REUTERS / Yuriko Nakao

UNLOCKING THE VALUE OF RESEARCH DATA

A REPORT FROM THE THOMSON REUTERS INDUSTRY FORUM
JULY 2013

THOMSON REUTERS™

“There is tremendous opportunity regarding data-repository curation, but we need to develop a business model.”

– Gerry Grenier, IEEE

With millions of scientists and scholars around the world busily pursuing their research, the accumulation of data is vast and growing. Digital technology facilitates the generation and storage of myriad forms of data generated in the course of research, even beyond anything that might necessarily wind up in a published paper. This material includes text, images, video, audio, graphs, animations, and much more. A recent estimate projected that, through 2020, the volume of data will grow by a factor of 44, from 0.8 zetabytes (or ZB, indicating 1 trillion gigabytes) as of 2009 to some 35 ZB at the end of this decade.

This continuous torrent feeds an ever-expanding, virtual reservoir—one that is not stored uniformly in one place but in various formats in scattered, disparate repositories of varying size across the globe.

Although only a fraction of such data is intended or destined for publication in journals, the potential utility of these separate, disconnected data stores is profound. Any given tidbit of data produced by one researcher might supply a missing puzzle piece to another—even one involved in seemingly unrelated work. The ability to harness such data and apply it in new ways and new directions holds the promise of substantially accelerating research and innovation.

The challenge, of course, lies in taming this great volume of research—in imposing uniformity and quality control, in providing universal access, and in maintaining proprietary rights and due credit for contributing researchers, among many other issues. These matters constitute just part of the changing landscape confronting publishers, funders, and other stakeholders, as research is increasingly disseminated through channels outside the traditional conduit of the peer-reviewed, subscription journal.

In short, the current unavailability of this data constitutes an impediment to the progress of research. The availability and accessibility of the material—properly maintained and curated—is essential for advancing science.

THE THOMSON REUTERS INDUSTRY FORUM

To address these and other evolving issues, Thomson Reuters has convened its Industry Forum, to serve as a platform for leaders in the scientific, technical, and medical (STM) publishing to discuss strategic trends, technological developments, and paradigm shifts in the field of scholarly communications, and to recommend industry initiatives in support of the attendant opportunities. The forum met on April 29 in Washington, DC, to discuss research data and other pressing matters.

CURRENT EFFORTS TO UNLOCK RESEARCH DATA

Recent years have seen the emergence of various initiatives that focus on advancing the availability and accessibility of research data. Examples include projects in astronomy, such as the publicly released data from the Sloan Digital Sky Survey (which, in turn, contributed to a "citizen science" project known as the Galaxy Zoo, in which members of the public assist in morphologically classifying thousands of newly observed galaxies recorded by the SDSS and other missions). In the life sciences, the Protein Data Bank stores and makes accessible extensive data on nucleic acids and other macromolecular structures. In neuroscience, the National Database for Autism Research collects and standardizes a range of data for general use.

In addition to these largely public projects, a number of commercial enterprises have emerged. One such company, figshare (http://figshare.com), was founded by Mark Hahnel (now a member of the Thomson Reuters Industry Forum), who first encountered the matter of proliferating data while a doctoral candidate in the life sciences, generating his own expanding store of research materials. The company represents his desire to make the totality of research output, as he puts it, "citable, sharable, and discoverable." Another pertinent tool arrived in 2012, when Thomson Reuters released the Data Citation Index, designed to afford access, along with citation metrics and other bibliometric measures, to datasets and repositories that currently lie outside the population of conventionally indexed materials.

An international group, the Research Data Alliance (RDA), emerged in 2013, with the stated mission of fostering efforts to "develop and adopt common tools and infrastructure, harmonize data standards, and apply policy and best practices" in the sharing and exchange of research data. As described by the RDA's US chairperson, Francine Berman of Rensselaer Polytechnic Institute, the RDA is an international community-driven organization focused on building out the social, organizational and technical infrastructure to accelerate research data sharing and exchange. In less than a year, the organization has attracted 700 members in 44 countries.

