Using RMap to Describe Distributed Works as Linked Data Graphs: Outcomes and Preservation Implications

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

Aaron Birkland
Tim DiLauro
Johns Hopkins University
3400 N. Charles Street / MSEL
Baltimore, MD 21218
+1 410-929-3722
apb@jhu.edu
timmo@jhu.edu

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

ABSTRACT

Today's scholarly works can be dynamic, distributed, and complex. They can consist of multiple related components (article, dataset, software, multimedia, webpage, etc.) that are made available asynchronously, assigned a range of identifiers, and stored in different repositories with uneven preservation policies. A lot of progress has been made to simplify the process of sharing the components of these new forms of scholarly output and to improve the methods of preserving diverse formats. As the complexity of a scholarly works grows, however, it becomes unlikely that all of the components will become available at the same time, be accessible through a single repository, or even stay in the same state as they were at the time of publication. In turn, it also becomes more challenging to maintain a comprehensive and current perspective on what the complete work consists of and where all of the components can be found. It is this challenge that makes it valuable to also capture and preserve the map of relationships amongst these distributed resources. The goal of the RMap project was to build a prototype service that can capture and preserve the maps of relationships found amongst these distributed works. The outcomes of the RMap project and its possible applications for preservation are described.

Keywords

Publishing workflows; linked data; data publishing; semantic web; RESTful API; digital preservation; scholarly communication; digital scholarship.

1. BACKGROUND

In recent years, the content that comprises the scholarly record has shifted from being primarily discrete text-based bounded objects, such as journals or books, to more dynamic and less "bounded" content that might include data, webpages, software, and more. In other words, the boundaries of the scholarly record are stretching beyond the traditional publication of outcomes to instead encompass additional outputs created during the process and aftermath of the work [10]. This means a scholarly work can be complex, dynamic, and consist of multiple distributed parts. An example of a typical map of the heterogeneous resources that comprise and describe a single work is shown in Figure 1.

These changes in scholarly communication have been facilitated by technological shifts that have diversified the kinds of content that can be produced during research and made it easier to share digital material. One consequence of this has been a movement towards more funders and publishers requesting that researchers maintain and/or share research outputs to support reuse, validation, and replication of their methods and results. In the US, for example, the Office of Science and Technology Policy's 2013 memorandum [8] highlighted the government's commitment to improving availability of data resulting from federally funded research. An example in publishing is Nature Publishing Group's policies requiring that authors make materials, data, code, and protocols available to readers on request[1].

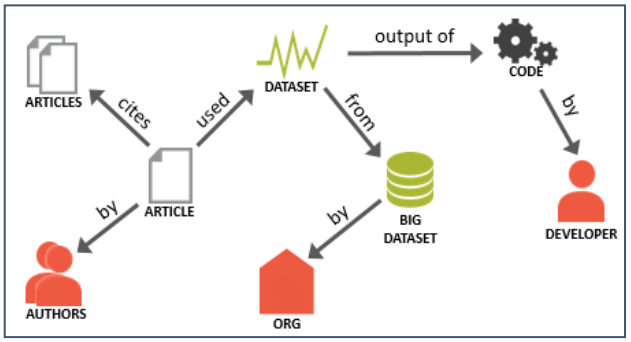

Figure 1 Multi-part Distributed Scholarly Work

Another consequence is the changes to publication workflows to support citing these new kinds of materials. For example, a lot of work has been done to support citing datasets in articles as first-class objects. Other initiatives have expanded this effort to include software citation [1] and citation of other kinds of resources, such as antibodies or model organisms [2]. Guidelines on data citation have been implemented by some publishers, though they are not yet consistently applied. One study shows only 6% of Dryad datasets associated with a journal article appear in the citation list for that article [11].

While this expansion of categories of citation is useful, there are many shortcomings attendant on attempting to shoehorn the rich network model of scholarly artifacts, contexts, and relationships into the structure of a journal article citation. First is the challenge inherent in the asynchronous nature of publishing the various components of a distributed work. Once an article is published in the traditional manner, the opportunity to connect or update related works has often passed, or at least become more difficult, with incentives for the author to update the connections greatly reduced. In an ideal scenario, all supporting material and outputs would be published and assigned identifiers before an article is published, but in reality this can be difficult to orchestrate and happens rarely. This means additional context shared post publication cannot typically be referenced from the published article. Second, the OCLC report on *The Evolving Scholarly Record* [10] describes how even after research outcomes are published, useful context and commentary are added to the work through presentations, blogs, and more in the "aftermath." These responses may never be published in an article with a DOI, but can provide important

[1] http://www.nature.com/authors/policies/availability.html

context to the work. A third challenge is that, while some publishers accept many kinds of works in their citation list, others are more restrictive. There are some works that cannot be published in a repository or easily assigned an identifier because their dynamic nature, scale, or copyright. If citations lists are limited to items with certain kinds of identifiers, for example, some components may not be included. Fourth, publisher or editorial boards often limit the total number, not just the type, of citations. Furthermore, there are sometimes simply too many objects for traditional citation to be practical. Finally, a linear citation list may not allow the researcher to clearly capture the role a resource played in the work or the nature of various contributions to the project.

