadata, such as *pmlp:hasCreationDateTime* and *pmlp:hasModificationDateTime*. Our initial implementation is demonstrated in Figure 2, with the "Creation Time" facet generated from the *pmlp:hasCreationDateTime* node field.

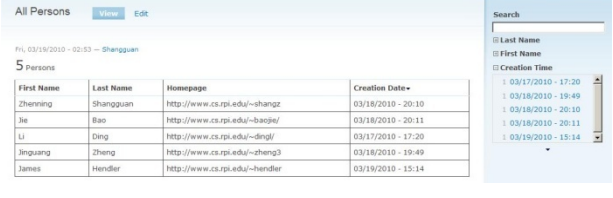

Fig. 2. Initial implementation of provenance-aware faceted search

5 Conclusion and Future Work

In this paper we described some of our ongoing effort to capture and publish provenance metadata in Drupal. With the help of various Drupal modules, such as CCK and

[5] Drupal clean URLs: http://drupal.org/getting-started/clean-urls. The provenance metadata can also be accessed via a similar URL pattern by appending "/pml" to the end of a node URL.

[6] Exhibit Drupal module: http://drupal.org/project/exhibit.

RDF CCK, node-level provenance can be captured by defining node fields and establishing mappings between them and RDF properties in the chosen provenance model, which is PML in our case. We also developed a provenance module in order to capture content-level provenance, which can be used to trace the revision history of nodes in a deployed Drupal website. Finally, our initial effort to facilitate faceted search with the help of the captured provenance metadata is also presented. We plan to provide a demonstration of our provenance-aware faceted search in one (or more) of our eScience applications at IPAW.

We identify two general directions for future work. First, our current implementation supports provenance encodings using only PML, without giving the user the ability to choose which provenance model to use. We plan to further enable the user to specify their desired provenance model when exporting the provenance metadata. Second, at present only the node-level provenance is exploited in the faceted search. However, content-level provenance might also contain useful properties to generate facets, such as revision time, date, and the user who made that revision. Moreover, content-level provenance can also be used to generate the edit history of the contents managed by Drupal. Therefore, another important aspect of our future work is to take advantage of the content-level provenance to further improve faceted search and generate edit history information of the nodes (pages) in Drupal. We also plan to leverage provenance data to help filter search results when result sets are large.

References

1. Moreau, L.: The Foundations for Provenance on the Web. J. Foundations and Trends in Web Science (2009)
2. Zhao, J., Goble, C.A., Stevens, R., Turi, D.: Mining Taverna's semantic web of provenance. Concurrency and Computation: Practice and Experience 20(5), 463–472 (2008)
3. Freire, J., Silva, C.T., Callahan, S.P., Santos, E., Scheidegger, C.E., Vo, H.T.: Managing rapidly-evolving scientific workflows. In: Moreau, L., Foster, I. (eds.) IPAW 2006. LNCS, vol. 4145, pp. 10–18. Springer, Heidelberg (2006)
4. Barga, R.S., Digiampietri, L.A.: Automatic capture and efficient storage of e-science experiment provenance. Concurrency and Computation: Practice and Experience 20(5), 419–429 (2008)
5. Kim, J., Deelman, E., Gil, Y., Mehta, G., Ratnakar, V.: Provenance trails in the wings/pegasus system. Concurrency and Computation: Practice and Experience 20(5), 587–597 (2008)
6. Simmhan, Y.L., Plale, B., Gannon, D.: Karma2: Provenance management for data-driven workflows. Int. J. Web Service Res. 5(2), 1–22 (2008)
7. McGuinness, D.L., Ding, L., Pinheiro da Silva, P., Chang, C.: PML2: A modular explanation Interlingua. In: ExaCt (2007)
8. Hartig, O., Zhao, J.: Using Web Data Provenance for Quality Assessment. In: Proceedings of the 1st Int. Workshop on the Role of Semantic Web in Provenance Management (SWPM) at ISWC, Washington, DC, USA (2009)
9. Hildebrand, M., van Ossenbruggen, J., Hardman, L.: /facet: A Browser for Heterogeneous Semantic Web Repositories. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 272–285. Springer, Heidelberg (2006)

10. Oren, E., Delbru, R., Decker, S.: Extending Faceted Navigation for RDF Data. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 559–572. Springer, Heidelberg (2006)
11. Schraefel, M.C., Karam, M., Zhao, S.: mSpace: interaction design for user-determined, adaptable domain exploration in hypermedia. In: AH 2003: Workshop on Adaptive Hypermedia and Adaptive Web Based Systems, pp.21–235 (2003)
12. Corlosquet, S., Delbru, R., Clark, T., Polleres, A., Decker, S.: Produce and Consume Linked Data with Drupal. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 763–778. Springer, Heidelberg (2009)