the institute (/)   The IEEE news source

Sign up for newsletter (/static/email-alert)

TECH TOPICS (/TECHNOLOGY-TOPICS)   TECH HISTORY (/TECH-HISTORY)   BLOG (/BLOGS)   SPECIAL REPORTS (/SPECIAL-REPORTS)

MEMBERS (/MEMBERS)   CAREER & EDUCATION (/CAREER-AND-EDUCATION)   RESOURCES (/RESOURCES)   🔍 (#)

Despite Growing Data, Infrastructure Stands Still

Why the gap puts research data at risk

By FRANCINE BERMAN 16 September 2014

Image: iStockphoto

This article is part of our September 2014 special report on **big data (/static/special-report-big-data)**, *covering technologies that support and make sense of the growing mountains of data, and several of its applications.*

Data increasingly drives innovation in virtually every area of inquiry. Whether the data helps to reveal the existence of the "**God particle (http://www.nytimes.com/2012/07/05/science/cern-physicists-may-have-discovered-higgs-boson-particle.html?pagewanted=all)**," the **discovery of a new planet (http://www.nasa.gov/content/nasa-kepler-results-usher-in-a-new-era-of-astronomy/)**, the **behavior of crowds (http://www.technologyreview.com/view/516241/music-festivals-bluetooth-monitoring-and-the-behavior-of-crowds/)**, or the **spread of disease (https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4493/49)**, it is key to discovery and innovation. Data is also a national priority around the world. In the United States, White House initiatives are focusing on **public access to research data (http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf)**, **big data**

(http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf), and **government open data (http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-)**.

Yet as data—big, small, or in-between—becomes more valuable to science and society, its longevity may **increasingly be at risk (http://www.nature.com/news/scientists-losing-data-at-a-rapid-rate-1.14416)**. This is because there has been inadequate planning and support for the hardware, software, organizational, and human infrastructure that enables its access and use. In particular, we need infrastructure to support the *stewardship* of research data (for access and use now), and its *preservation* (for access and use in the future). Without sufficient fiscal, community, and organizational support, **research data may disappear (http://www.sciencedirect.com/science/article/pii/S0960982213014000)**, potentially never to be restored.

So why is it so hard to support the infrastructure needed for valuable research data? Why is research data infrastructure such a hard sell?

To try to understand the problem, we can contrast the environments of data-driven research versus the environment in which we build research data infrastructure. Consider the following questions:

WHAT IS NEWSWORTHY?

In the research arena, new discoveries and results are newsworthy, such as the discovery of the **Higgs Boson (http://www.sciencemag.org/content/338/6114/1524.full)** and more accurate models for **predicting earthquakes (http://www.scec.org/m8/)**. These provide new insights, expand our horizons, and may ultimately create opportunities for innovative products and services.

In contrast, infrastructure is primarily newsworthy when it *fails*. For example, when valuable data such as social security numbers are lost or damaged, or when data systems break down. In particular, what is newsworthy is the "bad news." Well-running, reliable infrastructure for data stewardship and preservation is expected, unremarkable, and not on the radar for recognition.

WHAT'S THE VALUE PROPOSITION?

Research drives the innovation needed to advance science and society; it is critical for success. This is an important value proposition that enables investment in research, even when the results are not immediately applicable. In contrast, infrastructure *enables* research and innovation. This is a much weaker value proposition, making it harder to prioritize **data infrastructure investments (http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf)** over higher-profile or more urgent efforts.

WHAT'S THE FUNDING MODEL?

Almost all of our research funding is time-limited: Government agencies and foundations typically provide funding for a set period of time (often 3 to 5 years with or without limited renewal), and long-term research programs are typically funded by a variety of sources. In particular, research funding is generally not "continuous" from a single grant, agency, or source and almost all researchers' funding pipelines experience ebbs and flows.

In contrast, infrastructure funding must be continuous and long-term with no gaps. Systems must keep running and administrators must migrate valuable data smoothly over time from old technologies to new ones. Gaps in funding may put valuable data at risk. If we don't pay the light bill, the lights go out. If we don't pay the data bill, we are at risk of losing our data.

WHO IS RESPONSIBLE?

All across the world, research agencies provide critical support for research communities. In the United States, researchers are funded by the National Science Foundation, National Institutes of Health, Department of Energy, National Endowment for the Humanities, and other agencies. Researchers are also supported by non-profit organizations as well as research funding programs from the private sector.

Who is responsible for funding research data infrastructure? Typically this responsibility is shared among research agencies, institutions, non-profits and others but often as a lesser priority than the mainstream mission of these organizations. There is no "National Research Infrastructure Foundation" with a mission to build and maintain infrastructure that enables research, particularly data-driven research. Moreover, in many research-funding organizations, support for infrastructure must compete with funding for research **without expansion of funding levels (http://www.sciencemag.org/content/341/6146/616.full.pdf?**

keytype=ref&siteid=sci&ijkey=.e2Ezowko%2FxF2). (A free copy of this report is available on my **website (http://www.cs.rpi.edu/%7Ebermaf)**.) This is particularly challenging for digital research data, which may require near-constant care to be accessible for current and future use.

DATA INFRASTRUCTURE MATTERS

With primarily bad news, a weaker value proposition, a more challenging business model, and no group whose mainstream mission it is to plan and coordinate the data infrastructure needed for the research community, it is no surprise that we do not have a comprehensive plan for research data stewardship and preservation.

Nonetheless, good planning and strategic investment are critical to avoid falling behind. With focus and commitment, a national research data stewardship and preservation plan can be developed and used to drive new discovery and innovation. Without such a plan, we may be left behind wondering what we could have done with data that has disappeared.

Francine Berman is an IEEE Fellow, Hamilton Distinguished Professor of Computer Science at Rensselaer Polytechnic Institute, and Chair of Research Data Alliance/ United States.

Acknowledgements
Many thanks to Myron Gutmann for thoughtful comments on an earlier draft of this piece.

Photo: Francine Berman

RELATED

[(/ieee-roundup/members/achievements/four-takeaways-from-ieee-conferences-on-big-data)](/ieee-roundup/members/achievements/four-takeaways-from-ieee-conferences-on-big-data)

Four Takeaways From IEEE Conferences on Big Data [(/ieee-roundup/members/achievements/four-takeaways-from-ieee-conferences-on-big-data)](/ieee-roundup/members/achievements/four-takeaways-from-ieee-conferences-on-big-data)
Topics include sensor networks, privacy, and the role of humans

[(/blogs/standards2/standards-that-support-big-data)](/blogs/standards2/standards-that-support-big-data)

Standards That Support Big Data [(/blogs/standards2/standards-that-support-big-data)](/blogs/standards2/standards-that-support-big-data)
Covering a variety of applications

LEARN MORE IEEE ROUNDUP (/TAG/IEEE+ROUNDUP)   BIG DATA (/TAG/BIG+DATA)

CONNECT WITH US: (HTTP://WWW.LINKEDIN.COM/GROUPS/IEEE-INSTITUTE-3968449?HOME=&GID=3968449)   (HTTPS://WWW.FACEBOOK.COM/IEEETHEINSTITUTE)
(HTTP://TWITTER.COM/#!/IEEEINSTITUTE)   (/STATIC/RSS)

 (http://www.ieee.org)