All of these challenges suggest that there could be value in a service that can capture and preserve these evolving maps of relationships among the resources that form the scholarly work. One of the important tenants of the RMap Project is that this map itself can be considered a first class artifact of scholarly communication. For an increasing number of works, the published article is the tip of the iceberg. The definition of what encompasses a scholarly work has become much more complex than it once was. Understanding how the parts of a work relate to each other is important context for being able to preserve scholarship in a way that will allow it to be reused, replicated and validated.

2. THE RMAP PROJECT

The RMap[2] project was funded by the Alfred P. Sloan Foundation[3] and carried out by the Data Conservancy[4], Portico[5], and IEEE[6], starting in 2014. The goal of the project was to create a prototype API that could capture and preserve the maps of relationships amongst scholarly works.

The RMap team's work was developed in the context of a growing consensus that there is a need to capture the relationships amongst the components of complex scholarly works. The OCLC report on *The Evolving Scholarly Record* [10] identified the need for the expression of a set of relationships to bind together the pieces of a scholarly work. The Research Object collaboration has produced a set of tools and specifications for bundling together and describing essential information relating to experiments and investigations [3]. The RDA/WDS Publishing Data Services Working Group, in which the RMap team has participated, recently published recommendations for implementing a data to publication cross-linking service [5]. The working group also implemented a pilot aggregation and query service[7] and continue to develop the framework under the name *Scholix*[8]. More recently DataCite announced its Event Data service[9], which will support the registration and exchange of references between resources.

Some of these services focus on bilateral connections between objects, often with a circumscribed set of defined relationships between objects, and with allowable persistent identifiers for resources. RMap's focus is on the complete graph of resources that represent a compound work, with support for all identifiers and relationships that can be expressed as valid linked data. Through these graphs, bilateral relationships can also be identified. Over the last 2 years the RMap project has developed an API service that can act as a hub for capturing and preserving these maps.

RMap captures the resource maps as linked data[10] graphs, building on the features of the semantic web [4] and adopting the concept of an Aggregation from the Open Archives Initiative Object Reuse and Exchange[11] (OAI-ORE) standard. To support easy integration into existing data workflows, RMap employs a RESTful (Representational State Transfer) API [6]. Where available, RMap makes use of existing broadly adopted vocabularies (e.g. Dublin Core[12], Friend of a Friend[13], Open Provenance Model[14]) in its data model.

2.1 Objectives

As we have noted, RMap aims to capture and preserve links amongst the artifacts of scholarly communication and those who create, modify, employ, and annotate them [7]. 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 assertions recorded 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

The RMap data model utilizes the Resource Description Framework (RDF)[15] concepts of resources, triples, and graphs. The model includes three kinds of named graphs: *DiSCOs*, *Agents*, and *Events*.

2.2.1 RMap DiSCOs

RMap DiSCOs (Distributed Scholarly Compound Objects) are named graphs containing:

- A unique persistent identifier
- A list of 1 or more aggregated resource URIs (*ore:aggregates*) that form the aggregated work.
- An optional list of assertions about the aggregated resources. There are no constraints on the ontologies that can be used in these assertions, provided they form a connected graph with the aggregated resources at the root. These may be used to include additional context about each of the resources e.g. descriptive metadata, relationships to other resources, type, other identifiers, etc.
- An optional creator, description, and provenance URI to provide more information about the source of the DiSCO.

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

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

[4] http://dataconservancy.org/

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

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

[7] http://dliservice.research-infrastructures.eu/

[8] http://www.scholix.org/

[9] https://eventdata.datacite.org/

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

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

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

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

[14] http://openprovenance.org/

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

DiSCOs contain the ore:aggregates predicate, but do not otherwise follow the OAI-ORE model. For example, while OAI-ORE logically separates the concept of an Aggregation from the document that describes it (the "Resource Map"), a DiSCO combines these two notions into a single resource in order to make it easier to contribute data. Instead, much of the data that would typically be part of an OAI-ORE Resource Map is generated automatically as part of the API service and stored as RMap Events. As a result, the simplest form of a DiSCO is very easy to construct. An example of this is shown in Figure 2, which simply asserts that two resources form a compound object but does not further define the relationship between them. Beyond this users can add as much detail to the DiSCO as they see fit. The RMap team chose to keep the model simple and requirements to a minimum, but have also investigated what would be required to make the system fully compatible with OAI-ORE. It is estimated that the OAI-ORE model could be supported with several small enhancements if there were demand for this in the future.

```
<ark:/00000/03j9uf983h8fh8s>
        a rmap:DiSCO ;
        ore:aggregates <http://urlfordataset.org/part1>,
                       <http://urlfordataset.org/part2> .
```

Figure 2 Simple DiSCO as Turtle RDF

DiSCOs are immutable in that their identifier always corresponds to a specific set of assertions. When a DiSCO is updated, the previous version still exists and the new version is assigned a new identifier.

