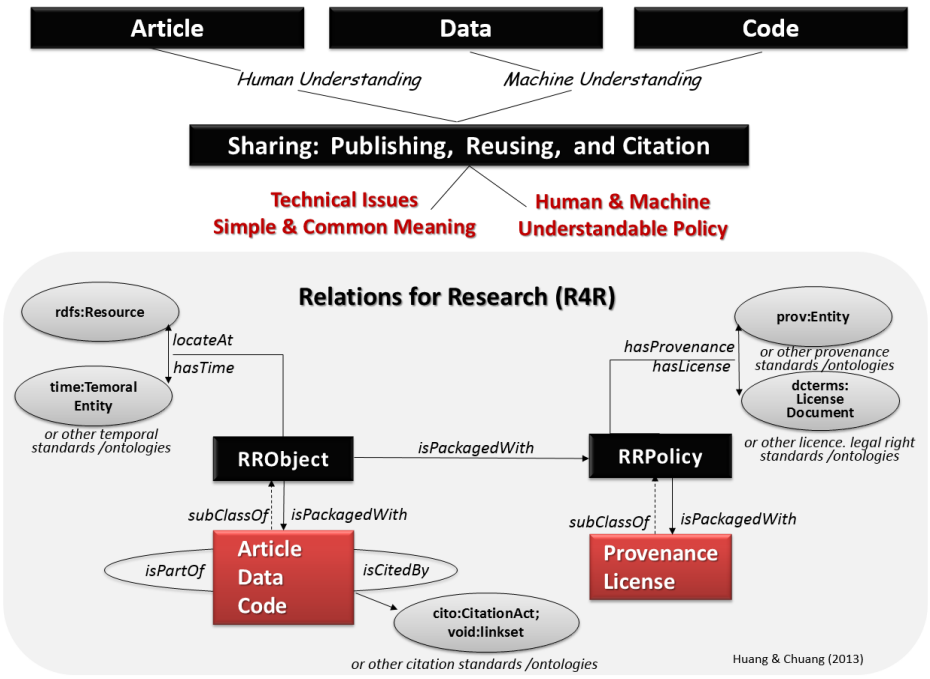
Relations for Research (R4R): A Conceptual Model for Publishing Research Articles, Data, and Code

Andrea Wei-Ching Huang
Institute of Information Science
Academia Sinica, Taiwan

Tyng–Ruey Chuang
Institute of Information Science
Academia Sinica, Taiwan

Abstract

This paper discusses the reasoning that it is the semantic relation between research articles, data and code that can support the current global demand for an Open Science. We argue that a conceptual model for relating the publishing of three kinds of research components: Articles, Data, and Code. Between these components they form various relations (e.g. article-article, article-data, article-code, data-data, data-code, code-code, and article-data-code), and they may be covered by machine-readable policies (i.e. provenance and license). Accordingly, we bring the Linked Open Data (LOD) approach into Scientific Data Repositories (SDRs) which currently have not managed and curated datasets in a semantically rich manner. We propose a conceptual model – Relations for Research (R4R) – as a guidance framework in such an approach.



Citation Information: Andrea Wei-Ching Huang and Tyng–Ruey Chuang (2013) Relations for Research (R4R): A Proposed Conceptual Model for Publishing Research Articles, Data, and Code, URL: http://andrea-index.blogspot.com/2013/12/R4R.html

1. Introduction

The year 2013 will come to an end soon. However, new progresses in scientific sharing and publishing are just beginning. The global demands on public access to research data have been endorsed by many government policies. The movement toward Open Science has also been welcomed by several key scientific publishing actors.

To name just a few, Nature Publishing Group has just announced the launch of an online data journal, *Scientific Data*, for the open access to detailed data descriptions. The data journal, *Earth System Science Data (ESSD)*, adopts a new form of the reviewing policy, which allows scientists and general public to review and comment articles. Later, these interactive comments plus author's responses and revisions are published and archived openly in fully citable and paginated forms. Web services like figshare, f1000research, or Research Compendia provide scientists new tools and alternative platforms to curate their research outputs.