"The focus of the RDA is impact and implementation," says Berman, "with the deliverables of its 'Tiger Team'-style Working Groups following the approach of *create-adopt-use*."

One of the documents informing the RDA's activities is a 2010 report to the European Commission from the High-Level Group on Scientific Data, "Riding the Wave: How Europe can gain from the rising tide of data."[1] Among other recommendations, the report urges the formation of a Collaborative Data Infrastructure that would answer the requirement of being, as the report notes, "flexible but reliable, secure yet open, local and global, affordable yet high-performance. There is no one technology that can achieve it all. So we need a broad conceptual framework for how different companies, institutes, universities, governments and individuals would interact with the system."

1. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

A June 2012 report from the Royal Society, "Science as an Open Enterprise,"[2] also examines the pertinent issues and offers recommendations for the exchange of research data. From the report's summary: "Openness facilitates a systemic integrity that is conducive to early identification of error, malpractice and fraud, and therefore deters them. But this kind of transparency only works when openness meets standards of intelligibility and assessability—where there is intelligent openness."

Another body deliberating these issues is the ICSU (International Council of Science) Data Publication Working Group (of which Thomson Reuters is a member), overseeing the ongoing development of the World Data System, as is described at the organization's website: "WDS strives to form a worldwide '*community of excellence*' for multidisciplinary scientific data, which ensures the long-term stewardship and provision of quality-assessed data and data services to the international science community and other stakeholders. Its concept aims at a transition from existing stand-alone components and services to a common globally interoperable distributed data system, with searchable common data directories and catalogues that incorporates emerging technologies and new scientific data activities."[3]

STAKEHOLDERS

The lack of available, accessible, and curated research data is adversely affecting stakeholders across the STM industry, and not just in academia, where scholars are now primarily judged on what amounts to a mere portion of their research output. Users and consumers of content—whether fellow researchers, analysts, news organizations, or the broader public—are only receiving the tip of the research iceberg at this time. Funders in the public and private spheres have a clear stake in maximizing the value of the research they underwrite. Service providers are missing out on a large business opportunity. For example, potential opportunities also await providers of cloud-based storage in offering tools and services for access. As forum member Gerry Grenier, senior director of Publishing Technologies at IEEE, recently observed, "There is tremendous opportunity regarding data-repository curation, but we need to develop a business model."

2. http://royalsociety.org/policy/projects/science-public-enterprise/report/
3. http://www.icsu-wds.org/working-groups/data-publication

CURRENT GAPS AND CHALLENGES

As abundant as the potential opportunities, so are the obstacles to unlocking the value of research data.

First, there are the challenges of providing uniform access to a broad variety of research outputs—the technical and computational limitations in making the data available, searchable and retrievable.

Another fundamental question centers on the quality of the data. Who will undertake the curatorial task of filtering the good from the bad, particularly for material that has not been subject to conventional peer review? Or, to frame the question more bluntly, who will pay the costs associated with this curation and quality control?

Various questions and uncertainties also pertain to the researchers who generate the data. What are the incentives and benefits, such as recognition, for making their research public? As was observed by forum member Alex Wade, director of Scholarly Communication for Microsoft, in a blog entry for BioMedCentral in 2011,[4] "It is clear that the sharing of scientific research data holds great promise for the scientific discoveries of the future. And yet the system of academic research achievement does not yet recognize or reward researchers for sharing their data. Change is afoot, however, and the next generation will look back on this decade as one of profound transformation in determining which parts of the scientific research process are recorded and how researchers are rewarded."

The establishment of community standards around the sharing and reuse of research data are signaled by initiatives such as the Amsterdam Manifesto on Data Citation Principles[5].

Additionally, uncertainty persists among researchers regarding the legal or proprietary distinctions in terms of which shared data is licensed or copyrighted versus what becomes part of the public domain. These matters must be resolved.

