Metrics and attribution: my thoughts for the panel at the ORCID-Dryad symposium on research attribution

Martin Fenner, Gobbledygook

May 31, 2013

This Thursday I take part in a panel discussion at the Joint ORCID – Dryad Symposium on Research Attribution. Together with Trish Groves (BMJ) and Christine Borgman (UCLA) I will discuss several aspects of attribution. Trish will speak about ethics, Christine will highlight problems, and I will add my perspective on metrics. This blog post summarizes the main points I want to make.

Oxford. Source: Wikimedia Commons

Scholarly metrics can be used in discovery tools, as business intelligence for funders, research organizations or publishers, and for research assessment. For all these scenarios – and in particular for research assessment – it is important to not only collect metrics for a particular journal publication, dataset or other research output, but to also link these metrics to the creators of that research output. That is why unique identifiers for researchers, and ORCID in particular, are so important for scholarly metrics, and this is also reflected in the ORCID membership of organizations such as Thomson Reuters, Elsevier/Scopus, Altmetric or F1000Prime who provide metrics in a variety of ways.

DORA

A good starting point for any discussion on metrics for research assessment is the San Francisco Declaration on Research Assessment (DORA) that was published las week, together with a set of editorials in several journals, including the *Journal of Cell Biology*, *Molecular Biology of the Cell*, *EMBO Journal*, *Science*, *Journal of Cell Science*, and *eLife*. The first three recommendations are a good starting point for the panel discussion:

1. *Do not use journal-based metrics, such as Journal Impact Factors, as a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions, or in hiring, promotion, or funding decisions.*

2. *Be explicit about the criteria used in evaluating the scientific productivity of grant applicants and clearly highlight, especially for early-stage investigators, that the scientific content of a paper is much more important than publication metrics or the identity of the journal in which it was published.*
3. *For the purposes of research assessment, consider the value and impact of all research outputs (including datasets and software) in addition to research publications, and consider a broad range of impact measures including qualitative indicators of research impact, such as influence on policy and practice.*

Persistent Identifiers

Before we can collect any metrics, we need persistent identifiers for research outputs. Most journal articles now come with a DOI, but we should make it easier for smaller publishers to use DOIs, as cost unfortunately is still an issue.

Persistent identifiers for data are a much more complex issue, as there a number of persistent identifiers out there (including DOIs, handles, ARKs and purls), in addition to all the domain-specific identifiers, e.g. for nucleotide sequence or protein structures. DataCite DOIs are probably the first choice for attribution, as this is their main use case and they have features that make attribution easier (e.g. familiar to researchers, funders and publishers, global resolver). There are many other use cases for identifiers for data (e.g. to identify temporary datasets in an ongoing experiment), and is of course possible to use several identifiers for the same dataset. CrossRef is of course also issuing DOIs for datasets on behalf of their members, and the publisher PLOS is for example using CrossRef component DOIs for figures and supplementary information associated with a journal article, and is making them available via figshare.

Particular challenges with persistent identifiers for research data include different versions of a dataset, and aggregation of datasets (e.g. whether we want to cite the aggregate dataset, or a particular subset). Persistent identifiers for other research outputs are an even bigger challenge, e.g. how to uniquely identify scientific software.

In addition to persistent identifiers for research outputs, we also need persistent identifiers for researchers. ORCID is obviously a good candidate, as it focusses on attribution (by allowing researchers to claim their research outputs and by integration in many researcher workflows). But it is clear that ORCID is not the only persistent identifiers for researchers, and that we need to link these identifiers, e.g. ORCID and ISNI.

Depending on how we want to aggregate the metrics we are interested in, we might also need persistent identifiers for institutions, for funding agencies and their grant IDs, and for resources such as particle accelerators or research vessels. Unfortunately much more work is needed in these areas.

Attribution

Attribution is then the next step, linking persistent identifiers for research outputs to their creators. Attribution is therefore essential for research assessment. The Amsterdam Manifesto on Data Citation Principles that came out of the Beyond the PDF 2 workshop in March are an excellent document, but are unfortunately missing the important step of linking persistent identifiers for data to the persistent identifiers of their creators.

One important issue related to attribution is the provenance of the claims. Has a researcher claimed authorship for a particular paper, is a data center linking creators to research data, or is a funder doing this? The ORCID registry is built around the concept of self-claims by authors, but will allow the other stakeholders to confirm these claims.

Metrics

Metrics for scholarly content fall into one of three categories:

- Citations
- Usage stats
- Altmetrics

Altmetrics is a mixed bag of many different things, from sharing on social media such as Twitter or Facebook to more scholarly activities such as Mendeley bookmarks or F1000Prime reviews. I therefore expect the altmetrics category to over time further evolve into 2-3 sub-categories.