High-level requirements of science reproducibility result in the coming of a new science publishing paradigm. This paradigm requires the packaging of articles, data and code, and encourages their joint publications. The initial task has been taken by some bio-medical science practices, and until recently, the *Executable Papers* of ScienceDirect in computer science implemented this vision online.

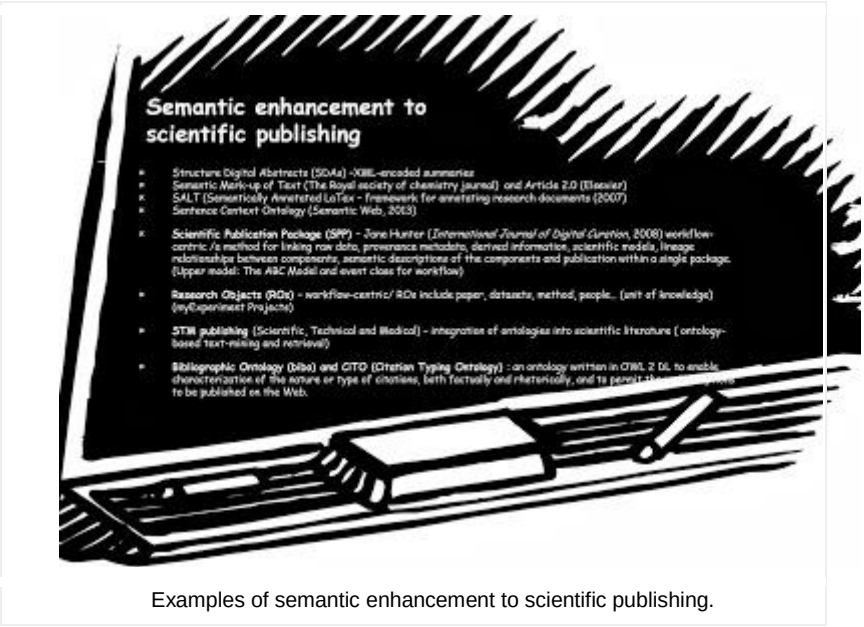
Thus rethinking the dilemma we face today is both for new possibilities and problems carried by "big data" and "open data". While we embrace the coming of big data and open research in a data-driven context, we have to tackle the data deluge problems caused by data generation, data sharing, and data publishing. Problems like data heterogeneity, interoperability, accessibility, citability, reproducibility as well as legal issues remain major challenges to the research communities.

Despite huge varieties existing in different domains, the difficulty falls into two main categories: technical issues and policy instruments. What we need are an intelligent openness strategy as outlined in Geoffrey Boulton's proposal that we present the scientific argument (the data and concept) together, as well as an integrated infrastructure for this new research paradigm.

2. The Reasoning

- **Why do we need to explore semantics and semantic relations when we publish research data?**

The emerging collaboration on scientific publishing, between Scientific Data Repository (SDR) community, Library Metadata community, as well as Linked Open Data (LOD) and Semantic Web communities, suggests that the semantic discourse has an important implication on research publishing. Among these, library communities have played a significant role in managing research data due to their expertise in managing metadata and data curation.



Examples of semantic enhancement to scientific publishing.

Yet, current state of metadata standards in the scientific context is not sufficient for data integration and reuse. More efforts need to be taken on scientific publishing and citation. Also, it is difficult to agree on a single metadata schema or standard in the open Web context. Challenges still remain in the technical complexity of mapping to achieve metadata interoperability.

In addition, linking has been a major feature of how scientific datasets can be managed. Among which, the problem of lacking dataset identity is the major obstacle to citation and metadata developments. In particular, metadata schemas for scientific data modelling can sometimes be too general or too specific in describing relations of multiple domains.

