Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 19 Issue 1 Version 1.0 Year 2019
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Enabling Researchers to Make their Data Count

By Ajit Singh

Department of Computer Science Patna Women's College

Abstract- Over the last years, many organizations have been working on infrastructure to facilitate sharing and reuse of research data. This means that researchers now have ways of making their data available, but not necessarily incentives to do so. Several Research Data Alliance (RDA) working groups have been working on ways to start measuring activities around research data to provide input for new Data Level Metrics (DLMs). These DLMs are a critical step towards providing researchers with credit for their work. In this paper, I describe the outcomes of the work of the Scholarly Link Exchange (Scholix) working group and the Data Usage Metrics working group. The Scholix working group developed a framework that allows organizations to expose and discover links between articles and datasets, thereby providing an indication of data citations. The Data Usage Metrics group works on a standard for the measurement and display of Data Usage Metrics. Here I explain how publishers and data repositories can contribute to and benefit from these initiatives. Together, these contributions feed into several hubs that enable data repositories to start displaying DLMs. Once these DLMs are available, researchers are in a better position to make their data count and be rewarded for their work.

Keywords: *crossref; research data count; citation; DLM; RDA; scholix; researcher; datasite; DOI, working group.*

GJCST-C Classification: *E.m*

Strictly as per the compliance and regulations of:

Enabling Researchers to Make their Data Count

Ajit Singh

Abstract- Over the last years, many organizations have been working on infrastructure to facilitate sharing and reuse of research data. This means that researchers now have ways of making their data available, but not necessarily incentives to do so. Several Research Data Alliance (RDA) working groups have been working on ways to start measuring activities around research data to provide input for new Data Level Metrics (DLMs). These DLMs are a critical step towards providing researchers with credit for their work. In this paper, I describe the outcomes of the work of the Scholarly Link Exchange (Scholix) working group and the Data Usage Metrics working group. The Scholix working group developed a framework that allows organizations to expose and discover links between articles and datasets, thereby providing an indication of data citations. The Data Usage Metrics group works on a standard for the measurement and display of Data Usage Metrics. Here I explain how publishers and data repositories can contribute to and benefit from these initiatives. Together, these contributions feed into several hubs that enable data repositories to start displaying DLMs. Once these DLMs are available, researchers are in a better position to make their data count and be rewarded for their work.

Keywords: crossref; research data count; citation; DLM; RDA; scholix; researcher; datasite; DOI, working group.

I. Introduction

Researchers who want to build on published research can reuse existing data to arrive at new conclusions. In addition, linking scholarly literature and data leads to increased visibility, discovery and retrieval of both literature and data, facilitating reuse, reproducibility and transparency. In a digital world where data can be more easily shared and documented, scholarly literature and its underpinning data are increasingly seen as inseparable.

At the same time, while the importance of data sharing is accepted, there are essential questions that still require an answer. For example, why should authors go through the effort of documenting and publishing datasets, if their career depends on the publication of articles and if there is no standard for metadata and basic attribution information around data? Several RDA projects are underway to provide answers to these questions by creating a framework to measure data reuse in a standardized fashion.

Finding the right way to measure the impact of shared data is crucial if research data is to be included as one of the scholarly outputs used for research evaluation. The current meritocratic system in academia relies heavily on the publication of scientific results in recognized academic journals, supported by an internantional editorial board and peer review system. The most commonly used metric to measure the impact of a publication is counting the number of times it receives a citation from other publications that are also peer reviewed and published in recognized journals.

The temptation to use the same metrics for data, and measure citations of datasets in articles, is certainly strong. However, the interaction and impact of research data is more complex than that. The very definition of what a citation for data is fuzzier than the equivalent for articles.

In this paper, I describe how the outputs of two RDA working groups (WGs), the Scholix WG and the Data Usage Metrics WG, can be used to assess data reuse and make data usage statistics and citations available. I will first outline how data repositories and publishers can expose article-data links using Scholix approaches and data usage metrics following the new code of practice for research data. I will then explain how they can consume this information to make DLMs available and help researchers get credit for their work.

II. Data Citation

a) *Scholix: aggregating article-data links to count data citations*

The goal of the Scholix WG was to establish a high-level framework for exchanging article-data links. It aimed to enable an open information ecosystem to understand systematically what data underpins literature and what literature references data.

The Scholix WG addressed this problem. Its goal was to improve the links between scholarly literature and research data as well as between datasets, thereby making it easier to discover, interpret, and reuse scholarly information. The Scholix initiative offers:

1. A universal, global framework that enables information about the links to be exchanged technical guidelines that specify how the interoperability framework works.
2. A common conceptual model, an information model, and open exchange protocols.
3. A community that discusses, develops and applies these specifications.

