

The International Journal of Digital Curation

Volume 8, Issue 2 | 2013

The Product and System Specificities of Measuring Curation Impact

Nicholas M. Weber and Andrea K. Thomer,
Graduate School of Library and Information Science,
University of Illinois at Urbana-Champaign

Matthew S. Mayernik, Bob Dattore, Zaihua Ji and Steve Worley,
National Center for Atmospheric Research

Abstract

Using three datasets archived at the National Center for Atmospheric Research (NCAR), we describe the creation of a 'data usage index' for curation-specific impact assessments. Our work is focused on quantitatively evaluating climate and weather data used in earth and space science research, but we also discuss the application of this approach to other research data contexts. We conclude with some proposed future directions for metric-based work in data curation.

Introduction

A quantitative evaluation or assessment of any phenomenon will try to answer two basic questions:

1. What should be counted?
2. How much should it count?

In scholarly communications, citations are typically what is counted, and their position, prevalence and popularity determine how much a citation should count as being evidence of research impact. To measure research impact, many statistical techniques like co-citation analysis (Small, [1973](#)), or the h-index (Hirsch, [2005](#)) have been developed to show how an individual can be evaluated via the citations made to their publications.

More recently, scholarly communications has started to innovate with these methods of analysis by questioning what is counted. In particular, alternative metrics (hereafter referred to as altmetrics) are beginning to leverage the various traces of activity on the social web in re-calculating research impact. Microblogging (aka “tweeting”) (Priem and Costello, [2010](#)), the prevalence of journal articles on social bookmarking sites (Haustein and Siebenlist, [2011](#)), research blogging (Shema, Bar-Ilan, and Thelwall, [2012](#)) and website page-views (Thelwall, [2012](#)) have all been explored as potential alternative, new impact indicators. These altmetric analyses don’t so much critique existing citation-based metrics as much as they offer a complementary means of impact assessment – one that provides a broader, more complete view of knowledge production in contemporary science (Priem, Piwowar and Hemminger, [2012](#)). So, we might say that altmetric studies question not just what is counted, but also how much and even why these new media traces count as evidence of research impact.

To date, most altmetric studies have been aimed at quantifying an individual’s impact on their community of practice (a notable exception is Bollen et al., [2005](#)). Our work here is focused on expanding that horizon, and asking whether or not we can develop assessment techniques that successfully quantify the impact of curation services developed by large groups of people and infrastructures funded by entire institutions. In a sense we’re promoting the same reconsideration of research impact as previous altmetric studies, but we’re doing so at a different level of granularity. We want to reconsider what it means for a service or a system to have research impact.

In this paper then, we have three ambitions:

- 1. To innovate with existing quantitative impact assessment techniques;**

By developing new indicators of how, when and under what circumstances research data are accessed, the curation community might also engage in a broader discussion about how these calculations can gauge the impact of services and infrastructures supporting data-intensive research. We also believe these metrics are an important step in making

curation work more visible to formal institutional reward structures and acknowledged in federal grant funding initiatives.

2. To explicitly and openly discuss the process of developing new research impact metrics;

We consider the process of developing new curation-specific metrics as distinct from (but informed by) traditional citation-based assessments of impact. Our work with the Research Data Archive (RDA) at the National Center for Atmospheric Research (NCAR) suggests that the quantification of curation impact will necessarily have “specificities” that make generalizable metrics a social hurdle as much as they are a technological one. Being explicit about these limitations and openly describing how new metrics were developed is an important first step in giving these techniques credibility within institutions of higher education and federal agencies.

3. To lay a baseline for future curation impact assessments, citation or otherwise.


As data citation initiatives mature within various scientific communities, we also believe it is important to create complementary techniques of impact assessment. This study is a first attempt at developing those methods based on indicators of how and when data are accessed for future use. We do not see our work here supplanting or replacing future efforts in data citation analysis, but instead complementing that work and laying a baseline for future comparisons that can make both efforts more useful for curation stakeholders.