Accordingly, we argue that the LOD approach provides a possible solution to the above problems. The LOD approach provides unique URIs for object identity;

1. Users are free to set different URIs to refer to an object (e.g. using links like *owl:sameAs*).
2. For the accessibility, identity links such as the URI links (in subject and object) help machines to find more data; the property links (in predicate) provide meaning and context for data to be linked.
3. In general, RDF (Resource Framework Description) links help to decrease the interoperability problem by *pointing data to the vocabulary they use,* and to *the definitions of related terms* in other vocabularies.

Thus the LOD approach assists scientific datasets to be accessible, to be related to other data sources, and to be linked between different datasets semantically. (see more details from Bizer, Cyganiak & Heath, 2007; Seneviratne, Kagal & Berners-Lee, 2009)

- **What kinds of policy mechanisms support the open research?**

Data reusing and remixing are part of the charm of open science. Yet, the data quality and usability are not easy to understand as long as provenance and licence information are not clear enough both for human and for machines.

While scientific data repositories need provenance metadata as the data preservation policy, the computational traceability has been required for the purpose of quality control and data reuse. Additionally, endless pages of copy right statements or license agreements are as complex as they can be. Thus people may have been violating others' rights without awareness of it. In such cases, if provenance and license information are machine-readable and further packaged with scientific datasets when they are travelling, policy-aware tools like Semantic Clipboard can thus help to detect license violations when exposing Creative Commons (CC) license metadata as RDF or RDFa.

(source: Semantic Clipboard http://dig.csail.mit.edu/2009/Clipboard/)

Accordingly, a portable and packaged metadata policy is the key to the open research. For a data collection to be open, they shall be freely downloaded, adapted, mixed with others, and re-hosted for other services. Being available and accessible on the Web by itself is not sufficient. A data collection must be easily ported to other computer systems, either on or on the Web, for it to be called open. The portable principle is also applied in modelling metadata for scientific data.

In other words, a portable and packaged research component should contain articles, data, code, as well as associated provenance and licence metadata as a completed knowledge package for research results reusing and remixing.

3. The Relations for Research (R4R) Conceptual Model

Identity functions for scientific publications require the dataset to be constructed as a semantically and logically concrete object. Thus we define two core classes: ***Research Related Object (RRObject)*** and ***Research Related Policy (RRPolicy)***. Three objects, Article, Data, and Code, are classified as subclasses of RRObject. Two classes, Provenance and License are subclass to RRPolicy. For object properties, we identify seven relations in between RRObject objects and RRPolicy objects. Here we only present a summary table below for the R4R conceptual model.

Table 1. A Summary of R4R.

Class	Property	Domain	Range
RRObject	locateAt	RRObject	rdfs:Resource (URL/URI/DOI/ISBN…)
Article	hasTime	RRObject	time:TemporalEntity
Data	isPartOf	RRObject	RRObject; void:Dataset
Code	isCitedBy / cite	RRObject	RRObject; cito:CitationAct; void:linkset
RRPolicy	isPackagedWith	RRObject	Article; Data; Code; RRPolicy
Provenance	hasProvenance	RRPolicy	prov:Entity
License	hasLicense	RRPolicy	dcterms:LicenseDocument

Our rationale for the design of the two core classes has been flexible in the definition of scientific publishing. RRObject is not necessary only for the publication purpose. RRObject can use the property of "*isPackagedWith*" to combine all related objects whether been published or not.

Instead of using direct statements about sharing research and publishing rights, we can use the *License,* subclass of the RRPolicy, to refer to well-known licenses. For instance, licenses include. the Creative Commons (CC) licenses for creative works; Open Data Commons Open Database License (ODbL) for databases and datasets, or your own Open Data Certificates; as well as the GNU General Public License (GNU GPL or GPL) for software source code. These licenses and certificates can be bundled and packaged with RRObject through the property of "*isPackagedWith*".

The ability to identify the relationship between articles, data and code is essential to a full understanding of the R4R design. To help explain the conceptual model, seven correlations are discussed.