Within the Scholix framework:

Data repositories, journals, and others provide information about the links between literature and data that they hold to community 'hubs' such as OpenAIRE,

Author: Assistant Professor, Department of Computer Science Patna Women's College, Patna-800001 Bihar, India.
e-mail: ajit_singh24@yahoo.com

Crossref and DataCite (with Crossref and DataCite working on a shared infrastructure). This supports and respects existing community-specific practices and the existing means of exchanging this information.

The community 'hubs' – which are natural places to collect and exchange information about the links between literature and data – commit to a common information model for exchanging the links that they hold and an agreed open exchange method enables this to occur.

The conceptual model (Figure 1) is about the link between two objects, such as a journal article and the underpinning data. Rather than describing in detail the properties of each of the two objects, the conceptual model focuses on the relationship between the objects. It also enables a record of who asserted the link and who made the link available.

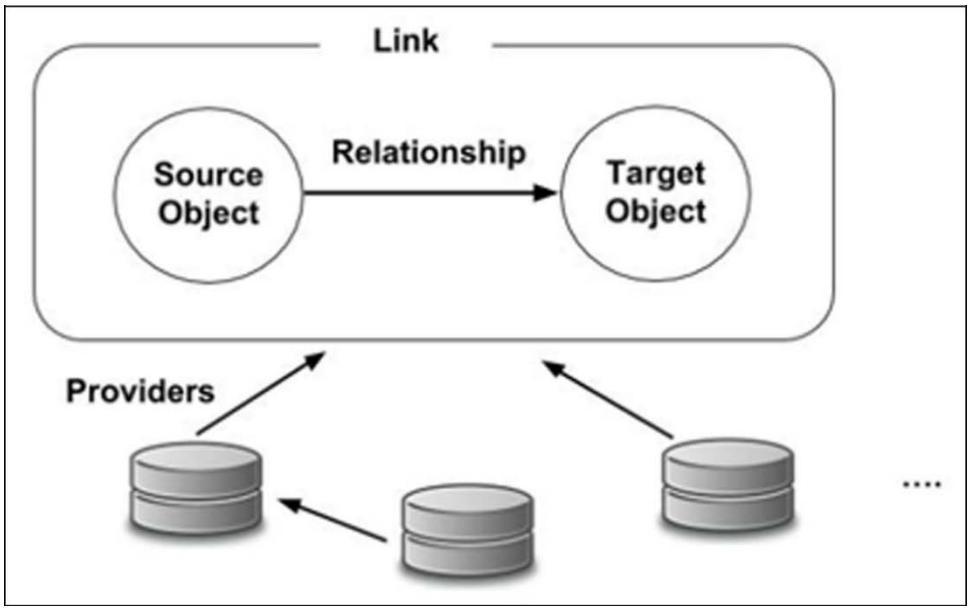

Figure 1: Scholix information model. Providers contribute links by sharing information about the source object (article or dataset), target object (article or dataset) and the nature and direction of the relationship.

b) Contributing data citations: publishers

As mentioned in the previous section, within the Scholix framework organizations contribute information through community hubs. The majority of scholarly publishers work with non-profit organization Crossref to share metadata about publications. These metadata records include comprehensive information about the items being registered, and increasingly include links to related scholarly artifacts such as data, software, protocols, and reviews.

As can be seen in Figure 2, Crossref provides two paths to registering data citations: references and relations. Relations are a way to associate related digital objects with each other through metadata. A publisher can register metadata with Crossref explicitly linking a dataset to a journal article. References are formal citations (such as would be provided in a bibliography) and are a type of relation but are provided separately within Crossref metadata.

Crossref members should deposit data citations as references if:

- The data citation includes a DataCite DOI
- They include data citations in their reference lists (recommended) Crossref members should deposit data citations as relations if:
- They want to capture specific relation types (e.g. is Supplemented By) beyond 'references'
- They are not able to supply data citations as references

In 2019 Crossref will be expanding citation support to allow publishers to explicitly identify data citations in line with the data citation roadmap for scientific publishers (Cousijn et al. 2018). This will allow for deposition of data citations with all types of persistent identifiers as references.

c) Contributing data citations: data repositories

Many data repositories actively curate and keep track of which articles are using the datasets they host. This is valuable information that is currently not always available to other organizations in the data community. For data repositories that use DataCite DOIs, the DOIs and accompanying metadata are registered with

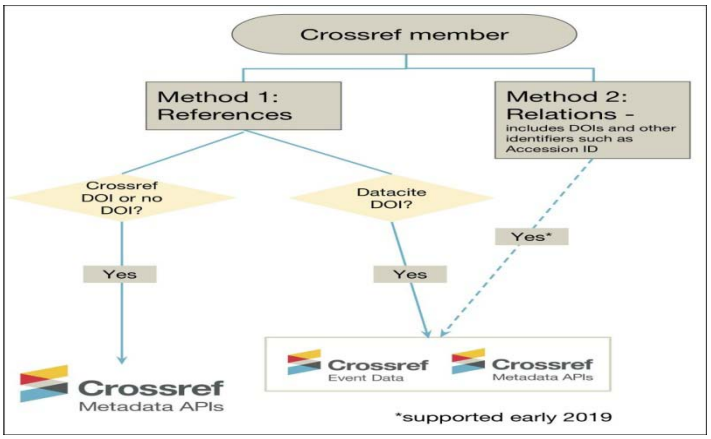

Figure 2: Depositing Data Citations with Crossref. Publishers can deposit data citations following two different methods: references or relations.

