Gauging openness, measuring impact

William Gunn
Mendeley

Section 8:
Scholarly communication

Gauging openness, measuring impact

William Gunn
Head of Academic Outreach,
Mendeley Ltd

This article examines the linked concepts of openness and usability as applied to scholarly works. Openness is used to mean many different things, from transparency about influence when used in a political context, to the lack of restrictions on use when used in a software context. In the scholarly domain, openness generally refers to unrestricted, free availability of a research product over the internet. A work is considered open if there are no permission or price barriers between the work and an individual seeking to make use of the work. However, there are different levels of openness, which are defined by the types of reuse permitted.

Sir Tim Berners Lee introduced the concept of 5 star open data back in 2006 to describe the continuum from a table rendered as a PDF through to data marked up as RDF and connected to the web of Linked Open Data (1). This system clearly explained the benefits of open data by demonstrating how more value was added at each successive step of openness.

A similar scenario is presented with scholarly works. The more open it is, the more useful it is to the author and the audience (2). The first level is simply availability online, as opposed to only as a printed copy. The next level is free to read - you can read the paper without any subscription barriers. A work which is explicitly openly licensed is even more open, but the variety of open licenses leaves many works encumbered with provisions that make it impractical to reuse other than on an item-by-item level (3). Using a license without those provisions would be a further level up. This is the level of the accepted standard license for open access works, CC-BY (4). With CC-BY, there are no explicit barriers to reuse, up to the point that simply tracking and attributing all the providers themselves becomes an unmanageable task. The final level would be fully open with no restrictions of any kind, as with CC0. Each of these levels raises the ceiling value for the amount of reuse possible, while making no statement about the desirability of the work, or the sustainability of the access. Simply put, a work that's more open has, in theory, higher usability than one that is less open. If an open work is also useful to a sufficient number of people, sustainability of access is generally easier to maintain than for closed works through the LOCKSS principle (5), because at least one copy will exist for each researcher who finds it useful. In contrast, closed works can fall into "orphan" status, where reproduction is desired but not permitted, because the rights holder can no longer be identified. Openness is particularly important for works where a long incubation time may be required before the work finds its full potential. Indeed, many great historical works would have been lost were it not for the diligent copying and recopying by centuries of scribes.

What kinds of reuse exist?

The ways in which research can be reused can be divided into five general categories based on application: inspiring new research, mining existing data for novel associations, application or implementation, contribution to the popular understanding, and meta-analysis. The various types of reuse and how these can be tracked for discovery and assessment, briefly discussed below, will be the subject of a forthcoming NISO whitepaper.

The first kind of reuse, inspiring new research, is well covered by the traditional databases which track citations, but is limited

★ make your stuff available on the web (whatever format)

★★ make it available as structured data (e.g. excel instead of image scan of a table)

★★★ non-proprietary format (e.g. csv instead of excel)

★★★★ use URLs to identify things, so that people can point at your stuff

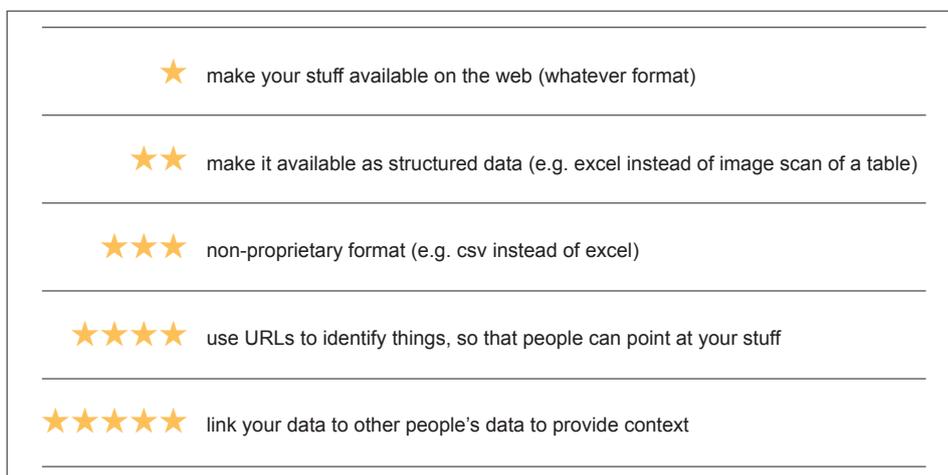★★★★★ link your data to other people's data to provide context

Figure 1: Tim Berners Lee's 5 star open data scale.

in that a subsequent piece of research points to a prior piece, but the prior piece does not reciprocally point back to the subsequent research it inspired. This type of reuse is inhibited through lack of access to the research. Additionally, the pointer is at the document level, which gives poor resolution of the details of the reuse. Another needed improvement for understanding citation behavior is to enrich a citation by adding distinguishing characteristics that would allow the different types of citations to be distinguished from one another. See the Citation Typing Ontology (CiTO) for the current work in this area (6).

