When counting is hard

September 2 2015, by Jennifer Lin

Counting is hard. But when it comes to research data, not in the way we thought it was (example 1, example 2, example 3. The Making Data Count (MDC) project aims to go further – measurement. But to do so, we must start with basic counting: 1, 2, 3… uno, dos, tres…

MDC is an NSF-funded project to design and develop metrics that track and measure data use, "data-level metrics" (DLM). DLM are a multi-dimensional suite of indicators, measuring the broad range of activities surrounding the reach and use of data as a research output. Our team, made up of staff from the University of California Curation Center at California Digital Library, *PLOS*, and DataONE, investigated the validity and feasibility of using metrics by collecting and investigating the use of harvested data to power discovery and reporting of datasets that are part of scholarly outputs.

To do this, we extended Lagotto, an open source application, to track datasets and collect a host of online activity surrounding datasets from usage to references, social shares, discussions, and citations. During this pilot phase we ran DLM against our sample test corpus of all datasets in DataONE member repositories (~150k). The overall profile of dataset usage appears to be significantly different from scholarly journal articles in the early DLM data collected.

Counting what we cannot see

Within this spectrum of indicators, citations – the single focus of this blog post – are considered by far the most interesting metric to both

researchers and data managers (Kratz and Strasser, doi.org/10.1038/sdata.2015.39). However, citations currently pose a difficult challenge for measurement. Article citation services are fairly well established – some openly available (PubMed Central), others require subscriptions to gain access (Scopus and Web of Science) or publisher membership to participate (Crossref). To date only one major data citation service exists across journals and it is relatively untested by the community, in part due to its subscription cost: Thomson Reuter's Data Citation Index. There are, however, other initiatives beginning to explore this arena such as BioCaddie, an NIH Big Data to Knowledge initiative. One of the biggest challenges in understanding dataset usage is associated with researcher practice in how datasets are cited. By and large, researchers do not cite datasets and where they are cited there is a great degree of variance in practice. Datasets may be mentioned in numerous places within the main body of the paper and/or formally cited in the reference section. There are emerging efforts to standardize this practice and enable the propagation of research data across scholarly communications by publishers, data repositories, and funders.

The DLM approach

Without access to a tested and openly available index for data citations across journal literature, what are we to do today? Since we cannot plug into a system that has already aggregated them (like most sources), we have to mine the literature directly and collect connections in the system. How does the DLM application do this and what content is targeted? The DLM application conducts full text searches across a corpus of content, looking for any mention of a dataset via its persistent identifier (DOI, ARK, etc.). Regardless of the location of the dataset persistent identifier – methods section, results, reference list, figure legend, etc., – the DLM application is able to find it. DLM uses publisher open APIs to do this. To date, we have implemented this search for all publications from *PLOS*, BMC, and *Nature* (all journals listed in nature.com). Additionally,

we ran the full text search on articles indexed in the Europe PMC corpus, a mirror of PubMed Central's corpus with 3.3 million articles available.

Researchers who publish in journals whose content is openly available (i.e., open access) or their corpus is openly searchable via an API have a citation advantage and a better chance of getting credit (Eysenbach G., doi.org/10.1371/journal.pbio.0040157). In a recent analysis of gene expression microarray studies, this phenomenon has also been identified in datasets (Piwowar, H. et al., doi.org/10.7717/peerj.175). At this point, we have verified this more widely in our pilot system with datasets representing a broader range of subject areas (life sciences and geosciences).

Counting events as 'sorta' counts

We found a number of interesting results from data citations collected on DataONE datasets. While full analyses of the data (over time and across channels) is still underway, we are able to share very preliminary findings:

- BMC – 339 events
- PLOS – 741 events
- Nature OpenSearch – 388 events
- Europe PMC – 2107 events

Data collected 12 Aug 2015

The table above shows the number of events collected and stored in DLM for matches found between a DataONE dataset DOI and a publication in the search sources. DLM counts events by detecting the presence of a persistent identifier for a dataset in a paper and picked up by the source API. This is a data citation, in the loose sense of the word.

