Better data-level metrics: Quality and impact

A workshop held as part of the BioMedBridges Symposium ‘Open bridges for life science data’ at the Wellcome Genome Campus, Hinxton, UK, 17–18 November 2015:

Welcome and introduction

Welcome to this workshop on better data-level metrics for quality and impact. Over the course of the next two hours we’ll be hearing from four different speakers about the various initiatives that are or have been working in this space. Then, at the end, we want to hear what you have to say, so we can feed your ideas and concerns back into those initiatives. But before we get to that, I just want to set the scene a little and explain why we wanted to hold this workshop.

As I’m sure you’re aware, the past decade has seen a sharp rise in expectations for data sharing across the whole research community. You will have heard the arguments: that sharing the underlying data helps to make the research more trustworthy, and that old data can be mashed up, recombined and re-analysed to extract new results.

But in those areas where data sharing is a new and troubling idea, simply telling researchers they have to share isn’t enough. For one thing it annoys them, and for another, if they really want to get out of sharing their data, they will either find a way to justify withholding it or else share it in such a state that no-one else could possibly make use of it.

There needs to be a change of culture, and for that to happen, data needs to be treated just the same as other scholarly outputs that researchers happily give away: their publications. In other words, there needs to be a similar set of controls in place to ensure that researchers get due credit, recognition and reward for the data they share. There needs to be a culture of data citation, with corresponding guards against plagiarism, and procedures for assessing the quality and impact of datasets.

For publications, the gold standard for assessing quality and impact is the peer review system. But again, over the past decade or so there have been rumblings about its scalability and effectiveness. Policy makers have started toying with the idea of using quantitative metrics, something that many academics are deeply unhappy about, and with good reason. I expect we can all quote a horror story of someone suffering at the sharp end of a misused journal impact factor at their institution.

But for data, there is more of a clean slate to work with, and very clearly not enough expert peer reviewers to do a thorough job of evaluating the vast amount of data being produced. We have an incentive to try and make metrics work for datasets, and perhaps that will show us the way to use them in a broadly acceptable way for regular journal publications as well.
Abstract

The drive towards accepting datasets as valid scholarly outputs, worthy of recognition and reward, has met with concerns about how to gauge the degree of their worth. With skilled peer reviewers for datasets few and far between, there has been considerable focus on the use of quantitative metrics to help assess data quality and impact. This presentation looks at work performed in this area by the European forum Knowledge Exchange, the UK Digital Curation Centre, and the international standards body CASRAI.

So now it’s time for me to change hats and tell you about my personal journey in this landscape. I used to work for the Digital Curation Centre, where I took an interest in data publication issues such as data licensing and data citation, and the latter was my way in. Between 2011 and 2013, organizations like CODATA made great strides building the case for data citations, and that seemed to trigger a renewed interest in whether those citations could be measured, and in the absence of widespread adoption, what other impact metrics could be used. So I started research into dataset-level metrics, and these three initiatives feature prominently in that research.

Contents

1 Knowledge Exchange 3

2 DCC How-to Guide 5

3 CASRAI Data-Level Metrics 7


Table 1: Metrics scenarios for data publishing models (adapted from Costas, Meijer, Zahedi, & Wouters, 2013, p. 16)

<table>
<thead>
<tr>
<th>Type</th>
<th>Tools</th>
<th>How used</th>
<th>Paper</th>
<th>Shared dataset</th>
<th>Published dataset</th>
<th>Data paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citations</td>
<td>Data Citation Index, Google Scholar, Scopus, DataCite</td>
<td>Size dependent</td>
<td>Y</td>
<td>Hard</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Direct average</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Source-based</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Social media</td>
<td>ImpactStory, Twitter, Facebook</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>indicators</td>
<td></td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Readership</td>
<td>Mendeley, CiteULike</td>
<td></td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>counts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Downloads, Views</td>
<td>Repositories, Journal platforms</td>
<td></td>
<td>Y</td>
<td>Hard</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

1 Knowledge Exchange

My first foray into dataset-level metrics was courtesy of the Knowledge Exchange. This is a forum that was set up to allow national providers of IT services to the academic sector to share good practice and tackle common issues in a collaborative way. The partners include

- CSC (IT Centre for Science), Finland
- Denmark’s Electronic Research Library (DEFF)
- German Research Foundation (DFG)
- Jisc, UK
- SURF, the Netherlands

In 2013, they published a state-of-the-art report ‘The Value of Research Data’,\(^1\) which looked at the kinds of metrics available, stakeholder perceptions of them, problems, challenges and possible solutions. I took part in the workshop they held to launch the report, somewhat confusingly also called ‘Making Data Count’.

