Lead, Lag or Get Out of the Index: Exploring Macro-economic Indicators of Data Use

Nicholas M. Weber
nmweber@illinois.edu

Andrea K. Thomer
thomer2@illinois.edu

Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Abstract

Developing techniques to better quantify, track and more accurately describe the impact of federally funded research is quickly becoming a reputable domain for information studies, including data curation. In previous papers we've suggested the adaptation of an existing Data Use Index to quantify data use in the Research Data Archive at the National Center for Atmospheric Research. In this poster we revisit those indicators to determine their ability to forecast or indicate changes in data use over time. Central to our exploration of these indicators is an economic approach to quantifying data use, which holds that patterns in data repository events (downloads, searches, browsing etc.) should be capable of both predicting and explaining variations in useover time. We present preliminary results from this analysis and conclude with some prospects for future work with macroeconomic indicators.

Keywords: data curation, infometrics, scholarly communications, evidence-based policy

Introduction

The transformation of digital datasets from the 'underlying' research material supporting formal journal publications, to first-class, citable and shareable research products has been a gradual process across the social and natural sciences. In recent months this process has sped up considerably, as funding agencies like NSF (Gutman, 2012) have announced that grant applications and policy reports will soon be required cite to research 'products' instead of simply 'journal publications'. This slight shift in language has the potential to profoundly impact the ways we quantify the productivity, impact and usefulness of research funded at a federal level (Lane, 2010; Mayernik, 2012; Parsons, Duerr, Minster, 2010).

Not Metrics of Use, But Indicators of Use

Measuring and tracking the use of datasets still remains largely dependent on the techniques of citation-based bibliometrics, but increasingly there are efforts to diversify both the types of scholarly materials that we measure (Priem and Hemminger, 2010) and the values we assign to such measurements (Piwowar, Carlson and Vision, 2012). In previous work (Weber et al., 2013), we proposed the construction of metrics that 'indicate' data use: Different from traditional measurements of direct citations or acknowledgements, these metrics are constructed from download, browse and search events in a data archive (see table 1). Combined, these indicators make up a Data Use Index (DUI) that allows for a more holistic understanding of how data are used and what impact they have on a community of researchers served by a data archive (Ingwersen and Chavan, 2011). In this analysis we've chosen to explore data use indicators in the Research Data Archive (RDA), at the National Center for Atmospheric Research (for a more thorough discussion of the RDA see Jacobs and Worley, 2009)

Acknowledgements: Thanks to Matthew Mayernik, Steve Worley, Robert Dattore and Zaihua Ji. This research was funded through the DCERC program, IMLS award # RE-02-10-0004-10.

Weber, N., & Thomer, A. (2013). Lead, lag or get out of the index: Exploring macro-economic indicators of data use. *iConference 2013 Proceedings* (pp. 949-952). doi:10.9776/13487

Table 1
Twelve indicators making up a Data Use Index for the RDA.

	RDA: **Indicator of use**	**Explanation**	**Type of indicator**
1	Unique Users (UU)	Unique users that downloaded data per time window	Coincident
1a	Unique Users - Advanced	UUs that accessed data programmatically	Coincident
1b	Unique Users - Assisted	UUs that accessed data via GUI or Service	Coincident
2	Number of Datasets	Number of Datasets assigned DS number	Coincident
3	Files DS	Number of files in Dataset per time window	Coincident
2	Download Frequency	Total number of files downloaded per time window	Leading
2a	Download Frequency - Advanced	Files downloaded by Advanced users	Leading
2b	Download Frequency - Assisted	Files downloaded by Assisted users	Leading
4	Homepage Hits	Dataset Homepage Hits per time window	Leading
4a	Homepage Hits - Direct Access	Dataset Homepage Hits per time window by users with direct access (link not indexed or retrieved by search)	Leading
4b	Homepage Hits - With Link	Dataset Homepage Hits per time window by users with link (from indexed list or retrieved by search)	Leading
5	Subset Requests	Subsets Requests per time window	Leading
6	Download Density	Average number of files downloaded per UU	Lagging
7	Usage Impact	Total number of downloaded files over total files in dataset	Lagging
7a	Usage Impact - Advanced	“	Lagging
7b	Usage Impact - Assisted	“	Lagging
8	Interest Impact	Total homepage hits per number of files in dataset	Lagging
9	Usage Balance	Files downloaded by number of homepage hits per time window	Lagging
10	Subset Ratio	Number of subset requests over total number files downloaded per time window	Lagging
12	Secondary Interest Impact	Homepage over UU	Lagging

The Economies of Data Repositories

Vertesi and Dourish first introduced the concept of data economies (2011) to information studies, and we later used this economic lens to explore data use between and within sub-disciplines of researchers involved in Earth Systems Science (Weber et al., 2012). In combining an economic perspective of data work and metric based indicators, we want to ask:

- Can we use macro-economic indicators to predict future data use in an archive setting?
- Can we explain patterns in data use which vary from temporal norms?
- Are there data use events that we’ve not collected that could better predict use and access patterns over time?