DiSCOs can have one of four statuses. *Active* means the assertions in the DiSCO are still assumed to be true. *Inactive* means the DiSCO has either been retracted or updated with a new set of assertions. Inactive DiSCOs can still be accessed publicly. When a DiSCO is updated, the previous version is automatically set to Inactive, but is still available to view in the version chain. *Deleted* means the DiSCO is retracted and the assertions are not publicly visible through the API, even though the data exists in the database. A *Tombstoned* status means the DiSCO has been removed from the database, but the provenance information persists as a record of the removal.

2.2.2 RMap Agents

RMap *Agents* are named graphs representing a person, process, or thing that is responsible for some action on the RMap database. Anyone who contributes data to RMap is required to have an Agent. Each new Agent is assigned a persistent identifier that is associated with changes to the database. Unlike the DiSCO model, Agents are mutable, so updates to the Agent graph will overwrite the previous version. Changes to the Agent graph are recorded as *Events*.

2.2.3 RMap Events

An RMap *Event* is automatically generated whenever a user makes any additions or changes to RMap. They are used to record and track the provenance and status of RMap DiSCOs and Agents. Each Event has a unique persistent identifier, and includes the URI of the RMap Agent that made the change, the type of change, URIs of any RMap objects affected, the timeframe of the Event, and optionally the specific API access key that was used to make the change. Events cannot be updated or deleted.

2.3 RESTful API

The primary interface for accessing the RMap database is a RESTful 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.

The RMap RESTful API includes over 30 functions for querying and generating data. For example, you can retrieve a list of triples that mention a specific resource, or a list of DiSCOs created by a specific Agent. Functions that generate lists of results can typically be filtered by date, creating Agent, and DiSCO status.

2.4 Web Application and Visualization Tool

In addition to the RESTful API, data can be navigated interactively through the RMap web application. This allows the user to look up DiSCO URIs and view either a tabular representation or a graph visualization (Figure 3) of the data. By clicking on resources in the visualization or data table, it is possible to drill into the data and view all triples and DiSCOs that reference that resource.

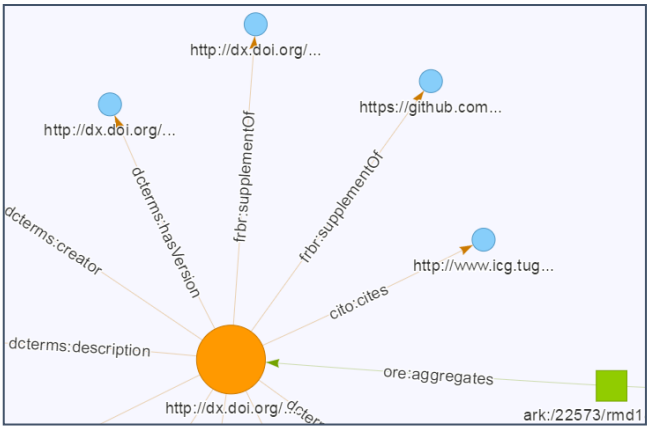

Figure 3 Part of RMap DiSCO visualization

2.5 Outcomes

Over the last two years, the RMap team has produced a working prototype RESTful API for managing and retrieving data in RMap. They have also built a web application for navigating the RMap data interactively. By logging into the web application using Twitter, ORCID, or Google authentication, users can generate keys for the RESTful API. Links to the tools and documentation can be found on the RMap website[16]. Also available is a versatile harvesting framework to support large scale harvesting and ingest of DiSCOs. The team has also explored options for an inferencing engine to support the mapping of equivalent identifiers.

Example DiSCOs were created using metadata from DataCite[17], NCBI's PubMed and Nuccore APIs[18], ACM's Transactions of Mathematical Software[19], Portico, and the complete collection of IEEE articles. In one example metadata relating to a single article was imported from IEEE, Portico, and DataCite in order to demonstrate how to navigate between different components of a work through overlapping DiSCOs. The RMap database continues to grow. At the time of writing the RMap prototype service contains over 4.5 million DiSCOs, comprised of over 230 million triples.