DataCite. Therefore, information about any journal publications related to a dataset can be included in the metadata records that are sent to DataCite. This additional information should follow the DataCite metadata schema which is aligned with the Scholix metadata schema (Burton et al. 2017b).

When these elements are added to the metadata that is registered with DataCite, the information about the links will automatically become openly available.

d) Contributing data citations: institutional repositories

For data centers that do not assign DataCite DOIs to datasets, OpenAIRE is currently the best place to deposit article-data links. Institutional repositories can export metadata descriptions of their datasets with links to articles as Dublin Core records or as Scholix records and register with OpenAIRE's Scholexplorer Service (Burton et al. 2017c) as a data source. Scholexplorer will bulk collect metadata records from the reposi-tory APIs; Scholexplorer is compatible with the OAI-PMH protocol or REST search APIs that allow collection of all records with a paging system (collecting by means of several calls) and with "last date of indexing" (incremental approach). Scholexplorer will then enrich its graph of article-dataset links with the ones collected from the repository, de-duplicate when necessary, and expose all links as Scholix records via APIs on behalf of the registered repository. All links exported by OpenAIRE carry provenance information about the data sources that provided the links (more than one source may have provided the same link), to ensure visi-bility of the contributing repositories and provide a degree of trust to the consuming services. OpenAIRE asks the database to display the Scholix logo on their website and indicate that it is harvested by Scholexplorer.

III. Data Usage Metrics

a) Standards for data usage metrics

Following the Scholix initiative and the related work of the RDA Data Citation WG, it was clear that there are broader metrics for data that the community needs to address. With the Scholix working group focusing on the relationships between articles and datasets and the Data Citation Working Group addressing challenges related to dynamic data citation, there was a need for a working group to define usage for data. The Data Usage Metrics WG started in Ocyober' 2018 and focuses on metrics that reflect usage of research data. The group is working to build a comprehensive list of use cases that covers the spectrum of types of 'usage metrics' that may apply to research data, build a recommendation for community guidance on what types of usage metrics should be applied at the data and repositories levels, and drive adoption of usage metrics across the research landscape. Specifically, the working group is aimed at outlining the barriers to adoption of data-level-metrics standards and current implementations of usage metrics across the data repository landscape. These conversations, surveys, and findings will aid in defining recommendations for types of data and associated metrics that repositories should be considering. The group works closely with the Make Data Count project and leverages the COUNTER code of practice for research data (mentioned below).

b) Contributing data usage metrics

This first release of the Code of Practice for Research Data specifically targets research data usage. The recommendations are aligned as much as possible with the COUNTER Code of Practice Release 5 for the major categories of e-resources (journals, databases, books, reference works, and multimedia databases).

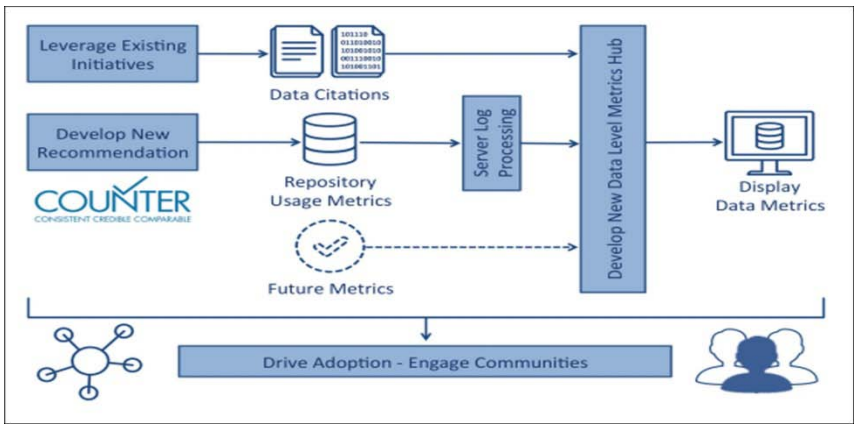

Figure 3: Framework of the Make Data Count project. Repositories process log files against the new Code of Practice and these processed files feed into the same hub as the article-data links collected following the cholix framework. All this information is made openly available to the community so organizations can develop and display DLMs.

and mainly concern views and downloads – called investigations and requests in the Code of Practice. Many definitions, processing rules and reporting recommendations apply to research data in the same way as they apply to other resources. The Code of Practice for Research Data enables the reporting of usage statistics by different data repositories following common best practices, and thus is an essential step towards realizing usage statistics as a metric available to the community to better understand how publicly available datasets are being reused.