For publishers, too, this new landscape presents a range of issues, further complicating the ongoing discussion regarding open access. For example, the financial implications: How will publishers who traditionally regard themselves as content providers negotiate the transition to a role focusing on service provision? How will businesses and systems accustomed to a pay-to-read model adapt to a pay-to-publish model?

Research assessment is also likely to face changes—both technological, as new metrics and indicators are applied to different kinds of research outputs, and cultural, as traditional touchstones or measures such as journal names and impact factor are challenged. The recent San Francisco Declaration on Research Assessment (DORA)[6] and its supporters embody one manifestation of this impetus for the reform of research assessment.

In all, new relationships, new dynamics, and new cultural behaviors are certain to be part of this changing world. But if any group seems likely to gain ascendancy, it is the authors themselves. As Peter Suber of Earlham College recently noted in Nature[7] in discussing open access (OA), "Of all the groups that want OA to scientific and scholarly research literature, only one is in a position to deliver it: authors....[A]uthors have primacy in the campaign for OA, and the single largest obstacle to OA is author inertia or omission."

4. http://blogs.biomedcentral.com/bmcblog/2011/05/17/paradigm-shifting-in-scholarly-communications/
5. http://www.force11.org/amsterdammanifesto
6. http://am.ascb.org/dora/
7. http://www.nature.com/nature/focus/accessdebate/24.html

RECOMMENDATIONS

Members of the Thomson Reuters Industry Forum agree that with a lack of an appropriate reward system in academia, it will be a long and hard battle to change the current scholarly publication process which is primarily catering to published journal articles. But why? As a first step, academia, publishers, and other service providers could advance the mechanisms for publishing research data with or without a research article, in a way and a format that allows scholars to use the data in the most productive way while giving credit to the original researcher. One idea discussed by the forum is a consortium of major publishers and associations—such as AAAS, Elsevier, Wiley, the Nature Publishing Group, IEEE, and selected major journals—to create standards by which to guide a transition toward new modes of communicating and sharing data.

Clearly, managing the process of unlocking research data will require cooperation and partnership; no one entity—governmental, academic, commercial—can be the sole driver. And communication will be key. It is our hope that the Thomson Reuters forum will provide a focal point for these exchanges, as we examine not only these aspects of research data but also the concurrent changes in the scholarly workflow, the enrichment of content, and other issues. There is much to discuss.

FORUM MEMBERS

Gerry Grenier
Sr. Director, Publishing Technologies, IEEE

William Gunn
Head of Academic Outreach, Mendeley

Mark Hahnel
Founder, figshare

Ian Mulvany
Head of Technology, eLife

Michael R. Nelson
Analyst, Technology & Internet Policy, Bloomberg Government

Michael O'Brien
Technology/Innovation Consultant

Aldo de Pape
Business Development Manager, Digital Science

Mark Patterson
Managing Executive Director, eLife

Jasper Simons
Vice-President, Product & Marketing Strategy, Thomson Reuters

Alex Wade
Director, Scholarly Communication, Microsoft

Mark Ware
Director, Mark Ware Consulting

ABOUT THOMSON REUTERS

Thomson Reuters is the leading source of intelligent information for professionals around the world. Our customers are knowledge workers in key sectors of the global economy. We supply them with the intelligent information they need to succeed in fields that are vital to developed and emerging economies such as law, financial services, tax and accounting, healthcare, science and media.

Our knowledge and information is essential for drug companies to discover new drugs and get them to market faster, for researchers to find relevant papers and know what's newly published in their subject, and for businesses to optimize their intellectual property and find competitive intelligence.

NOTE TO PRESS:

To request further information or permission to reproduce content from this report, please contact:

Laura Gaze
Phone: +1 203 868 3340
Email: laura.gaze@thomsonreuters.com

THOMSON REUTERS
REGIONAL OFFICES

North America
Philadelphia +1 800 336 4474
+1 215 386 0100

Latin America
Brazil +55 11 8370 9845
Other countries +1 215 823 5674

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

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

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

1003903

THOMSON REUTERS™