COLLABORATIVE SCIENCE

SOLVING THE ISSUES OF DISCOVERY, ATTRIBUTION AND MEASUREMENT IN DATA SHARING

ESSAY

OCTOBER 2012

*Scientific breakthroughs will be powered by advanced computing techniques that help researchers explore and mine datasets. Digital data are both the products of research and the foundation for new scientific insights and discoveries that drive innovation. **

JEANNETTE WING, ASSISTANT DIRECTOR
NATIONAL SCIENCE FOUNDATION
COMPUTER & INFORMATION SCIENCE & ENGINEERING DIRECTORATE

EXECUTIVE SUMMARY

Twenty-first century research is more data-intensive than ever due to the proliferation of digital technologies and the demand for answers in today's era of fast-paced innovation. Similarly, the movement toward collaborative (aka "open") innovation is affecting scientific research, bringing scientists from different disciplines together in their pursuit of solutions to today's challenges. In an effort to make research data more transparent and ensure that "any data obtained with federal funds be accessible to the general public,"* funding organizations and professional associations are now requiring researchers to provide data management plans when filing for grants and to make the results of their work available to the larger research community. The benefits are obvious: data can be used by other researchers with different objectives, results can be reproduced more easily and accurately, researchers receive the credit they're due, and data producers have a new channel by which to promote their work.

Historically, however, there have been disparate repository systems of varying levels of credibility for housing datasets, making it extremely difficult to discover, attribute, and measure research, and at the same time offering little incentive for researchers to submit their findings for others' use.

Knowing the growing importance of data sharing, Thomson Reuters Scientific & Scholarly Research analysts researched and studied the situation in search of a better solution. They determined a system was needed to help researchers more easily discover data relevant to their work, attribute data in a way that appropriately acknowledges intellectual debt, incorporate data usage in measurements of impact and influence, and contribute to the use of data citation information for future funding awards or even academic tenure and promotion decisions.

As part of the effort to address the problem, Thomson Reuters started working with prominent champions of data sharing, including leading research libraries and digital repositories. Because its Web of Knowledge℠ platform already indexes peer-reviewed literature in journals, books and conference proceedings, the project team sought to develop a way to not only connect disparate datasets and allow for more unified searching across this emerging scholarly landscape, but also to connect and display that information alongside more traditional literature indexed in Thomson Reuters databases. The result was the Data Citation Index which the company launched on the Web of Knowledge platform in 2012 as a single source for dataset discovery that also places datasets in context of attribution and usage metrics within the broader research landscape.

* National Science Foundation (May 10, 2010): Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans;
http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928&org=NSF&from=news

*Researchers from numerous disciplines need to work together to attack complex problems; openly sharing data will pave the way for researchers to communicate and collaborate more effectively. **

ED SEIDEL, ACTING ASSISTANT DIRECTOR
NATIONAL SCIENCE FOUNDATION
MATHEMATICAL AND PHYSICAL
SCIENCES DIRECTORATE

THE PUSH TOWARD DATA SHARING

In 2005, the National Science Foundation (NSF) in the United States published a manifesto on the importance of collecting and curating datasets for ongoing use. Six years later, the NSF was the first organization to mandate that all funding proposals include a two-page "data management plan" describing how the project will conform to NSF policy on sharing of datasets. With data management a growing priority for NSF, many funders and organizations followed suit. Proposal requirements were revised to include a data sharing component, increasing the transparency and reproducibility of funded or published projects.

Similar efforts are underway around the world. In the United Kingdom, the Economic and Social Research Council (ESRC) and National Environment Research Council (NERC) have begun implementing requirements for data management plans, along with the European Commission FP7, Wellcome Trust (UK), and the Australian National Data Service (ANDS) and Research Council (ARC).

Data sharing has increasingly become a global imperative due to the changing nature of research and available data. In its 2005 paper, NSF defines data as "any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations…generated by various means including observation, computation or experiment." The proliferation of digital technology allows data to come in forms unthinkable even 20 years ago, increasing data volume and variety. Although data has always been a part of research, these new forms make research increasingly data-intensive.

