The RMap Project: Capturing and Preserving Associations amongst Multi-Part Distributed Publications

Karen L. Hanson
Portico
100 Campus Drive, Suite 100
Princeton, NJ 08540
+1 609-986-2282
karen.hanson@ithaka.org

Tim DiLauro
Johns Hopkins University
3400 N. Charles Street / MSEL
Baltimore, MD 21218
+1 410-929-3722
tim.dilauro@jhu.org

Mark Donoghue
IEEE
445 Hoes Lane
Piscataway, NJ 08854
+1 732-562-6045
m.donoghue@ieee.org

ABSTRACT

The goal of the RMap Project is to create a prototype service that can capture and preserve maps of relationships amongst the increasingly distributed components (article, data, software, workflow objects, multimedia, etc.) that comprise the new model for scholarly publication. The demonstration will provide a tour of some of the features of the initial web service prototype. This will include examples of Distributed Scholarly Complex Objects (DiSCOs) and associated provenance data in RMap, as well as some of the options that users might have for interacting with the framework.

Categories and Subject Descriptors

H.3.5 [**Information Systems**]: Online Information Services – *web-based services, data sharing*. H.3.7 [**Information Systems**]: Digital Libraries – *collection, dissemination*.

General Terms

Management, Documentation, Standardization.

Keywords

Publishing workflows, linked data, data publishing, semantic web, REST API, digital preservation, scholarly communication, digital scholarship

1. BACKGROUND

In recent years, the content that comprises the scholarly record has become more dynamic and less "bounded." Formerly, even digital artifacts of the scholarly record were more or less discrete objects, such as journal articles or books, usually encapsulated in a single file. Increasingly, the primary unit of scholarly communication is evolving into a multi-part distributed object that often includes data and software [3] (see Figure 1).

As one indicator of this trend, we see publishers working to adopt new approaches for dealing with these forms of publication. The movement towards publishing and citing datasets as standalone objects has been particularly prominent in recent years, and is but one component of an ongoing shift in scholarly publishing.

Further, the many, and dynamic, relationships amongst the components comprising a distributed scholarly complex object are first class objects in themselves. Not only does the scholarly community require preservation of publication, data, and other artifacts of scholarly research (whether preserved separately or together), it also requires the preservation of the relationships amongst them. In addition, in the scholarly environment, proper preservation also mandates models and information graphs that account for the provenance of the assertions of those relationships.

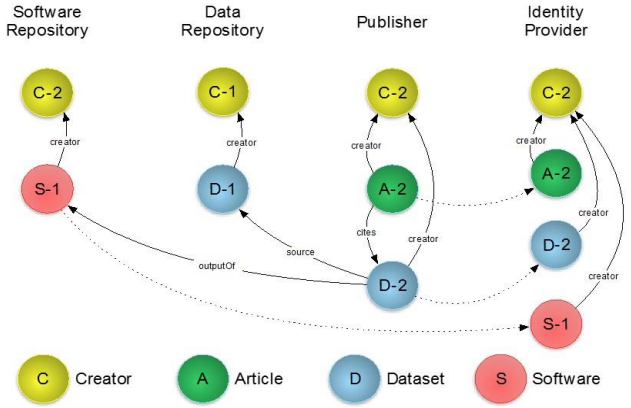

Figure 1 Multi-part Distributed Publication

2. THE RMAP PROJECT

The RMap Project[1] is a two-year Alfred P. Sloan Foundation[2]-funded initiative undertaken by the Data Conservancy[3], Portico[4], and IEEE[5]. The goal of the project is to create a prototype service that can assemble the maps of relationships amongst the distributed components of a modern scholarly publication, and preserve those maps over the long term.

The project builds on the features of the semantic web [1] and linked data[6], adopting concepts from the Open Archives Initiative

JCDL'15, June 21–25, 2015, Knoxville, Tennessee, USA.
ACM 978-1-4503-3594-2/15/06.
http://dx.doi.org/10.1145/2756406.2756952

[1] http://rmap-project.info/rmap/

[2] http://www.sloan.org/

[3] http://dataconservancy.org/

[4] http://www.portico.org/

[5] http://www.ieee.org/

[6] http://www.w3.org/standards/semanticweb/data

Object Reuse and Exchange (OAI-ORE).[7] To simplify integration with publishing and scholarly workflows, RMap employs a RESTful (Representational State Transfer) API [2] and makes use of existing well-known vocabularies (e.g. Dublin Core[8], Friend of a Friend[9], Open Provenance Model[10]) in its data model.