Setting: The RDA at NCAR

The Research Data Archive (RDA) is a repository of atmospheric and oceanographic observational data, weather prediction model output, gridded analyses and reanalyses, climate model output, and satellite-derived data that has been curated by staff in the Computational and Information Systems Laboratory at NCAR for over 40 years (Jacobs and Worley, 2009). The holdings of the RDA are dynamic; many datasets are routinely updated, and new datasets are added each year, with total holdings currently exceeding 1.3 Petabytes.

One of the motivations for this study is to find ways to assess the performance of the RDA beyond a generic “total number of users served” statistic. In particular, we want to highlight and make visible the nuanced or craft-like work that goes into curating heterogeneous large-scale datasets in this environment. Software engineers working in the RDA have a more complex set of responsibilities than their title implies, including at least two activities not mentioned in the digital curation life-cycle model:

1. The creation of data services, which spans a wide range of activities from creating customized sub-setting and format conversions for multi-terabyte-sized datasets to “data rescue” for content stored on out-dated magnetic tapes;

- 
2. Archival content development, which includes activities focused on improving data organization, quality checks on data values, assuring archival completeness of documentation, extensive metadata harvesting to drive local discovery and access, and detailed dataset evaluations in response to user questions and the potential data errors (Jacobs and Worley, [2009](#)).

These two curation activities are currently assessed quite differently. The effectiveness of data services are usually measured through user satisfaction surveys administered on an annual or semi-annual basis; while archival content development is typically evaluated through systems log-analysis, or web analytics that attempt to directly correlate the volume of data downloaded with the quality of the data being served.


Separately, these two techniques are effective for exploring when and how often data hosted by an archive are consumed, but they are also exceptionally labour-intensive and it is often difficult to generalize about “impact” from survey or log-analysis data alone (Bollen et al., [2008](#); Henneken et al., [2009](#)). These techniques also have a difficult time capturing the nuanced work of data curators, including how shifts or changes in services impact end user consumption. This leaves the services and infrastructures, such as those developed by staff at the RDA, invisible to promotion or tenure awards at an individual level, and often ignored or overlooked by federal funding at an institutional level.

A Data Usage Index

One previous attempt at making curation work more visible is Ingwersen and Chavan’s ([2012](#)) Data Usage Index (DUI). Using a combination of web-analytics and log-analysis, this index consists of 14 quantitative indicators that capture different ways that data are discovered and accessed in an archival setting (See Table 1 for full description). The DUI was originally developed to measure the use of species occurrence records from the Global Biodiversity Information Facility (GBIF) database, and was effective in showing how changes within that infrastructure impacted user activity over time.

Ingwersen and Chavan state that the DUI should be adaptable to a new research domain, but that in doing so, ‘...one needs to take into account the fundamental characteristics of datasets and their usage patterns’ within that domain ([2011](#)). In adapting the DUI from a biodiversity setting, we found a number of differences between the ways that users performed searches, but also in their very orientation to “using” climate and weather data.

To return to our original discussion of what counts and what is counted in any impact assessment; in the DUI what is counted are data access events (e.g. downloads, searches etc.), but what should count will be unique to the system and the type of data products being analysed. We refer to the differences between what is counted and what counts in a DUI assessment as the *system* and *product* specificities of measuring research impact.




	Indicator of use	Explanation
$s(u)$	Searched records	Number of records searched/viewed (by IP address) in unit
$d(u)$	Download frequency	Number of downloaded records from unit
$r(u)$	Record numbers	Number of records in (period; dataset(s); geographical and/or species) unit
$S(u)$	Search events	Number of different searches (by IP address) in unit
$D(u)$	Download events	Number of different downloads from unit
$R(u)$	Dataset number	Number of datasets in (period, geographical and/or species) unit
$s(u) / S(u)$	Search density	Average number of searched records per search event
$d(u) / D(u)$	Download density	Average download frequency per download event
$d(u) / r(u)$	Usage impact	Download frequency per stored record per unit
$s(u) / r(u)$	Interest impact	Searched records per stored record per unit
$d(u) / s(u)$	Usage ratio	Ratio of download frequency to searched records in unit
$D(u) / S(u)$	Usage balance	Ratio of download events to search events for unit (in %)
$U(u) / r(u)$	Usage score	Ratio of unique downloaded records (U) to record number (in %)
$l(u) / r(u)$	Interest score	Ratio of unique searched records (I) to record number (in %)

Table 1. The Indicators from Ingwersen and Chavan's Data Usage Index.

