Making the case for sharing with indicators of research data impact

Alex Ball

22 October 2015

This is a transcipt of a talk given as part of the Library Connect webinar 'How to Assist Researchers in Sharing their Research Data'. A recording of the full webinar[1] is available.

Abstract. A talk aimed at librarians, providing an overview of the services and tools that are available to help researchers to see data sharing as a natural and rewarding part of their workflow.

Title slide

Hello, my name is Alex Ball. I currently work for the University of Bath as a Research Data Librarian, but before that I worked for the UK Digital Curation Centre, where I maintained a research data management tools catalogue and produced guidance on topics such as data citation and data impact metrics.

In my talk today I'll be looking at some of the services and tools that are available to help researchers to see data sharing as a *natural* and *rewarding* part of their workflow. I should say at the start, in deference to our hosts, that one tool I won't be talking about is Mendeley Data. That's not a value judgement, it's simply because you'll be getting a much more expert insight into that platform from Joe later on.

Motivation

If you're listening to this, I probably won't have to try that hard to convince you that data sharing is a good thing for the transparency, quality and advancement of science generally. The trouble is that the most tangible benefits go to the people on the receiving end, not to those giving it away, so it can feel like an uphill struggle to convince researchers on the ground to share their data.

Research funders have been applying pressure the best way they can, by making funding conditional on researchers taking appropriate steps to share their data, but mandates only get you so far.

[1] http://libraryconnect.elsevier.com/library-connect-webinars?commid=175949

A truism I have found from working with researchers is that if you force them to do something, they will do it grudgingly and badly. To get them to do it well, they must really *want* to do it themselves. So the emphasis of this talk is about services and tools that make researchers feel good about data sharing by giving them back some hard numbers, numbers that show that the effort they put into making their data available is having a real effect.

Researchers get judged on the quality, quantity and impact of their outputs. Of these, impact is probably the one that worries them the most as it's the one they have least control over. So to speak to those concerns, the services and tools we'll be looking at provide some indicators that can be used to build a story of impact. You might notice some cautious phrasing there. None of the tools I'll mention measure impact in a direct and unambiguous way; they are more suggestive than that. There are all sorts of reasons why a dataset might be cited, downloaded from a repository, or talked about a lot on Twitter, but in each case if the number is remarkable it's worth looking deeper to see if the reason is linked to cases of impact.

Apologies if this is all too obvious to you already, but seeing as how many universities blindly use Impact Factors to measure article impact, I think we need to be clear so that data impact does not go the same way. Indeed, such are the concerns about this that there are initiatives underway to agree on best practices regarding the collection and usage of dataset metrics. I'm involved one of them, CASRAI's Dataset-Level Metrics group[2], which is concentrating on precisely defining certain metrics so they can be compared more easily across services. If you're familiar with the COUNTER standards for e-journal usage statistics, that's the kind of consistency we want to achieve. Our group is also working closely with the NISO Altmetrics Initiative[3], which has a wider remit looking at all sorts of altmetrics and types of output.

It will take some time before we have an agreed set of standards for measuring the impact of datasets, but in the meantime, we mustn't ignore the power of a high score for instilling in researchers a warm glow about data sharing.

I'll start by looking at some services that institutions can offer to their researchers.

Thomson Reuters Data Citation Index

The Data Citation Index[4] is one of the scholarly indexes that make up Thomson Reuters' Web of Science. It works in much the same way as the other indexes, but it has some unique features. It supports a hierarchy of records, so datasets can be grouped into data studies and whole repositories, or broken down into nanopublications. This helps with problems of granularity, so that, say, a citation of a nanopublication also counts as a citation for the whole dataset. Speaking of citations, it doesn't just look for them in reference lists: it also counts less formal citations that might occur in a data access statement, acknowledgements section or abstract. The citation count from each peer index is given, and you can follow the network of citations to see how many were attracted by papers that cited the dataset, for example.