Macroeconomic Indicators

In traditional macro-economic analysis, there are many types of indicators (e.g. Performance, Technical, Directional or Temporal etc.), but most common for investigating broad, temporal trends are the rather recursively named 'Economic Indicators' (Moore, 1983). For this analysis, we've chosen to analyze search and download events from datasets hosted by the RDA using the three most common temporal Economic Indicators:

Leading indicators are used to forecast how slight changes in patterns of use or disruptions / enhancement in access to data may in fact foreshadow larger shifts in the 'economy' of the data archive.

Lagging indicators are usually post-hoc calculations that attempt to correlate metrics with how or why a certain event happened the way that it did. When used comparatively with leading indicators, lagging indicators can substantiate or refute predicted changes in an economy.

Coincident indicators give a snapshot of the here and now of an economy. Traditionally coincident indicators are representative of a current economic state not yet affected by leading indicators, and not yet represented by lagging indicators. In this sense, the leading and lagging indicators give a sense of certainty to coincident indicators.

Analysis of RDA Indicators

Method

We first categorized our 12 usage indicators according to their corresponding economic indicators (see table 1, column 4). We expect that discovery- and access-related events will act as leading indicators; simple data such as the number of registered users for the archive, or the amount of files within a given dataset will act as coincident indicators; and lagging indicators will combine these leading and coincident measures to explain patterns of use, and measure the relative impact of a dataset from the RDA.

We then calculated these 12 usage indicators for three of the most heavily used datasets hosted by the RDA; these also represent the diversity of the RDA's holdings, as they include various types of climate data, such as model output data, reanalysis data, and observational data collected by field campaigns. It's important to note that while each of these datasets are assigned *one* unique identifying number (e.g. ds083.2), they are composed of numerous files, which users often subset or download in various combinations. After calculating the indicators listed in table 1, we then graphed leading, lagging and co-incident indicators for each dataset to explore shifts or variations in data use in the RDA.

Preliminary Results

Some of our most interesting results have been within our leading indicators: Increases in download frequency by type of user access often has relationship with total amount of data downloaded per month. So, months in which programmatic (advanced) user access rises we see a much higher overall download frequency- whereas increases or decreases in assisted user access seems to negligibly affect the total download frequency.

This seemed puzzling given that we expected assisted user access, which is aided greatly by the data curators at NCAR, to positively effect the overall amount of data consumed by archive users. However, our leading and co-incident indicators revealed a an interesting relationship between the number of subset requests and advanced user download rates. As sub-set requests increased the overall download frequency decreased. So, the implication might be that user services (like sub-setting) actually lead to more selective data use, and overall that assisted access might actually decrease the total amount of data downloaded in a given month. One conclusion we might draw from this analysis is that data archives promoting their success using a raw metric such as the total amount of data downloaded per month may actually be confusing the inefficiency of their architecture, with their impact on a user community.

Future Work

Our analysis is ongoing, but future work will include a more thorough analysis of the data available from the RDA, as well as an exploration of the predictability of these indicators in the form of lagging indicators. This work is also a first attempt at understanding use and impact amongst federally funded research data. We expect that as more datasets and archives are analyzed, these indicators will become more stable, and better able to accommodate an economic modeling of the impact of research data.

References

Gutmann, M. (2012, September). Academic Data: A Funder's Perspective. Presentation at the 2012 Wolfram Data Summit. Washington, D.C.

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. doi: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.

Lane, J. (2010). Let's make science metrics more scientific. Nature, 464, 7288, 488-489. doi:10.1038/464488a

Moore, G (1983) Why the Leading Indicators Really Do Lead. Business Cycles, Inflation, and Forecasting, 2nd ed. p. 339 - 352. http://www.nber.org/chapters/c0707

Mayernik, M.S. (2012). Data Citations: Initiatives, Issues, and First Steps. Bulletin of American Society for Information Science and Technology, 8(5): 23-28. http://www.asis.org/Bulletin/Jun-12/JunJul12_MayernikDataCitation.pdf

Parsons, M.A., Duerr, R., & Minster, J. B. (2010). Data Citation and Peer Review. Eos Transactions, AGU, 91(34). doi:10.1029/2010EO340001

Piwowar, H., Carlson, J., Vision, T. (2012). Beginning to track 1000 datasets from public repositories into the published literature. Proceedings of the American Society for Information Science and Technology. 48(1), 1-4. DOI: 10.1002/meet.2011.14504801337

Priem, J., & Hemminger, B. M. (2010). Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. First Monday, 15, 7. http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2874/2570

Vertesi, J., & Dourish, P. (2011). The value of data: considering the context of production in data economies. Proceedings of the ACM 2011 conference on Computer supported cooperative work, 533–542.

Weber, N., Baker, K., Thomer, A., Chao, T., Palmer, C. (2012). Value and context in data use: Domain analysis revisited. Proceedings of the 75th Annual Meeting of the American Society for Information Science and Technology. Baltimore, MD.

Weber, N., Thomer, A., Mayernik, M., Dattore, R., Zaihua, J., Worley, S. (2013). The product and system specificities of measuring impact: Indicators of data use in research data archives. Proceedings of the 8th International Digital Curation Conference. Amsterdam.