1. **Article-Article:** This is the most conventional relation that has been used in scientific publishing and citation through the bibliography. We use "*isCitedBy*" to provide a general relation between article and article, and refer to CiTO ontology for further semantics of various relation types.
2. **Article-Data:** Data or datasets collected, created, and derived for a research itself "*isPackagedWith*" the Article. Article can also "isPackagedWith" Data. Furthermore, according to CiTO, Article can *cite* Data, and Data can cite Article.
3. **Article-Code:** The relation of Article and Code share the same logic with Article and Data for "*isPackagedWith*" and "*isCitedBy*".
4. **Data-Data:** Data can be "*isPartOf*" other Data based on the granularity and scalability of the dataset. However, the relation of "isPartOf" is transitive and reflexive. Data can also "*cite*" Data.
5. **Data-Code:** Code can be "*isPartOf*" Data; Data can be "isPartOf" Code. However, according to CiTO, Article can cite Data, and Data can cite Article. Although "isPartOf" and "isCitedBy" share some similar semantics, we use "isPartOf" for being capable of representing transitive relation.
6. **Code-Code:** The relation of Code and Code share the same logic with Data and Data for "*isPartOf*" and "*isCitedBy*".
7. **Article-Data-Code:** The relations between the three kinds of RRObject can raise some interesting questions. For example, when CodeA "*isPartOf*" DataB, and CodeA "*isCitedBy*" ArticleC, what relation between DataB and ArticleC can we say about? Is there a concise term we can use to express such a relationship?

The relations described above for Article, Data and Code are such some of the initial steps prepared for some clarifications of research components. Note that, the proposed R4R conceptual model here is not yet a formal ontology, as we have not finalized the detailed vocabulary to be used. In the meantime, we also need to elaborate the model with some use cases.

In addition, data citation standards and practices are still in progress. Most models and tools of provenance for web databases and scientific workflow are still in the experimental level. Definition and relation of R4R may need to be refined in the near future. However, we here describe R4R in the form of a conceptual model so as to enter into discussion with research communities. We expect to develop and extend R4R later into a formal ontology.

4. Related Works and Discussion

Scientific Publication Packages (SPP) is similar to our view in packaging textual publications, raw data, derived products, algorithms, and software altogether. The major differences are two: (1) SPP has taken those research components in one concept, namely Data, based on the extension of the ABC model, a model for the library, museum and archival domains. (2) SPP does not consider license packaging. In contrast, the notion of Research Objects (ROs) as first class citizens for sharing and publishing is similar to our R4R design.

However, both ROs and SPP are workflow and life-cycle centric. While SPP emphasises data preservation and publishing, ROs focuses more on aggregation. Both ROs and R4R notice the necessity of packaging and licensing issues, but only R4R includes licensing in the core model. ROs packages the workflow with data, results and provenance; while R4R packages articles, data, code, provenance, license, and their semantic links. One specific difference distinguished R4R from the other two is that R4R stresses the importance of relations between research components for research publications and citations.

In addition, existing vocabularies and ontology which are similar to R4R are the MESUR ontology and the Semantic Publishing and Referencing Ontologies (SPAR). MESUR and SPAR share the same purpose with R4R in that they describe the scholarly publishing tasks, but MESUR concerns more on scalability and the bibliometric usage, while SPAR provides core vocabularies and semantics for publishing and referencing in eight ontology modules.

A further consideration is that it would be interesting to explore more on relationships between provenance and event, since provenance information provides relations of research components changed over time (e.g. editing history). In the editorial note of the Preservation Metadata: Implementation Strategies (PREMIS) Ontology, it describes that "*Digital provenance often requires that relationships between objects and events are documented*".

SPP, ROs and MESUR all apply event concepts for describing life-cycle of objects. Three out of eight ontology modules in SPAR have event concepts (for example, by taking citation as an event, or taking care of event occurring in the publishing process and workflow). As R4R is still in its preliminary stage of design, we may consider to include event concepts in the future (note that our event based concept is more toward the "relation" property which focuses more on spatio-temporal relationships and cause-effect relationships explained in *here*). As for an overall view of mapping between different provenance vocabularies, this can be seen from the task of W3C Provenance Incubator Group.