[2] http://casrai.org/Research_Dataset-Level_Metrics

[3] http://www.niso.org/topics/tl/altmetrics_initiative/

[4] http://wokinfo.com/products_tools/multidisciplinary/dci

PLoS Article-Level Metrics (Lagotto)

Lagotto is the name of the open source software that underlies the Public Library of Science's Article-Level Metrics service. It can track views, downloads, comments, ratings, notes, citations and mentions on social media, blogs and Wikipedia, and compile them for display on web pages or in custom reports.

This software is aimed at publishers to help enhance their websites. Beyond PLoS it has also been taken up by Copernicus, eLife Sciences, and Pensoft, and is available as a plugin for use in PKP's Open Journal System. CrossRef have even experimented with using it to provide statistics on articles with DOIs across the board.

Although Lagotto was initially aimed at journals, there are moves to adapt it for datasets. The first step on this road was the Making Data Count project[5], a partnership between PLoS, DataONE and the California Digital Library. Last month DataCite announced they would be taking that work forward and piloting a dataset-level metrics service, so that is one to watch.

PlumX

PlumX[6] collates a wide range of statistics about the reach that various types of scholarly outputs have, including datasets and software. These statistics are grouped into five categories:

- *usage* refers to the number of times a resource has been viewed, downloaded, or in the case of software, the number of code contributors;
- *captures* refers to the number of times someone has marked it as being of interest in a service such as Mendeley, Delicious or GitHub;
- *mentions* refers to blog posts that mention the resource, or the number of comments or reviews it has attracted;
- *social media* refers to the number of recommendations – likes, plus-ones, upvotes and so on – the resource has received on social media;
- *citations* has the obvious meaning, with the data coming from Scopus, CrossRef and elsewhere.

The service provides a diagrammatic representation of these figures, known as a Plum Print, that can be embedded in an institutional repository record or an author's own website.

Altmetric

Altmetric[7] is superficially similar but uses a less diverse set of metrics to generate a single score. It concentrates on counting the number of times a resource is mentioned

[5] http://wmdc.lagotto.io/

[6] http://plu.mx/

[7] http://www.altmetric.com/

in blogs, social media, open review platforms, newspaper websites and grey literature, or added to a Mendeley or CiteULike library. Each mention is then weighted according to source, author, and intended audience before contributing to the final score. It can also collect download information from some sources though these don't count towards the score.

Again, the service provides a diagrammatic representation known as a badge or doughnut for embedding in web pages. Unlike PlumX, researchers can use Altmetric for free on an individual basis, while the institutional version provides additional reporting tools and aggregate statistics.

Even though the primary target is journal articles, there are examples of it being used for data on figshare, for example.

ImpactStory

ImpactStory[8] is aimed directly at researchers, and as such is a subscription service costing 60 US dollars per year. Like the last two services, it enables the researcher to build up a profile of their outputs and track relevant statistics, but its unique selling point is that it tries harder to put them into context. It compares them against other resources of the same type from that year and flags up those metrics in which the resource has done comparatively well.

For example, ImpactStory tracks the number of times a dataset in figshare has been viewed, downloaded or shared. If the dataset was in the top 50% of datasets uploaded to figshare that year for shares, it would be marked as highly recommended. If it didn't rank quite so highly but was still up from zero, it would be marked as recommended. Mousing over the flag would give you the absolute figure and the relevant percentile as a kind of short impact story, hence the name of the service. ImpactStory uses some other descriptors like viewed, discussed and cited, and distinguishes what it thinks of as scholarly and public attention.

Other notable services

There are just a few other services I want to give a brief mention to.

- ResearchGate[9] is a social networking site for researchers. For outputs added to a user's profile, it tracks the number of times each has been viewed *on* and downloaded *from* the ResearchGate site and also tracks citations. These, along with the user's activity on forums and the number of followers they have, contribute to an RG Score.

