Open Linked Data in Policy-based Repositories

Chien-Yi Hou
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, United States
chienyi@unc.edu

Richard Marciano
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, United States
richard_marciano@unc.edu

Abstract—**Since the creation of the World Wide Web and the emergence of search engines, the web has become the major media for researchers to explore and share data. In 2009, Berners-Lee started a movement to encourage researchers to put their data on the web. He wanted researchers to open and link their data to the public to increase the use and reuse of content. However, the lack of proper mechanisms to assist researchers publishing data on the web has prevented them from effective sharing. We create mechanisms to help researchers open and link their data using the integrated Rule-Oriented Data System, iRODS. iRODS is a data grid software that has been widely used to manage research data in large-scale European, American, and Asian national research projects, such as at the "Bibliothèque Nationale de France". iRODS comes with a business rule engine which allows researchers to create rules to manage and process data. In this paper, we show how to configure iRODS rules to transmit and open linked data in a distributed data cloud setting.**

Keywords-linked data; policy-based data management.

I. INTRODUCTION

Publishing research results on the web has become the most common approach for researchers to share their research findings because with robust search engines, everything you put on the web gets revealed to the world automatically. That is not necessarily the case though for the data that researchers use to analyze and eventually link to their publications. These data are usually not available after experiments are conducted and might not get retained. The main reason is that these data are stored in digital repositories that are not exposed to the web. In Berners-Lee's TED talk [1] in 2009, he advocated the idea of sharing "raw data now". He encouraged researchers to share not only the study results but also the data to produce the study results. The technology created to support this idea is Linked Data [2]. Linked Data provides the mechanism for researchers to create associations between data or information and other related data or information. It creates the relationships between data and information. It helps aggregate data and information together. Researchers could do a "one-stop shop" to find all the relevant data if they were linked together.

Sharing raw data with the public is an important idea to bring data to life. The same dataset might be used across multiple disciplines and reveal interesting or even unexpected findings. However, the last thing researchers want to do is further burden their daily research. Opening and linking one's data should be so easy and seamless that one almost forgets its existence! In this paper, we would like to show how to automate this process by using policy-based data management software to manage and publish data. The goal is to fit data sharing processes into one's data management cycle seamlessly. In the 2nd section, we will introduce the data grid software used in this paper and discuss how researchers are using it. In the 3rd section, we will show you how to open Linked Data by using rules. In the conclusion, we will discuss our findings and next steps.

II. POLICY-BASED DATA MANAGEMENT

The last four years or so have seen the emergence of policy-aware infrastructure. New collaboratives have emerged with a focus on distributed preservation frameworks that are driven by community-based management policies, comprising auditing, replication of content, automatic extraction and association of metadata, validation of checksums, format migration, and trustworthiness. Policies are typically rules describing actions that take place in repositories. This trend was highlighted in the 2008 Communications of the ACM Magazine [3], where the need for repositories to incorporate mechanisms that implement and automate policies and regulations was identified. Emerging data cyber-infrastructure management environments and systems include well-known and widely-used systems such as LOCKSS, DSpace, Fedora, and iRODS.

The iRODS approach we focus on and discuss in greater detail in this section, supports the notion of extensibility with a scalable rule-based engine, allowing the registration of new server-side distributed user-defined workflows.

A. iRODS Overview

iRODS [4], the Integrated Rule-Oriented Data System, is a community-driven, open source, data grid software that aims to help researchers manage large sets of data. iRODS has been used widely by scientists to manage data in large-scale European, American, and Asian national research projects. It is also used as a distributed file system to manage and share data across different locations. Researchers will

 ISBN: 978-1-61208-220-2

need to have accounts in order to access data stored in iRODS.

B. iRODS Rules

The rule engine inside iRODS gives researchers the capability to specify their data management policies within iRODS. The basic components of an iRODS rule are hooks, conditions, actions, and recovery actions. iRODS hooks are operations that happen during the data manipulation process. There are over 70 hooks now in existence. For example, a hook named *acPostProcForPut* will be triggered after you upload a file. When the rule is triggered, the rule engine will check the conditions of that rule. If the conditions are satisfied, the actions in the rule will be executed. The action could be a single procedure or a chain of procedures. If the actions somehow fail, recovery actions can be executed if specified.

iRODS rules could be executed at three different modes: (1) immediate execution, (2) delayed execution, and (3) periodic execution. For example, you can specify a rule that sends you an email immediately or three minutes after a file is deleted. You also can run a rule every month to verify the integrity of your data. iRODS provides flexible rule design principles and many hooks that span the data management lifecycle. These features provide the opportunity to plug data sharing processes into researchers' data management processes.