Specificities

Oliver Williamson originally used “asset specificity” to describe economic transactions where one firm acted irrationally, or unexpectedly, when trading goods with another firm (1981). Williamson observed that some firms required specific assets, like a particular material, tool, or type of human expertise in order to achieve a desired outcome. A firm requiring these assets had a specificity that locked them into certain transactions, and certain ways of doing business that seemed completely irrational to an unknowing marketplace.

As an example, consider an architectural firm that designs a skyscraper to be made entirely of white marble. Any construction company that they hire to build the skyscraper will be necessarily beholden to a few specific sites in Tuscany, Italy where



Carrera marble (the only kind of white marble strong enough for this scale of construction) is quarried. Carrera marble then is an asset specificity of this building's design: it locks a construction company into certain ways of working, and necessarily limits their choice in acquiring a competitive price on the materials they need to accomplish a task. Without a nuanced understanding of the larger context in which both firms are operating, their actions seem irrational. However, we can understand and begin to better accommodate these types of behaviours if we can find ways to account for and record specificities that constrain marketplace actions.

Malone et al. (1987), and more recently Haythornthwaite (2006), refined Williamson's concept of asset specificity and added new applications of the term, such as institutional, knowledge, structure and system specificities. These types of specificity more explicitly account for external factors that shape the way groups, teams and organizations produce new knowledge, and are limited in their organization and collaboration by specificities introduced through networked information technologies.

When evaluating usage patterns and the characteristics of datasets served by the RDA, we noted two types of specificities that constrained scientists accessing these materials: system and product specificities.

System specificities include the architecture and organization of data hosted by an archive. These specificities limit the way a user can search, browse or access a dataset. Whether data is accessible through a graphic user interface or through a command line tool like 'curl' is an example of system specificity. These externalities shape the way a user can interact with an archive's content, and consequently these specificities are manifest in the user-logs that record how often, and what amount of data a scientist can access in the RDA.

Product specificities are the properties of a dataset – the file structure, format, and size – that affect the way a user can interact with an archive in consuming and discovering data. An example from the RDA is a dataset containing observations made at a NOAA weather station. This dataset will likely contain variables like precipitation or wind speed that are recorded at a sub-daily rate, and a file corresponding to each sub-daily recording. To retrieve a meaningful or complete set of records, an end user often has to consume thousands of files in a single session. Based on file count alone, the downloading an entire weather station's data would seem like a user had consumed a massive amount of data. In reality, the volume of these files might equal only a single gigabyte in size. These externalities make file size, download counts, or even download frequency a product specificity for impact assessments.

Both product and system specificities shape the way that users interact with or access the content of a data archive. In turn, metrics that are developed based on user-archive interactions will necessarily reflect these specificities. In the next section we explore the creation of a DUI unique to the Research Data Archive at NCAR, and note some generalizations that can be gleaned from this process. We then operationalize the DUI to study three datasets hosted by the RDA and discuss the limitations of this work in light of the product and system specificities described above.

Towards a DUI for the RDA

Building on Ingwersen and Chavan's previous work, the first step in adapting a DUI to a new research environment is to define a unit of analysis. This unit must determine:


1. An appropriate level of granularity at which there is a meaningful group of data, or a 'dataset'; and
2. An appropriate time window in which to capture robust user-system interactions.

For the RDA's DUI, we chose to define a dataset as any data product that was issued a unique identifier. Since many datasets in the RDA are dynamic, and will have new content added at regular intervals, we believed that a monthly time window would yield a high enough volume of user-archive interactions for the purposes of our case study.

What to Count? Usage-Based Indicators

Ingwersen and Chavan's DUI was made up of a series of indicators that were derived from events recorded in a system's user-log data, such as the number of files a user downloaded or the number of unique user queries performed in a given month. We similarly rely on these traces of data access to calculate groups of indicators that make up the RDA's DUI (see Table 2).

