Provenance Support for Content Management Systems: A Drupal Example

Aída Gándara and Paulo Pinheiro da Silva
agandara1@miners.utep.edu, paulo@utep.edu*

University of Texas at El Paso, Computer Science Department,
El Paso, Texas 79968, USA

Abstract. Provenance helps with understanding data but without proper tools to share and access content, its reusability is limited. This paper describes the CI-Server framework currently being used to help scientific teams seamlessly share data and provenance about scientific research. CI-Server has been built using Drupal, a content management server workbench, with a focus on publishing and understanding the semantic content that is now available over the Web. By focusing on an open framework, scientists publish provenance related to their scientific research then leverage the semantic knowledge to understand and visualize the information.

1 Introduction

Regardless of how useful provenance is for capturing knowledge related to scientific research, how provenance is managed, e.g. how to access provenance-related information, can greatly affect its reusability. For example, for some scientists, research is performed on a single workstation and the results, data and data-related provenance, are stored on the same system. Consequently, most information including provenance is restricted to only scientists with specific privileges to access that workstation. As a result of such isolated environments, data and provenance are not shared. Web portals, Web sites focused on collecting and sharing data and resources, normally within a particular domain, provide a solution for scientists to share their data and make it available to other scientists. For example, the Earthscope Data Portal[3] is a Web portal built to enable sharing and discovery of geological data. One drawback to portals is that in many cases publishing data on them is a manual process; a user interactively uploads directly to a portal location or requests administrative support from the portal's webmaster. Portals can be quite unique in their presentation and usage, i.e. the management of resources and the process of uploading files is different for every portal. As searches span across multiple Web portals, scientists are forced into understanding the multiple cultures of different Web portals. The GEO portal[4], for example, has a similar focus as the Earthscope portal, yet the interface and

* This work was supported in part by DHS grant 2008-ST-062-000007 and by NSF grant HRD-0734825.

structure of the site is quite different. The distinct steps of understanding the culture of the portal and manually uploading information can be distracting to scientists' needs to share and discover information. One proposed solution is to unify several related data archives onto one, e.g. the Earthscope portal. Our solution is to focus more on building a structure into content management servers so tools can access the needed information, data and provenance-related data, without having to access one portal or understand the nuances of each portal's interface.

2 Provenance Support for Drupal

Drupal is an open source content management server framework used to build Web sites and Web portals. The tool supports user security and multiple levels of configuration, e.g. menu calls, forms, and event hooks. The Drupal Development Community is currently a very active component of Drupal because developers share software solutions that can be enabled in different Drupal implementations.[2] We have built additional functionality in Drupal, providing provenance support based on the PML notation[5] in an open-portal based infrastructure we call the CI-Server. Our implementation uses and extends various modules provided by the Drupal Development Community in an effort to facilitate the sharing of information and reuse of provenance for scientific research.

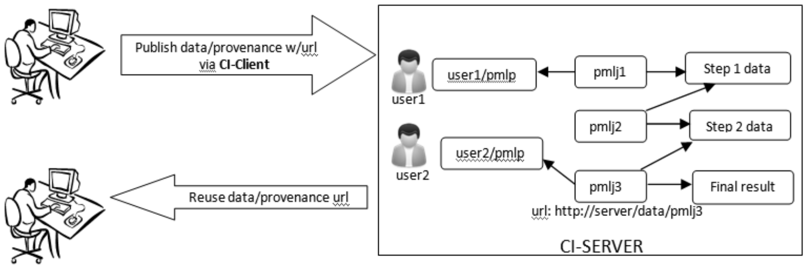

Fig. 1. The CI-Server Framework enables sharing of provenance.

Figure 1 illustrates the provenance support enabled by the CI-Server framework. Via a CI-Client module and CI-Client API that extend and expose internal server functionality, scientific tools can be enabled to seamlessly publish data and provenance (pmlj) files, moving data from the scientist's workstation to any Drupal based CI-Server. This avoids the manual step of file uploads or the nuances of understanding different portal interfaces. In Figure 1, the top scientist has published pmlj1, pmlj2, pmlj3, called a PML nodeset, and some corresponding data. The pmlj documents are semantic documents, written in OWL. These documents are built with knowledge about how a scientific process occurred and they rely on links to available resources, e.g. data. The CI-Server uses modules to support file management and url aliasing, enabling users to upload content and then access it via url links. Since the pmlj documents contain references to entities that it is capturing provenance about, provenance knowledge is immediately available to traverse as a knowledge set. Furthermore, the CI-Server manages content and information that is often useful in capturing provenance. For example, user content on Drupal can be used to document source related information, e.g. who published a data file. The CI-Server builds pmlp nodes for users on the system. To see a user's public information, the user's pmlp page would be accessed via a dynamically created OWL-based pmlp node. Building provenance dynamically with internal CI-Server knowledge, aids in the collection of provenance and avoids scientists from having to supply that information repetitively. Figure 1 shows that pmlj1 captures the knowledge that user 1 was involved in creating Step 1 data and that pmlj2 captures the knowledge that Step 1 data was an input to create Step 2 data and finally that pmlj3 captures the knowledge that user 2 and Step 2 data were used to create the Final result. Because pmlj3 is identified with an URL, the scientist reusing the information can use the url link to access the entire provenance nodeset and visualize it using context-related scientific tools. A PML nodeset, for example, can be visualized via Probe-It[1] by simply providing the nodeset's URI.

References

1. Nicholas Del Rio and Paulo Pinheiro da Silva. Probe-it! visualization support for provenance. In *Proceedings of the Second International Symposium on Visual Computing (ISVC 2)*, pages 732–741, Lake Tahoe, NV, 2007. Springer.
2. Drupal community innitiatives. http://drupal.org/community-initiatives.
3. Earthscope data portal. http://earthscope.data.portal.
4. Geo-portal. http://geoportal.org.
5. Deborah McGuinness, Li Ding, Paulo Pinheiro da Silva, and Cynthia Chang. PML2: A Modular Explanation Interlingua. In *Proceedings of the AAAI 2007 Workshop on Explanation-aware Computing*, Vancouver, British Columbia, Canada, July 22-23 2007.