2.1 Objectives

The RMap Service will record and preserve links amongst the artifacts of scholarly communication and those who create, modify, employ, and annotate them. Its purpose in doing so is to facilitate the discovery and reuse of those artifacts, to demonstrate the impact and reuse of research, to make those demonstrations available to those making curatorial decisions about collection and preservation of digital research artifacts such as software and workflows, and to inform those curatorial and other choices with solid provenance information about the representations in RMap.

Key design objectives of the RMap service in support of these goals are to

- support assertions from a broad set of contributors
- integrate with Linked Data
- leverage existing data from other scholarly publishing stakeholders (publishers, identifier providers, identity authorities, data and software repositories)
- provide some support for resources lacking identifiers

2.2 Data Model

Major components of the data model, constructed upon the Resource Description Framework (RDF)[11] concepts of resources, triples and graphs, are *Statements*, *DiSCOs*, *Agents*, and *Events*.

RMap *Statements* are essentially reified RDF triples and map closely to the class of *RDF Statement*. In addition to an identifier they provide status, and Event (provenance) information.

RMap *DiSCOs* (Distributed Scholarly Complex Objects) are graphs representing aggregations of related scholarly resources (see Figure 2). For example: A single DiSCO might represent, an article, its related datasets, and software – as well as any useful context metadata describing those resources. DiSCOs also have an associated identifier, status, and Event (provenance) information.

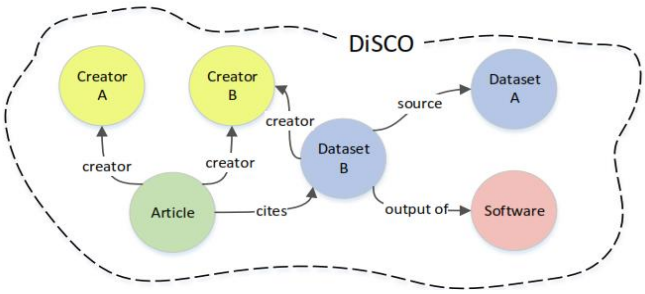

Figure 2. Example of a DiSCO

An RMap *Agent* is a person or thing (or group of these) responsible for some action. In RMap this can include, but is not limited to, authors, funders, publishers, and administrators. RMap distinguishes between scholarly agents (e.g., author, funder, publisher, data processing software) and system agents (RMap component, user, etc.).

An RMap *Event* captures provenance within RMap. This includes the system agent responsible for a particular activity, the activity itself, the timeframe thereof, and the associated context. Linked, for example, to a DiSCO at the time of its creation or update, Events make it possible to trace the provenance of all assertions within and about a DiSCO.

2.3 REST API

The primary interface for accessing the RMap database is a REST API. The features of a RESTful API include programming language independence and conformance to web architecture metaphors. Both are important in facilitating the integration of the RMap service into heterogeneous publisher, researcher, funder, and other institutional workflows.

3. CONCLUSIONS

By being part of publisher, researcher, funder, and other scholarly workflows and by aggregating data from multiple sources, RMap aims to support third party discovery as well as facilitate the capture of information about scholarly artifacts that is not easily captured elsewhere.

4. ACKNOWLEDGMENTS

The RMap Project is funded by the Alfred P. Sloan Foundation. The authors wish to acknowledge the contributions of their RMap project colleagues: Sayeed Choudhury, Johns Hopkins University, The Sheridan Libraries, Associate Dean for Research Data Management; Kate Wittenberg, Managing Director, Portico; Gerry Grenier, Senior Director, Publishing Technologies, IEEE, Portico colleagues Jabin White, Sheila Morrissey, Vinay Cheruku, Amy Kirchhoff, John Meyer, Stephanie Orphan; and IEEE colleagues Renny Guida, Ken Rawson, Ken Moore.

5. REFERENCES

[1] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 28-37.

[2] Fielding, R. T. 2000. "Architectural Styles and the Design of Network-based Software Architectures". Dissertation. University of California, Irvine. Retrieved 26 January 2015 from https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf

[3] Lavoie, B., Childress, E., Erway, R., Faniel, I., Malpas, C., Schaffner, J., and van der Werf, Titia. 2014. The Evolving Scholarly Record. Dublin, Ohio: OCLC Research. Retrieved 26 January 2015 from http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-evolving-scholarly-record-2014.pdf

[7] http://www.openarchives.org/ore/

[8] http://dublincore.org/specifications/

[9] http://xmlns.com/foaf/spec/

[10] http://openprovenance.org/

[11] http://www.w3.org/TR/rdf11-concepts/