	Indicator	Explanation
1	Unique users	Unique users that downloaded data during a time window
1a	— Programmatic	Unique users that accessed data programmatically
1b	— Assisted	Unique users that accessed data via GUI or RDA Service
2	Number of datasets	Number of datasets assigned a dataset (DS) number by RDA
3	Files DS	Number of files in dataset per time window
4	Download frequency	Total number of files downloaded per time window
4a	— Programmatic	Files downloaded programmatically
4b	— Assisted	Files downloaded by assisted users
5	Homepage hits	Homepage hits of dataset per time window
5a	— Direct link	—
5b	— Query	Homepage hits of dataset per time window by users with a link from an indexed list or retrieved by search
6	Download density	Average number of files downloaded per unique user



	Indicator	Explanation
7	Usage impact	Total number of downloaded files over total files in dataset
7a	— Programmatic	Usage impact score for programmatic users
7b	— Assisted	Usage impact score for assisted users
8	Usage balance	Files downloaded by number of homepage hits per time window
9	Interest impact	Total homepage hits per number of files in dataset
10	Secondary interest impact	Total homepage hits over unique users
11	Subset ratio	Subset requests over total number files downloaded

Table 2. The RDA's DUI Indicators.

What Counts? A Case Study of Three Datasets

We selected three different datasets from the RDA to test our proposed DUI indicators. These three datasets are a representative sample of the RDA's diverse holdings: one is a set of global observational data (ds540); the next is a popularly analysed dataset derived from a numerical weather prediction centre (ds083); and the last is a complete and exceptionally large global atmosphere and ocean reanalysis dataset (ds093.0-6). In Table 3, we have normalized the user log data and fitted the scores to a complete set of indicators for two separate one-month time windows. We choose two separate time periods that were 16 months apart in order to emphasize the stability of certain indicators, such as unique users, download frequency and homepage hits.

	ds540.0-1: 3/2011	ds540.0-1: 7/2012	ds083.2: 03/2011	ds083.2: 05/2012	ds093.0-3: 3/2011
Unique users	46	45	987	976	88
Download frequency	264	373	374962	335422	3528
Files DS	433	473	22221	25504	195616
Homepage hits	685	588	6749	6907	1655
Subset requests	145	35	n/a	42	175
Download density	5.73913043	8.28888888	379.900709	343.67008	40.0909090
Usage impact	0.60969976	0.78858351	16.8736485	13.151740	0.01803528
Interest impact	1.58198614	1.24312896	0.30371145	0.2708202	0.00846043
Download ratio	2.59469697	1.57640750	0.01799915	0.0205919	0.46910430
Usage balance	0.38540146	0.63435374	55.5581567	48.562617	2.13172205
Subset ratio	0.55	0.09383378	n/a	0.0001253	0.04960318
Datasets	2	2	1	1	3
Secondary interest impact	14.8913043	13.0666666	6.83789260	7.0768442	18.8068181

Table 3. Indicator scores for three datasets from the RDA.



Results

Ds083.2 is undoubtedly one of the most popular datasets in the RDA, as reflected by the number of unique users it attracted in both time intervals. Interestingly this dataset's importance in the RDA is also reflected in the usage impact and usage balance indicators. Users of this dataset have a much higher download density than either of the other two datasets, suggesting that users of ds083.2 show a high amount of interest in the dataset and access a large portion of the dataset per time window. This latter point is an important one: ds083.2 is exceptionally popular because it includes observations from the Global Forecast Systems (GFS) and its users are likely to systematically download new additions to the dataset on a regular basis.

Interestingly, ds093 increased greatly in popularity over the 16-month period of our observation (a trend that has continued). While it increased nominally in file size, the number of unique users and the usage balance tripled, and the download density more than doubled. This leads to a secondary, but nonetheless compelling value of the DUI indicators: they hold the potential to both compare impact across an archive, as well as track fluctuations of use, popularity and impact of an individual dataset over time. This has important implications for the amount of staff time that is devoted to curating this particular dataset, as it appears to have an expanding user base.