C. iRODS Metadata

In order to make Linked Data more useful, it is essential to provide rich information to describe the data, e.g., who the creator of these data is and what these data are about. iRODS provides the capability to create metadata to describe an individual object, a collection, a user account, or even a resource that is used to store the data. You can define your own metadata as AVU (Attribute, Value, Unit) triplets to describe your "subjects". The capability to ingest metadata into iRODS is very useful when we need to extract information to create relationships for Linked Data.

D. iRODS Use Cases

Because of the flexibility to design policies for your own digital repository, iRODS has been used in many research projects. Hedges [5] implements preservation policies on iRODS to manage research data. The SHAMAN project [6] also uses iRODS to design polices for preservation. Walling and Esteva [7] integrate their procedures into iRODS to automate the metadata extraction process while ingesting data. The PoDRI project [8] uses iRODS policies to manage the interoperability between Fedora, Flexible Extensible Digital Object Repository Architecture, and iRODS. Integrating open Linked Data processes with iRODS data management is a new idea.

III. OPEN AND LINKED DATA BY POLICIES

Data grids have been used in international research projects [9][10] to manage large-scale data, but sharing data is mainly restricted within projects or groups. Researchers usually only share their research results with the public as publications but not as raw data. These raw data are invisible to web crawlers and search engines. In order to make these raw data accessible and usable by the general public, we need a mechanism to publish data to the web and automate the open Linked Data process in researchers' daily data management.

iRODS provides researchers the flexibility to incorporate policies and procedures as iRODS rules into their data management routine. This feature gives them the opportunity to integrate data publishing processes with data management processes. The process to publish Linked Data will be triggered automatically without further actions after initially setting it up and thus it becomes part of the data management process. The content to be published as Linked Data is data already stored in iRODS or data that will be ingested into iRODS. First of all, we need to define the actions to trigger the "publish Linked Data" procedure. Let us assume we only publish the data when the data is opened to the "public", then data will be opened when the file's access permission is set to "public". "Public" access permission here means that anonymous users have read permission to the data. This kind of situation could take place in a couple of different scenarios. The first one is when the file's access permission is being changed to "public". The second scenario is when data are ingested to a directory with public access permission. Below are the two open linked data usage scenarios:

- Scenario 1 (Figure 1.): After researchers change the access permission, a rule named *acPostProcForModifyAccessControl* will be triggered. We add a condition to check the access permission. If the permission is set to "public", the rule will initiate the "publish Linked Data" procedure. There should also be corresponding rules to check the permission while researchers remove the public access permission, a "close" Linked Data procedure would need to be called to remove the link from the web.
- Scenario 2 (Figure 2.): A rule named *acPostProcForPut* will be triggered every time researchers upload a file to iRODS. If the access permission of the target directory is set to "public", then the access permission of the uploaded file will be "public" as well. In this case the "publish Linked Data" procedure will be called to publish data to the web.

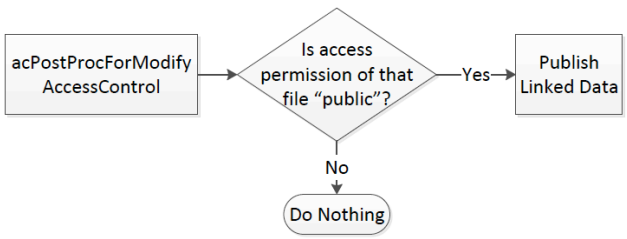

Figure 1. Open linked data scenario 1

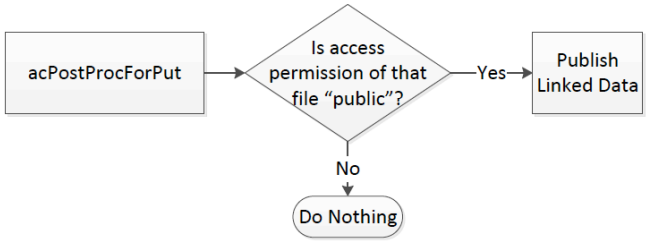

Figure 2. Open linked data scenario 2

According to Bizer [11], there are three basic steps to publish a dataset as Linked Data on the Web. They are: (1) assigning URIs to the entities, (2) setting RDF links to other data sources on the Web, and (3) providing metadata about published data. In order to fulfill these three steps, researchers will need to provide enough information to describe their data. The information will be stored as iRODS user-defined metadata and will be used by the "Publish Linked Data" process in Figure 1 and Figure 2. The advantage of storing this information as user-defined metadata is the flexibility to link data based on user-defined metadata. Different datasets might link to different data sources on the web. iRODS also provides a unique iRODS URI to each individual object that could be used to access to the object. Researchers will use the published Linked Data information to reveal the data's existence and use the iRODS URI to get the data. Data will be stored in iRODS where we will apply other management policies like integrity checking. The details of how the "Publish Linked Data" process works will be included in future work.