I won’t go through the whole thing but there are some points I want to draw out, as I think they might be useful to inform our discussion today. The first is this table (Table 1), which provides a classification of different types of metrics and where the data for them might come from.

It distinguishes three types of citation-based metrics:

- **Size-dependent** means metrics that look at raw performance, for example, the total number of citations of a resource or the $h$-index of a researcher.

- **Direct average performance** means metrics that either take the average performance of a group of outputs or measure the performance of an output or person relative to that of its peers.

- **Source-based performance** are ones that measure the performance of a publication venue rather than a person or output. It’s basically there to give a home to the Journal Impact Factor.

It also distinguishes three types of altmetrics beyond citations:

- **Social media indicators**: mentions, likes;

- **Readership counts**: that is, the number of people reached;

- **Downloads and views**: that is, the number of times a resource is accessed.

I should note that this is not quite how the table appears in the report. The report makes a strong distinction between citation counts on one hand and altmetrics on the other, which is why it doesn’t drill into the different ways altmetrics can be used. That’s a pity, because it would highlight that ImpactStory, for example, provides relative metrics and would thus fall under direct average performance.

A note of hope is that so long as data are published properly through an archive or data journal, as opposed to being thrown up on a project website, there is no impediment in principle to measuring the performance of datasets in an analogous way to journal papers.

The second thing worth extracting from that report are the challenges it identifies.

- **Vicious cycle** – *We need to find metrics that can bootstrap the system and turn this into a virtuous circle.*

  ![Diagram](Diagram.png)

  Scholars don’t share because it is not rewarded and not worth the time. Data sharing/reuse transactions are few, making metrics hard or impossible.

  Without metrics, it is hard to find a basis for rewarding data sharing.

- **Need to take access restrictions and embargoes into account** otherwise we will end up disenfranchising those working with sensitive data.

- **Need to take scale of data sharing within disciplines into account** – *since there is vastly more astronomical data in circulation than clinical psychology data.*
• Why measure the impact of research data

• Impact measurement concepts

• Impact measurement services

• Current issues and challenges

• Tips for raising research data impact

Figure 1: How to Track the Impact of Research Data with Metrics

Lastly, I want to pick out some of the recommendations that came out of the report and the associated workshop.

• Share good practice regarding data quality review in disciplines.

• Standardize and formalize data citation – that’s happening through initiatives like FORCE11 and the Research Data Alliance.

• Ensure robust linking between data and publications – again, the RDA is working on technical solutions, and the DCC among others are providing advocacy.

• Immediately start collecting metrics that might be used – I expect we’ll hear more on that from Martin and Sarah.

• Work out which metrics are appropriate for data – and that’s happening in the NISO and CASRAI initiatives.

2 DCC How-to Guide

Attending that workshop was one of the first things I did to prepare for writing a practical guide to data impact metrics (Figure 1). I should explain that the Digital Curation Centre produces guidance for the UK Higher Education sector on matters to do with digital curation and research data management. My colleague Monica Duke and I were commissioned to write a How-to Guide on the subject in January 2013, but back then there wasn’t a great deal we could say, so it wasn’t until June this year that we actually published it.

The central part of the guide is a run-down of the various tools providing dataset-level metrics: what they measure, at whom they are aimed, how much they cost, and so
on. I won’t repeat all of that now, though, as the guide puts it well enough. Instead, I’d like to offer some reflections on the difficulties we had writing it.

¶ The first issue was one of scope. We started off with the title ‘How to Measure the Impact of Research Data’, but then someone said, it’s not really possible to measure impact directly. You can use metrics to indicate or demonstrate or suggest impact, but you can’t get a reliable measurement of it because lots of things can cause the figure you get. You need an element of human interpretation. For example, if an dataset is mentioned a lot on social media, is that because it is having a positive impact on the discipline, because it has an oddball or titillating title or abstract, or because people are shocked at how bad it is?

§ So we changed the title to ‘How to Track the Impact of Research Data’. But then people pointed out that impact has a huge number of dimensions, most of which weren’t covered by the metrics we presented in the guide. So to avoid mutating it into a guide on writing impact statements for data, we changed the title one last time to ‘How to Track the Impact of Research Data with Metrics’.