Ds540 is much smaller than the other two datasets in our case study and consequently its indicator scores were lower, which seems to indicate that it receives less attention in the archive as a whole. However, the secondary interest impact score of ds540 is quite high – indicating that it is of very high interest to a small number of repeated users. One explanation for this very high score is that the community of users for this dataset (which consists of historical observational data) is likely to be climate-model developers. Although there are very few climate model development projects in the world, their work has an enormous impact on the field of climate science overall. Thus, in the case of ds540, the secondary impact score indicates that there is an additional value of this dataset that is not well represented by the index as a whole. We'll return to the issue of size as a function of attention later in this paper, but we do recognize the need for a weighting scheme that can smooth the effect of size on metrics developed for archives hosting datasets that vary in volume.

Discussion

Data usage indicators typify how data are discovered or accessed, and we believe that the DUI as a whole can give curators valuable insight regarding the impact of data on a community of users. Over time, we also believe these indicators can be useful tools for understanding which datasets within an archive would most benefit from additional curation efforts.

Creating new impact assessments can also provide the opportunity to make curation work more visible to two particular stakeholders. Firstly, indicators that signal the impact of a dataset can be used to illustrate the value of a repository to research funding agencies on behalf of a data producer. Secondly, and of equal importance, impact indicators can be used internally by repository staff to assess the effectiveness and value of their own services, systems and workflows. In combination, these indicators can inform the ways that a particular piece of architecture should be

Acknowledgements

This research has been made possible through the support of the Institute for Museum and Library Services (Award Number: RE-02-10-004-10) Data Curation Education in Research Centers. The first author was supported by the Advanced Study Program at NCAR. NCAR is funded by the National Science Foundation.

References

- Bollen, J., Van de Sompel, H., Smith, J.A. & Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing & Management*, 41(6), 1419-1440. [doi:10.1016/j.ipm.2005.03.024](https://doi.org/10.1016/j.ipm.2005.03.024)
- Bollen, J. & Van de Sompel, H. (2008). Usage impact factor: The effects of sample characteristics on usage-based impact metrics. *Journal of the American Society of Information Science*, 59. [doi:10.1002/asi.20746](https://doi.org/10.1002/asi.20746)
- Bollen J., Van de Sompel, H., Hagberg, A. & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE* 4(6). [doi:10.1371/journal.pone.0006022](https://doi.org/10.1371/journal.pone.0006022)
- Haustein, S. & Siebenlist, T. (2011). Applying social bookmarking data to evaluate journal usage. *Journal of Informetrics*, 5(3), 446-457. [doi:10.1016/j.joi.2011.04.002](https://doi.org/10.1016/j.joi.2011.04.002)
- Haythornthwaite, C. (2006). Articulating divides in distributed knowledge. *Information, Communication & Society*, 9, 761-780. [doi:10.1080/13691180601064113](https://doi.org/10.1080/13691180601064113)
- Henneken, E.A., Kurtz, M.J., Accomazzi, A., Grant, C.S., Thompson, D., Bohlen, E. & Murray, S.S. (2009). Use of astronomical literature: A report on usage patterns. *Journal of Informetrics*, 3(1), 1-8. [doi:10.1016/j.joi.2008.10.001](https://doi.org/10.1016/j.joi.2008.10.001)
- Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569-16572. [doi:10.1073/pnas.0507655102](https://doi.org/10.1073/pnas.0507655102)
- Ingwersen, P. & Chavan, V. (2011). Indicators for the Data Usage Index (DUI): An incentive for publishing primary biodiversity data through global information infrastructure. *BMC bioinformatics*, 12 (Supplement 15), S3. [doi:10.1186/1471-2105-12-S15-S3](https://doi.org/10.1186/1471-2105-12-S15-S3)
- Jacobs, C. A. & Worley, S. J. (2009). Data curation in climate and weather: Transforming our ability to improve predictions through global knowledge sharing. *International Journal of Digital Curation*, 4(2), 68-79. [doi:10.2218/ijdc.v4i2.94](https://doi.org/10.2218/ijdc.v4i2.94)