BENEFITS OF DATA SHARING

In a world where the grassroots open-source technology movement has spawned increasing levels of digital altruism, data sharing offers many benefits to the research community.

- **Advancing scholarship.** Using existing data as the basis of breakthrough ideas is nothing new. Four hundred years ago, Johannes Kepler, assistant to Danish astronomer Tycho Brahe, discovered the laws of planetary motion by poring over Brahe's carefully cataloged observations. This is just one example of the compounding effect of scientific research. What is new, however, is the unprecedented ability to share data with potentially hundreds or thousands of researchers globally for the entire lifecycle of the data.
- **Increasing transparency.** With access to original datasets, researchers can verify results more readily and accurately. Transparency also allows researchers to understand how data was compiled and how conclusions were drawn, making the ability to determine if a particular dataset is an appropriate source for secondary or tertiary research much easier.
- **Promoting work in new ways.** As data sharing increases, datasets widely used by other researchers can potentially indicate separate contributions to the research community outside of the article or more traditional publication that reviews the findings. This could eventually reshape the way funders and researchers approach the scope and structure of data collection for a given investigation, while providing researchers a new avenue in which to promote their work.
- **Curbing the double-funding tide.** From a funder's perspective, one of the most frustrating situations is learning that one's organization has funded the same or a very similar initiative as another group (or even at one's own organization), yet in isolation—resulting in unwanted replication of efforts rather than building off or refining prior work. With a streamlined, authoritative source for datasets and digital citation, funders can see what others have funded to avoid "double-funding" the same project. By the same token, funders can ensure that work they have funded is properly attributed to them and publically available for researchers to access and build upon.

"Data is not like a narrative—you don't know its value and utility until it's used. In this regard, we're doing everything we can to promote sharing of data from our archives. The Data Citation Index from Thomson Reuters is helping to turn the tide."

JOHN KUNZE, ASSOCIATE DIRECTOR
CALIFORNIA DIGITAL LIBRARY

DISCONNECTED RESOURCES

Although the sharing of this form of data offers many advantages, such datasets have been largely underleveraged by the community due to the lack of discovery, attribution and usage measurement associated with the data.

Datasets live around the world in disparate repositories, many of which organically grew over the past decades to fit the needs of different disciplines. As a result, not all repositories are structured the same way and, indeed, they may vary significantly by (or within) disciplines or even type of study. Taken together, these varying networks frustrate a streamlined discovery process. Moreover, dataset discovery lacks the key context routinely available in more traditional literature such as articles, books, and proceedings—most notably the citation connections: Did the work influence others? Who, how many, and at what time? And how exactly did they draw upon the data in their work?

Although the organic growth of repositories allowed for great innovations and approaches to data management, the variance also led to the "silo" environment of disconnected resources. Specific ramifications of this included:

- **Inability to discover relevant datasets.** Because data repositories have been siloed, a researcher conducting a particular search would have to query multiple repositories in search of relevant data. Lacking a single source to identify datasets across multiple jurisdictions, data often go undetected, and researchers have had little hope that their contributions would be discovered.
- **Lack of standard attribution conventions.** Because data are a relatively new addition to a researcher's source list, there is no current standard for citing the information. Researchers referencing data consequently struggle with the issue of appropriately acknowledging intellectual debt.
- **Difficulty in measuring data use and impact.** Lacking citation conventions and a discovery tool with the ability to search multiple repositories has meant that tracking and measuring data usage and impact was a somewhat dance in the dark. Although dataset downloads and other alternate metrics are sometimes available, they are housed within the siloed repositories rather than one central resource. Moreover, even in new forms of digital scholarship, citations remain the strongest "credit" authors can give to honor the intellectual debt owed to their peers, offering the most reliable linkages on which to measure impact.
- **Challenges in seeing the full funding picture.** Funders have similarly been impacted by the siloed repository environment by not having a clear view into the results or findings from the research they support and the subsequent research to which an initial grant has contributed. Their challenge has more to do with where their money is going and the impact it has on science, which is of equal concern.