IV. CONCLUSION AND NEXT STEPS

We have demonstrated some simple usage scenarios on how to publish Linked Data using policy-based data management software, but there are still many factors that we need to consider when dealing with more complex data. For example, data that require IRB (institutional review board) approvals might need some pre-processing to remove the personal identities or ask other researchers to obtain similar IRB approvals. Policies to avoid accidently violating the privacy of data will need to be defined.

Our next step is to identify required information to describe data in order for it to be published. Different types of data will have different requirements to create relationships, but we would like to find a general set of information that could be used to describe most of the data and create linkages. Additional information could be considered as add-on but is not necessary.

Opening research data to the public is becoming popular and it is an important approach to get the best out of the investments that are used to generate or acquire these data. The technology of Open Linked Data provides researchers a mechanism to share data. By using Policy-based Data Management systems, we will be able to build policies within the system to help researchers publish Linked Data to the web. This approach could encourage researchers to share more raw data because the sharing procedure has been built into data management process and could be modified by researchers when needed. It has the potential to save a lot of effort, encourage the reuse of research data, and open up the field of study.

ACKNOWLEDGEMENTS

This work is supported by a Research and Demonstration grant from the Institute of Museum and Library Services, (IMLS LG-06-09-0184-09) "Policy-Driven Repository Interoperability.

REFERENCES

[1] Berners-Lee, T. (2009) on the next Web, TED talks, 2009(Feb). http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

[2] Bizer, C., Heath, T., Idehen, K., and Berners-Lee, T. (2008). Linked data on the web (LDOW2008). Proceeding of the 17th International Conference on World Wide Web WWW 08, 2008 (September), 1265. ACM Press.

[3] Berman, F., Got Data? A Guide to Data Preservation in the Information Age, Francine Berman, Communications of the ACM, December 2008, Vol. 51, No. 12, pp. 50-56, http://mags.acm.org/communications/200812/

[4] Rajasekar, A., Moore, R., Hou, C.-Y., Lee, C. A., Marciano, R., De Torcy, A., Wan, M., et al. (2010). iRODS Primer: Integrated Rule-Oriented Data System. Synthesis Lectures on Information Concepts Retrieval and Services (Vol. 2, pp. 1-143).

[5] Hedges, M., Hasan, A., and Blanke, T. (2007). Management and preservation of research data with iRODS. Proceedings of the ACM first workshop on CyberInfrastructure information management in eScience CIMS 07 (pp. 17-22). ACM.

[6] Innocenti, P., Ross, S., Elena, M., Wilson, T., Ludwig, J., and Pempe, W. (2009). Assessing digital preservation frameworks: the approach of the SHAMAN project. MEDES 09 Proceedings of the International Conference on Management of Emergent Digital EcoSystems.

[7] Walling, D., and Esteva, M. (2011). Automating the Extraction of Metadata from Archaeological Data Using iRods Rules. International Journal of Digital Curation, 6(2), 253-264.

[8] Pcolar, D., Davis, D. W., Zhu, B., Chassanoff, A., Hou, C. Y., and Marciano, R. (2010). Conceptualizing Policy-Driven Repository Interoperability (PoDRI) Using iRODS and Fedora. Information Sciences, 25. Morgan and Claypool Publishers.

[9] Hoschek, W., Jaen-martinez, J., Samar, A., Stockinger, H., and Stockinger, K. (2000). Data Management in an International Data Grid Project. (R. Buyya and M. Baker, Eds.) Manager, 1971, 77-90. Springer-Verlag.

[10] Johnston, W. E. (2002). Computational and data Grids in large-scale science and engineering. Future Generation Computer Systems, 18(8), 1085-1100.

[11] Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked Data - The Story So Far. (T Heath, M. Hepp, and C Bizer, Eds.) International Journal on Semantic Web and Information Systems, 5(3), 1-22. Elsevier.

 ISBN: 978-1-61208-220-2