We are all familiar with citation-based metrics for journal articles. We currently see the long-overdue shift from journal-based citation metrics to article-level metrics (see #1 from the DORA statement above for the reasoning), and as the technical lead for the PLOS Article-Level Metrics project I of course welcome this shift in focus. We also see a trend towards opening up reference lists that will make citation-based metrics much more accessible, and the JISC Open Citations project by David Shotton and others is an important driver in this, as is the Open Bibliographic Data project by OKFN. Until open bibliographic data become the norm, we have to deal with different citation counts from different sources. PLOS is collecting citations from Web of Science, Scopus, CrossRef and PubMed Central, and the citation counts are highly correlated overall (e.g. R2= 0.87 for CrossRef and Scopus citations for 2009 PLOS Biology papers), but for some papers differ substantially. Similar to persistent identifiers, reference lists of publications should become part of the open e-infrastructure for science and not depend on proprietary systems. This makes citation metrics more transparent and easier to compare, and fosters research and innovation, in particular by smaller organizations.

The data citation community has adopted the journal article citation model, and we are starting to see more citations to datasets. Even though data citations look similar to citations of journal articles, many essential tools and services still don't properly handle datasets. The Web of Knowledge Data Citation Index is an important step in the right direction, as is the new DataCite import tool for ORCID. Something that we should pay closer attention to is the citation counts of the paper(s) associated with a dataset. Maybe the major scientific impact is in the data, but scientific practice still dictates to the cite the corresponding paper and not the dataset itself (one of the reasons we see data journals being launched). The DataCite metadata can contain the persistent identifier of the corresponding journal article, thus making it possible to associate the citation count of the corresponding paper with the dataset. This approach is particularly important for datasets that are always part of a paper, as is the case for Dryad. One important consideration is that contributor lists may differ between journal article and dataset, or between related datasets.

Another problem with data citation is that citation counts might not be the best way to reflect the scientific impact of a dataset. We are increasingly seeing usage stats for datasets, and DataCite for example has started in January to publish monthly stats for the most popular datasets by number of DOI resolutions. The #1 dataset in March was the raw data to a figure in a F1000Research article, hosted on figshare.

Similar to citations we see a strong trend for usage stats to move from aggregate numbers for journals to article-level metrics. COUNTER has released a draft code of practice for their PIRUS (Publisher and Institutional Repository Usage Statistics) standard in February, and increasing numbers of publishers and repository infrastructure providers such as IRUS-UK and OA-Statistics are providing usage stats for individual articles.

One challenge with usage stats, in particular with Open Access content, is that an article or other research output might be available in more than one place, e.g. publisher (or data center), disciplinary repository and institutional repository. For PLOS articles we don't know the aggregated usage stats from institutional repositories, but we know that 17% of HTML pageviews and 33% of PDF downloads happen not at the PLOS website, but at PubMed Central.

Altmetrics provide new challenges, but they are also a more recent development compared to usage stats and citations. Similar to usage stats they are easier to game than citations, and for some altmetrics sources (e.g. Twitter) standardization is still difficult. Altmetrics not necessarily measure impact, but sometimes rather reflect attention or self-promotion. We have just started to look into altmetrics beyond the numbers, e.g. who is tweeting, bookmarking or discussing a paper or dataset. Altmetrics provide the opportunity to show the broader social impact (as Mike Taylor from Elsevier explains it) of research, e.g. changing clinical practice or policies.

Contributions

One important aspect to attribution is contribution, i.e. what is the specific contribution by a researcher to a paper or other research output. An International

Workshop on Contributorship and Scholarly Attribution was held together with the May 2012 ORCID Outreach Meeting to discuss this topic. Authorship position (e.g. first author, last author) is used in some metrics, but overall the contributor role is still poorly appreciated in most metrics. David Shotton has proposed a Scholarly Contributions and Roles Ontology (ScoRO), and is suggesting to split authorship credit in percentage points based on relative contributions, but I haven't seen these numbers used in the context of metrics.

Conclusions

Persistent identifiers for people, attribution and metrics are closely interrelated and we have seen a lot of exciting developments in this area in the last two years. The widespread adoption of ORCID identifiers by the research community will have a huge impact on scholarly metrics. But with all the excitement we should never forget that a) there will never be a single metric that can be used for research assessment, and b) that scientific content will always be more important than any metric. I look forward to a great panel discussions on Thursday, and welcome any feedback via comments, Twitter or email.

May 23, 2013: Post updated with minor corrections and additions.