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

[17] https://www.datacite.org/

[18] http://www.ncbi.nlm.nih.gov/home/api.shtml

[19] http://toms.acm.org/

A short extension to the project is supporting the exploration of representing SHARE[20] and Open Science Framework[21] data as DiSCOs in RMap.

3. PRESERVATION IMPLICATIONS

The goal of the RMap project was to develop a framework for capturing and *preserving* maps of relationships. Since RMap DiSCOs can be exported as RDF text format, exporting and preserving RMap DiSCOs can follow a typical preservation pathway for plain text. As the project has unfolded, however, some other potential preservation use cases have been identified.

While the pathways to preservation of articles produced by publishers are well understood, the other components of the scholarly works described previously are typically not preserved in the same repository. Even if all of the components of the work are available in other repositories, it is unlikely that the map of the connections between all of the parts will be available in a form that is accessible to all repositories. This means none of the components show a full picture and the complete work is difficult to assemble. Using RMap as a hub to represent these connections between the distributed components of the works, could help ensure all components of the work can be found and preserved.

Where metadata and components are distributed across different kinds of platforms, it is possible that one or more of the resources will eventually be lost or altered. Even if all resources are preserved, it is highly likely that one of the resources will reference a URL that has moved or no longer exists and will produce a 404 "not found" error when accessed. One study showed that the problem of *reference rot* already affects one in five articles [9]. Add to that equation a variety of non-article resources that are not necessarily peer reviewed or conforming to any fixed publication path, and the problem of reference rot may be even more problematic. Even if there is a new equivalent link available, there is often no easy way for anyone to indicate a new location. Not only does RMap provide an opportunity for links to be updated and identifiers added, one useful enhancement to the framework might be to interface with the Internet Archive's Wayback Machine[22] APIs to associate Memento links with web URLs that do not use a persistent URI.

Finally, during the first phase of the project, the RMap team generated some DiSCOs using Portico content. Each DiSCO showed which resources were preserved by Portico for a single article. Combining similar data from other repositories could be useful for identifying preservation gaps and overlap for different kinds of work.

4. CONCLUSIONS

The RMap project has produced a framework for generating maps of the components of a distributed scholarly work. 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. Some applications of RMap could also support improved preservation of distributed scholarly compound works.

5. 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, Vinay Cheruku, Amy Kirchhoff, John Meyer, Stephanie Orphan, Joseph Rogowski; and IEEE colleagues Renny Guida, Ken Rawson, Ken Moore.

[20] http://www.share-research.org/

[21] https://osf.io/

[22] http://archive.org/web

6. REFERENCES

[1] Ahalt, S., Carsey, T., Couch, A. et al. 2015. NSF Workshop on Supporting Scientific Discovery through Norms and Practices for Software and Data Citation and Attribution. Retrieved April 2016 from https://softwaredatacitation.org/Workshop%20Report

[2] Bandrowski, A., Brush, M., Grethe, J. S. et al. 2015. The Resource Identification Initiative: A cultural shift in publishing [version 2; referees: 2 approved]. *F1000Research* 4, 134. DOI= http://doi.org/10.12688/f1000research.6555.2.

[3] Bechhofer, S., Ainsworth J., Bhagat, J. et al. 2013. Why Linked Data is Not Enough for Scientists. *Future Generation Computer Systems* 29, 2 (February 2013), 599-611. DOI= http://doi.org/10.1016/j.future.2011.08.004

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

[5] Burton A., and Koers, H. 2016. *Interoperability Framework Recommendations.* ICSU-WDS & RDA. Publishing Data Services Working Group. Retrieved 29 June 2016 from http://www.scholix.org/guidelines

[6] 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

[7] Hanson, K. L., DiLauro, T. and Donoghue, M., 2015. The RMap Project: Capturing and Preserving Associations amongst Multi-Part Distributed Publications. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '15). ACM, New York, NY, USA, 281-282. DOI= http://dx.doi.org/10.1145/2756406.2756952

[8] Holdren, J.P., 2013. *Increasing access to the results of federally funded scientific research. Memorandum for the heads of executive departments and agencies.* Office of Science and Technology Policy, Executive Office of the President, Washington, DC. Retrieved 20 April 2016 from https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[9] Klein, M., Van de Sompel, H., Sanderson, R., et al. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE,* 9, 12. e115253. DOI= http://doi.org/10.1371/journal.pone.0115253

[10] Lavoie, B., Childress, E., Erway, R., Faniel, I., Malpas, C., Schaffner, J. and van der Werf, Titia. 2014. *The Evolving Scholarly Record*. OCLC Research, Dublin, Ohio. Retrieved 20 April 2016 from http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-evolving-scholarly-record-2014.pdf

[11] Mayo, C., Hull, E.A. and Vision, T.J. 2015. The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository. *Zenodo*. DOI= http://dx.doi.org/10.5281/zenodo.32412