5. Conclusion

As discussed at the beginning, research challenges on technical complexity and policy instrument for an Open Research are still not well understood nor agreed on in many aspects. We hope this discussion and reasoning on why a LOD approach can serve as a semantic solution to various problems. And R4R can provide a conceptual framework to relating major components and relations to scientific publication and citation. To conclude, we summarise them in the following:

1. A research publication has to be packaged with both RRobject (article, data, and code) and RRPolicy (provenance and license) in a portable and policy-aware manner.
2. A research model that highlights relations between various research components has been proposed. We propose a Linked Open Data approach into scientific publishing process.

3. We argue that as long as we take the challenge of technical complexity and policy instrument together, as long as research publications stay open, the use of shared and linked semantics will help scientific research continue to grow in a new way.

So at the coming of the New Year 2014, we are optimistic about Open Science, as well as the semantic relation between research articles, data and code. We see a lot of efforts on shared semantics which hold machine understandable promises that data can be represented and reasoned semi-auto or automatically. We see a strong potential in the scientific research, the same scientific referents can be reused, reanalysed and remixed effectively from distributed agents (human and machine) of various domains. In other words, finding new use of known scientific facts, linking new relationship of established research, or generating new science are under the way.

6. References

1. Altman M, Arnaud E, Borgman C, Callaghan S, Brase J, Carpenter T, Chavan V, Cohen D, Hahnel M, & Helly J. Out of Cite, Out of Mind: The Current State of Practice, Policy and Technology for Data Citation. *Data Science Journal* [Internet]. 2013;12:1–75.
2. Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., ... & Goble, C. (2011). Why linked data is not enough for scientists. *Future Generation Computer Systems.*
3. Bizer, C., Cyganiak, R., & Heath, T. (2007). How to publish linked data on the web. Retrieved October, 20, 2013
4. Chuang, T.R. (2013) Packaging and Distributing Data Collections for the Web, *Open Data on the Web*, 23 - 24 April 2013, London, Retrieved from http://www.w3.org/2013/04/odw/odw13_submission_44.pdf
5. Cox, A. M., & Pinfield, S. (2013). Research data management and libraries: Current activities and future priorities. *Journal of Librarianship and Information Science*.
6. Haslhofer, B., & Klas, W. (2010). A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys (CSUR),* 42(2), 7
7. Hunter, J. (2008). Scientific Publication Packages–A selective approach to the communication and archival of scientific output. *International Journal of Digital Curation*, 1(1), 33-52
8. Marcial, L. H., & Hemminger, B. M. (2010). Scientific data repositories on the Web: An initial survey. *Journal of the American Society for Information Science and Technology*, 61(10), 2029-2048
9. Qin, J., Ball, A., & Greenberg, J. (2012). Functional and Architectural Requirements for Metadata: Supporting Discovery and Management of Scientific Data. In *Twelfth International Conference on Dublin Core and Metadata Applications*. University of Bath.
10. Rodriguez, M. A., Bollen, J., & Van de Sompel, H. (2007, June). A practical ontology for the large-scale modeling of scholarly artifacts and their usage. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (pp. 278-287). ACM.
11. Peng, Roger D. "Reproducible research in computational science." *Science* (New York, Ny) 334.6060 (2011): 1226-1227
12. Seneviratne, O., Kagal, L., & Berners-Lee, T. (2009). Policy-Aware content reuse on the web. In *The Semantic Web-ISWC* 2009 (pp. 553-568). Springer Berlin Heidelberg.
13. Shotton, D., Portwin, K., Klyne, G., & Miles, A. (2009). Adventures in semantic publishing: exemplar semantic enhancements of a research article*. PLoS computational biology*, 5(4), e1000361.
14. Wynholds, L. (2011). Linking to scientific data: Identity problems of unruly and poorly bounded digital objects. *International Journal of Digital Curation*, 6(1), 214-225
15. Yarmey, L., & Baker, K. S. (2013). Towards Standardization: A Participatory Framework for Scientific Standard-Making. *International Journal of Digital Curation*, 8(1), 157-172.