(Some may define data citations in the formal sense based on requirements such as location, metadata identification, etc.) This brings us to our first finding, which while manifestly evident, is still worth mentioning given the lack of real information on article and linked data connections. The counts represent the set of journals made available in the search. If we are to know the objective count of data citations, we cannot simply sum up all the events collected across citation search sources. Data from publisher-specific sources is unique as they are limited to the exclusive corpus. But Europe PMC data covers the corpus of multiple publishers, some of which have already been covered in the current sources. At the same time, this archive is limited to articles deposited by participating journals, leaving out the majority of articles from subscription-based publishers. In our current configuration with the existing citation source list described, the fullest set would constitute a unique list of dataset-article links from the results of Europe PMC and *Nature* OpenSearch. Early evidence shows that the DLM approach to searching the Europe PMC archive is effective in extracting linked publication connections for OA publishers here, making it redundant to poll individual publishers if already indexed. As we continue to develop the system, future opportunities include expanding the search source list to text mine a broader set of publisher content (Crossref TDM, etc.) and exploring connections with other emerging data literature linking efforts (OLDRADA, DLI Service, etc.).

However, some events captured are not dataset citations in the formal sense. We discovered two types that are arguably not dataset citations, adding unwanted noise. In rare instances, DLM pulls article corrections in a data citation search where the data reference is part of a formal literature update – in this instance it is treated as a separate publication. This has occurred in both *PLOS* and *Nature* corpus.In the former instance, *PLOS* publishes a data access statement along with every article. The new policy which began April 2014 has elevated the visibility and prominence of the underlying data, which may have

contributed to the added corrections. Additionally, *Nature*'s article PDFs published before 2014 are consistently included in the *Nature* OpenSearch count along with online article itself. For example, the Dryad dataset on Somatic deleterious mutation rates (10.5061/dryad.t8q7t) pulls back two events from *Nature*, the associated article and the article PDF. Thus, we are double counting dataset citations in these articles.

Unlike the search sources, DataCite counts capture a different set of elements altogether for our sample corpus: the components in the dataset package are each counted separately, along with the article it was originally part of if applicable. For example, the Dryad dataset doi.org/10.5061/dryad.3qd54 returns a DataCite count of 8 eventsbecause the dataset has 7 component datasets (10.5061/DRYAD.3QD54/1 – 10.5061/DRYAD.3QD54/7) that is linked to one associated article. This tally provides us with meaningful information on the dataset, though the numerical count of events in itself should not be directly included in the data citation calculation.

Counting in the early days (before the abacus)

Data citation is a very messy business in the absence of widespread and consistent data citation practices and DLM data provides further verification of the challenges and downstream impacts. Citations to data are primarily inline and, in rare instances, are located in reference lists. We will investigate the frequency of reference list citations further with the Europe PMC corpus. The presence of datasets are almost always mentioned as part of the research process, though described in a wide range of ways, the least of which make it extremely hard to automatically tag or classify. On rare occasions, the dataset may be handled in the same manner as article references. These are 'wild west' days for data citation as publishers continue to consider policies and practices on how to implement data citations in their submission

systems. Data repositories can also have a significant impact on the data citation and publication end. Currently all the top cited DataONE datasets are associated with Dryad Digital Repository. Dryad actively partners with publishers to integrate data and manuscript submissions workflow, thereby facilitating data-article linkage. Dryad also provides clear instructions to researchers on how to cite datasets deposited in their repository.

Measurement is recognized as difficult, but counting proves to be its own challenge. The MDC pilot has begun to test the design of data metrics and its preliminary results have already begun to offer a richer view into the ways and degree in which researchers are really using scholarly data in the wild. Next up, we will begin to examine usage statistics, another fascinating area which we are eager to dive into.

More information: For full project background information and the latest progress updates, please visit the main MDC project page: mdc.lagotto.io.

This story is republished courtesy of PLOS Blogs: blogs.plos.org.

Provided by Public Library of Science