- Google Scholar[10] allows researchers to set up profiles and 'claim' scholarly works. It will then use its citation counts to calculate a total citation count, an

[8] https://impactstory.org/
[9] https://www.researchgate.net/
[10] https://scholar.google.co.uk/

h-index, and i10-index for the past 5 years and for all time. There is no explicit support for claiming datasets, but you can fudge it by entering the dataset details manually using the 'other' category and putting the identifier in the 'Report number' field.

- Microsoft Academic Search[11] creates profile pages for all authors it identifies from the published literature, and these can be curated by the community. These profile pages also carry annual and total citation counts. The focus is very much on journal and conference papers, so you can't use it for tracking datasets directly, but you could use it for *data papers*.

Data papers and data journals

If you don't know what data papers are, they are papers that describe how a dataset was collected, what it contains and how it has been structured, but does not draw out any scientific conclusions. You would most often find them in dedicated data journals such as *Earth System Science Data*[12] or the *Journal of Open Archaeological Data*[13].

Data papers are intended both to provide the necessary documentation to enable reuse, and to act as a proxy for the dataset in traditional bibliometric systems. So if a researcher is comfortable dealing with citation counts and things like the *h*-index, data papers might be the way to get them interested in getting credit for their data.

There are moves to make it easier for researchers to write data papers. In the UK, a group including Oxford University, Faculty of 1000 Research and Wiley are running a project called 'Giving Researchers Credit for their Data'. The title's a bit opaque but it's developing an API and helper application to allow researchers to take a dataset that has been accepted into an archive or repository, package it up with metadata and documentation, and submit it to a data journal in a matter of clicks. A report on the first phase of this work[14] is available from figshare.

figshare

I've mentioned figshare[15] a few times, so I think its worth looking a bit closer. At Bath, we have a handful of researchers who are really keen on using figshare; for them, it has set a standard for ease of use against which our university systems must compete.

Researchers can sign up for free and then start uploading datasets, images, multimedia files, code, or even documents. At a minimum, they have to provide a title, author, an abstract and some keywords, though there is an institutional version of figshare avaiable that allows the institution to apply a different set of requirements.

[11] http://academic.research.microsoft.com/
[12] http://www.earth-system-science-data.net/
[13] http://openarchaeologydata.metajnl.com/
[14] http://doi.org/10.6084/m9.figshare.1483297
[15] http://figshare.com/

In exchange for that metadata, their resource is made available online with a sample citation and a DOI, and the site tracks how many times it is viewed or shared with other users. Behind the scenes it tracks the number of downloads; in future it promises to track citations as well.

Final thoughts

I think those of us managing institutional data repositories can learn a lot from figshare's approach. The way to get researchers on side is to make it as easy as possible for them to share data, but the way to keep them sharing, and influence how they share for the better, is to give them that feedback that shows it makes a difference.

So yes, show them how often their dataset has been viewed or downloaded, as we've seen DRUM has done. Make that information available through an API so services like ImpactStory can use it as well. Provide them with a stable citation and identifier for their dataset, so citations are easier to track.

That way, not only will they get quantitative feedback about the attention their datasets are attracting, but with any luck, they will see that the datasets they have documented better, described better, and linked up with publications, will get more attention. And that will make them want to share *more*, and share *properly*.

Further information

I hope that has been useful. If you want to know more about this topic I recommend looking at the How-to Guide I wrote for the Digital Curation Centre with Monica Duke.[16]

There's a whole other talk I could give you on making your data repository citation-ready, but if you want to know more about that, I recommend first reading the FORCE11 Joint Declaration of Data Citation Principles,[17] then having a look at the How-to Guide Monica and I wrote on data citation, particularly the second half.[18]

Thanks for listening.

[16] Ball, A. & Duke, M. (2015). *How to track the impact of research data with metrics.* Edinburgh, UK: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/how-guides/track-data-impact-metrics

[17] FORCE11, Data Citation Synthesis Group. (2014). *Joint declaration of data citation principles.* Retrieved from https://www.force11.org/datacitation

[18] Ball, A. & Duke, M. (2015). *How to cite datasets and link to publications.* Edinburgh, UK: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/how-guides/cite-datasets