HISTORICAL LACK OF INCENTIVE

Submitting data to a repository has been a time-consuming process with questionable, or at least largely unrealized, benefit to date. Repositories require researchers to submit metadata, and in most cases researchers must "clean" the data by preparing or formatting it so that it can be used by others.

Additionally, even those who have understood the impact of accessible data and are willing to invest the time to prepare it for a repository have encountered a second layer of discouragement: the same lack of discovery, attribution and usage measurement that would not only aid their research but also open new avenues to promote the work they have produced in formal datasets.

The same discovery challenges have likewise made it difficult to incentivize researchers to submit their data. Thomson Reuters focused on solving these key issues as part of its initiative to contribute to more collaborative science and to offer a place and process for the proper attribution to researchers for their work.

DATA CITATION INDEX

Working with prominent champions of data sharing, including leading research libraries and digital repositories, such as the California Digital Library, ICPSR, Protein Data Bank, PANGAEA, and UK Data Archive, among others, Thomson Reuters developed the Data Citation Index. Designed to be a single source of data discovery for the sciences, social sciences, and arts and humanities, Data Citation Index fully indexes a significant number of the world's leading data repositories of critical interest to the scientific community, including over two million data studies and datasets. The records for the datasets, which include authors, institutions, keywords, citations and other metadata, are then connected to related peer-reviewed literature indexed in the Web of Knowledge.

By indexing repository data and displaying it alongside other research outputs, Data Citation Index solves three of the major issues that frustrate the discovery process and in turn discourage researchers from submitting their data to repositories.

- **Discovery.** Raw repository metadata and descriptive metadata feeds are analyzed and augmented during the indexing process. Once indexed, Data Citation Index records are integrated into the Web of Knowledge platform, making it a single, searchable platform for researchers looking for particular data. Searches performed incorporate all search-filtering capabilities available in the Web of Knowledge, refining the index's search results in more useful ways. Once a data record has been selected, it links directly to the data housed within its respective repository.
- **Attribution.** Each result page contains a *How to Cite this Resource* link with a recommended citation format. Providing standardized citations will help establish citation conventions, paving the way for better measurement of data impact over time and contributing to faster, more accurate indexing while also increasing search productivity. Thomson Reuters EndNote® supports the same citation standards offered in Web of Knowledge for streamlined, consistent attribution of datasets during the writing process.
- **Measurement.** Integration with the Web of Knowledge and the resulting linkages from data to literature allow the Data Citation Index to present measures traditionally associated with published literature to datasets. Researchers can measure the influence of their data and publications through the *Times Cited* counter on each index record, creating a valuable representation of influence for future funding awards or even academic tenure and promotion decisions.

FUNDERS AND THE FUTURE OF DATA SHARING

The emphasis on data sharing promises only to increase in the next decade.

Infrastructures are being built to house and share data for researchers, educators and the general public. In 2008, the National Science Foundation funded over $95 million in data preservation and infrastructure development projects under the DataNet initiative, including the Data Observation Network for Earth (DataONE) and the Data Conservancy headquartered at Johns Hopkins. Additionally, the Australian National Data Service (ANDS) and Australian Research Data Commons are pioneering national initiatives, and organizations such as European Data Infrastructure (EUDAT) are working to create an electronic data-sharing infrastructure supporting all of Europe.

As data sharing and data management become important prerequisites for funders, the pressure on researchers to share their data will increase. However, researchers should be given opportunities to understand the contributions their data are making in their field and to the research community at large. In this way, the pull of funders to make data available will be well met by the push of researchers eager to uncover new ways of measuring the value of their work and tracing its footprint to find additional data sources, articles, or even collaborators to accelerate the time to create new work.

For more information, visit:
go.thomsonreuters.com/datacitationindex

Science Head Offices:

Americas
Philadelphia +1 800 333 4474
+1 215 386 0100

Europe, Middle East, Africa
London +44 20 7433 4000

Asia-Pacific
Singapore +65 6775 5088
Tokyo +81 3 5218 6500

For a complete office list, visit:
science.thomsonreuters.com/contact

10/2012

THOMSON REUTERS