Tracking mining of datasets, the second category of reuse, is often done via tracking the papers which describe them (7). However, more datasets are appearing on sites such as Figshare and Dryad, which assign DOIs (Digital Object Identifiers) to the data directly (8), instead of just a paper describing the data. Creating URIs (Uniform Research Identifiers) which point to the data directly promotes the data to equal standing with a research paper, because the data can now be referenced directly and can accrue reuse separately from the paper. As with citation of papers, access to data is a barrier to reuse, but technical skills and equipment to handle the data are also needed.

When you move out of the scholarly realm and into applications, there are less explicit mentions of the original works themselves. Detection of a reuse event in a commercial application can be done via looking for references in patent applications or publications arising from academic/industry collaborations, but this only shows first-order impact at best. As you move further away from the publication into the inventions or policies that it may have enabled or informed, the trail gets very difficult to follow, even as the raw number of possible reuse events grows. This is where individual efforts such as the implementation of a Becker Model analysis (9) become necessary, though this is prohibitive to do at scale.

Looking at the reuse of a scholarly work by the public is done much as with an application or implementation. The main source of reuse events in this category are mentions in popular media, although there is a significant "long tail" of lay communities online which discuss research: patient communities, space aficionados, citizen scientists, and teachers in non-professorial roles. Interestingly, PubMed Central reports that the majority of the page views to research papers hosted there come from non-institutional domains (10). Another

notable feature of reuse within the public domain is that the direction of flow is reversed: external events such as natural disasters, celebrity endorsements, or other news events often drive increased public reuse events (11, 12), whereas availability of a technology facilitates the application.

Meta-analysis is its own category of reuse. There is a growing movement to conduct and publish replication studies of existing work, such as the Reproducibility Initiative and the Reproducibility Project: Cancer Biology, a partnership between the Reproducibility Initiative and the Center for Open Science. The aims of these projects are to understand and promote replication of research as a type of reuse. The replication studies contain pointers to the original research and explicitly identify which experiments were carried out and what the results were. This enables the creation of a separate discovery layer, to highlight and identify the more reproducible or the most reusable work, facilitating downstream commercial application or reduction to practice.

Bootstrapping discovery of reuse

Open Access and Open Data have now become funder priorities across the world. Because funding agencies such as the NIH and Wellcome are now paying for openness in order to maximize the reuse potential of their funded outputs, it has become important be able to aggregate reuse events and to understand their relative impacts. Detecting a reuse event is challenging with current technology, primarily because reuse events don't always point back to the original item. To serve these needs, the Association for Research Libraries, with funding from Sloan and the Institute for Museum and Library Services, is building the Shared Access Research Ecosystem, an event aggregator, which will consume data sources which report on research events. Additionally, the scholarly metadata organization CrossRef is working on a service called Prospect, which aims to facilitate text and data mining of proprietary content (i.e. the data is open at the one star level, but efforts are made to make it as usable as possible). Together with technologies such as Mendeley and Impact Story, we are developing an ever clearer understanding of the importance and value of openness to the research world and society at large.

References:

1. 5 star Open Data. Available at: http://5stardata.info/

2. Piwowar, H.A., Day, R.B.S. & Fridsma, D.S.B. (2007) "Sharing detailed research data is associated with increased citation rate", PLOS ONE, Vol. 2, No. 3, e308.

3. Dryad (2011) "Why does Dryad use CC0?", Dryad news and views on WordPress.com. Available at: http://blog.datadryad.org/2011/10/05/why-does-dryad-use-cc0/

4. Open Access Scholarly Publishers Association (2012) "Why CC-BY?". Available at: http://oaspa.org/why-cc-by

5. LOCKSS, "Preservation Principles". Available at: http://www.lockss.org/about/principles

6. S. (2013), "CiTO, the Citation Typing Ontology". Available at: http://www.essepuntato.it/lode/ http://purl.org/spar/cito

7. Piwowar, H.A. (2010) PhD Thesis: "Foundational studies for measuring the impact, prevalence, and patterns of publicly sharing biomedical research data", D-Scholarship@Pitt (Database). Available at: http://etd.library.pitt.edu/ETD/available/etd-04152010-115953/

8. Piwowar, H.A. & Vision, T.J. (2013) Data from: "Data reuse and the open data citation advantage". Available at: http://doi.org/10.5061/dryad.781pv

9. Holmes, K.L. & Sarli, C.C. (2013) "The Becker Medical Library Model for assessment of research impact – an Interview with Cathy C. Sarli and Kristi L. Holmes", Research Trends, Issue 34. Available at: http://www.researchtrends.com/issue-34-september-2013/the-becker-medical-library-model/

10. Plutchak, T.S. (2005) "The impact of open access", J. Med. Libr. Assoc., Vol. 93, No. 4, pp. 419-421.

11. Bauer, M.W., Allum, N. & Miller, S. (2007) "What can we learn from 25 years of PUS survey research? Liberating and expanding the agenda", Public Underst. Sci., Vol. 16, No. 1, pp. 79–95.

12. Monzen, S. et al. (2011) "Individual radiation exposure dose due to support activities at safe shelters in Fukushima Prefecture", PLOS ONE, Vol. 6, e27761.