IV. Consuming Data Usage Statistics and Citations

The citations and usage statistics contributed by data repositories and publishers are made openly available to the community via APIs. Crossref and DataCite developed Event Data, a shared underlying infrastructure that holds (among other things) all citations that are contributed as part of article and dataset metadata. Crossref and DataCite each have their own API through which they make these citations available.

- Services such as Scholexplorer retrieve data citations from the Crossref Event Data service using this Scholix API endpoint: http:/api.eventdata.crossref.org/v1/events/scholix.
- Scholexplorer combines this information with the citations that are provided to OpenAIRE.
- Views and downloads processed against the COUNTER Code of Practice are sent to DataCite and any repos-itory or research data service can consume usage statistics for a given dataset DOI from an Event Data Query API provided by DataCite (https://support.datacite.org/docs/eventdata-guide). The API combines citations and other events into one API call.

V. Conclusions

Measuring data (re)use and the development of DLMs are crucial if data is to become a first-class research output. Both the Scholix and Data Usage Metrics WGs are making significant contributions in this area by developing clear guidance on how to collect and share data usage statistics and article-data links. Whereas the Scholix WG has reached the end of two very successful 18 month working group terms, the Data Usage Metrics only just started and will continue the work on DLMs and the adoption thereof.

In this paper, I described how data repositories and publishers can contribute to and participate in these initiatives. The openness of the systems developed offers an infrastructure for collaboration using accepted standards. Community organizations, publishers, data repositories, and service providers can rely on common guidelines and standards to share (re)use information they collect about datasets. The most important next step is for as many organizations as possible to standardize usage counts and contribute usage and citations to the open infrastructure hubs.

Acknowledgements

I would like to thank the Co-ordinator cum Editor, IQRC Journal, Patna Women's College, Patna, IND.

References Références Referencias

1. Bierer, B, Crosas, M and Pierce, H. 2017. Data authorship as an incentive to data sharing. N Engl J Med, 376. DOI: https://doi.org/10.1056/NEJMsb1616595
2. Borgman, C. 2012. The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63: 1059–1078. DOI: https://doi.org/10.1002/asi.22634

3. Burton, A, et al. 2017a. The Scholix Framework for Interoperability in Data-Literature Information Exchange. D-Lib Magazine, 23(1/2). DOI: https://doi.org/10.1045/january2017-burton
4. Burton, A, et al. 2017b. Scholix Metadata Schema for Exchange of Scholarly Communication Links. Zenodo. DOI: https://doi.org/10.5281/zenodo.1120265
5. Burton, A, et al. 2017c. The data-literature interlinking service: Towards a common infrastructure for sharing data-article links. Program, 51(1): 75–100. DOI: https://doi.org/10.1108/PROG-06-2016-0048
6. Cantu-Ortiz, F. 2017. Research Analytics: Boosting University Productivity and Competitiveness Through Scientometrics. Boca Raton: CRC Press. DOI: https://doi.org/10.1201/9781315155890
7. Cousijn, H, et al. 2018. A data citation roadmap for scientific publishers. Scientific Data, 5. DOI: https://doi. org/10.1038/sdata.2018.259
8. Fenner, M, et al. 2018. Code of practice for research data usage metrics release. PeerJ Preprints, 6(e26505v1). DOI: https://doi.org/10.7287/peerj.preprints.26505v1
9. Kratz, J and Strasser, C. 2015. Making data count. Scientific Data, 2. DOI: https://doi.org/10.1038/sdata.2015.39
10. Make-Data-Count. 2018. Implementing the COUNTER Code of Practice for Research Data in Repositories. Github. Available at: https://github.com/CDLUC3/Make-Data-Count/blob/master/getting-started.md [Last accessed 30 August 2018].
11. Mongeon, P, et al. 2017. Incorporating data sharing to the reward system of science. Aslib Journal of Information Management, 69: 545–556. DOI: https://doi.org/10.1108/AJIM-01-2017-0024
12. Piwowar, H and Vision, T. 2013. Data reuse and the open data citation advantage. PeerJ, 1(e175). DOI: https://doi.org/10.7717/peerj.175
13. Rauber, A, et al. 2016. Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use. TCDL Bulletin, 12(1).
14. Silvello, G. 2018. Theory and practice of data citation. JASIST, 69(1): 6–20. DOI: https://doi.org/10.1002/ asi.23917