§ Even so, we couldn’t avoid the fact that we were entering into politically sensitive territory. In our early reviews we were criticized for lack of balance: we’d said why you might want to use metrics but hadn’t mentioned the dangers of going too far. It was pointed out that reliance on metrics encourages a game play mentality, where people are encouraged or indeed forced to score points rather than do great and innovative things. Some even questioned whether impact was the right thing to measure but frankly the guide was not the right place to argue that particular point. But it did highlight for us the need to consider a suite of metrics and not to rely on just one or two.

¶ Here are some other issues that came up in the review process.

We were asked to include additional warnings about measuring Twitter: tweet counts might be

- inflated by auto-tweets from repositories, publicity engines, or fraudster accounts
- suppressed if tweeting ‘not the done thing’ in a community – *this is the case among German academics, so I’m told*

We were encouraged to consider how metrics might be used to create fairer Credit models

- Need to go beyond *the* download stats for *the* dataset itself
- Consider the network of reuse: transitive credit – *can we take into account metrics relating to derived datasets?*

Lastly, I had a comment on something I’d written about Impact and quality, which was that

- Researchers can do more about ensuring the quality of their outputs than they can about impact, since the latter depends on many factors outside their influence
- *But that doesn’t mean that quality is any easier to measure.*

Might it be more tractable, though? Hopefully we’re about to find out.
3 CASRAI Data-Level Metrics

Earlier this year I was asked to join the CASRAI Dataset Level Metrics Interest Group, led by Rebecca and Sünje.

For anyone who doesn’t know, CASRAI is the Consortium Advancing Standards in Research Administration Information and as the name suggests it develops standards for the exchange of research administration information.

The Dataset Level Metrics Interest Group is initiating a range of projects to support the interoperability of administrative information, or what you might call paradata, about research datasets. So the kinds of things we’re interested in producing are terminology and vocabularies, sets of clearly defined metrics, and proof-of-concept implementations.

The way the CASRAI process works is that they first agree a clear and focused use case, then work out what information the various actors and systems need in order to achieve success, and lastly agree on a precise specification for that information so that the various stakeholders can interoperate with it and use it consistently.

These are the use cases we’ve identified so far are:

- A publisher wanting to display metrics for datasets underlying articles so authors have an idea of the impact of their data
- A funder wanting to evaluate funded datasets using a common set of impact-related metrics
- A university administrator wanting to evaluate the impact of current and potential staff using a common set of impact-related metrics
- A data repository (domain-specific or generalist) wanting to demonstrate to stakeholders the quality and impact of their holdings
- A general data repository wanting metrics or criteria for evaluating the quality of incoming datasets – they need to be clear so they can be communicated to depositors, they need to be applicable to a wide range of disciplines, they need to be stringent enough to ensure a useful level of quality, but not so strict as to discourage deposits.

This last use case is the one we’re dealing with first, so we have set up the Quality Dataset Level Metrics for Repositories Working Group which will work in a three-month sprint to specify those metrics or criteria.

If you think we should look at, please let me know.

- CICSNC Data Stewardship Maturity Matrix
- NERC Data Value Checklist
- FORCE11 FAIR Principles
• UK Data Archive data review
• University of Edinburgh Deposit Checklist

¶ In our initial brainstorm we came up with the following question set.

• What potential impact/reuse do you anticipate from this dataset?
• How has the dataset been licensed?
• Are there legal or access requirements for the data?
• How has your dataset been documented for reuse?
• Is the data structured and labelled?
• To what degree is your dataset reproducible, *in the sense that the processed data could be derived again from the raw*?
• Have you provided additional disciplinary-associated metadata?
• Has the data been peer reviewed (externally)? What was the result?
• What kinds of format(s) is your dataset in?
• Was the data collected according to an accepted methodology? *If so, have you referenced it?*
• Does your submission have integrity?

Again, we would value your input on whether that’s a reasonable set of questions to ask, what form we should collect the answers in, and perhaps the weighting we should give to each.

¶ If you find this whole area fascinating and would like to be more actively involved, that would be great too: see the starred roles in the list below:

**WG Chair** Alex Ball – *responsible for WG membership*

**WG Scribe** Rebecca Lawrence (notes in meetings); David Baker/Chair/Point Person(s) (blog posts)

**WG Point Person(s)***  TBD – *subject specialist answering queries from CASRAI*

**WG Facilitator/Coordinator** CASRAI Secretariat (David Baker) – *runs the sprint and schedules meetings*

**WG Modeller** CASRAI Secretariat (Paul Kiel) – *formalizes the outputs for adoption*

**WG Participants***  

---

University of Bath. For more information, please visit http://www.bath.ac.uk/

CASRAI Quality DLMs for Repositories WG: http://www.casrai.org/Quality_DLMs_for_Repositories