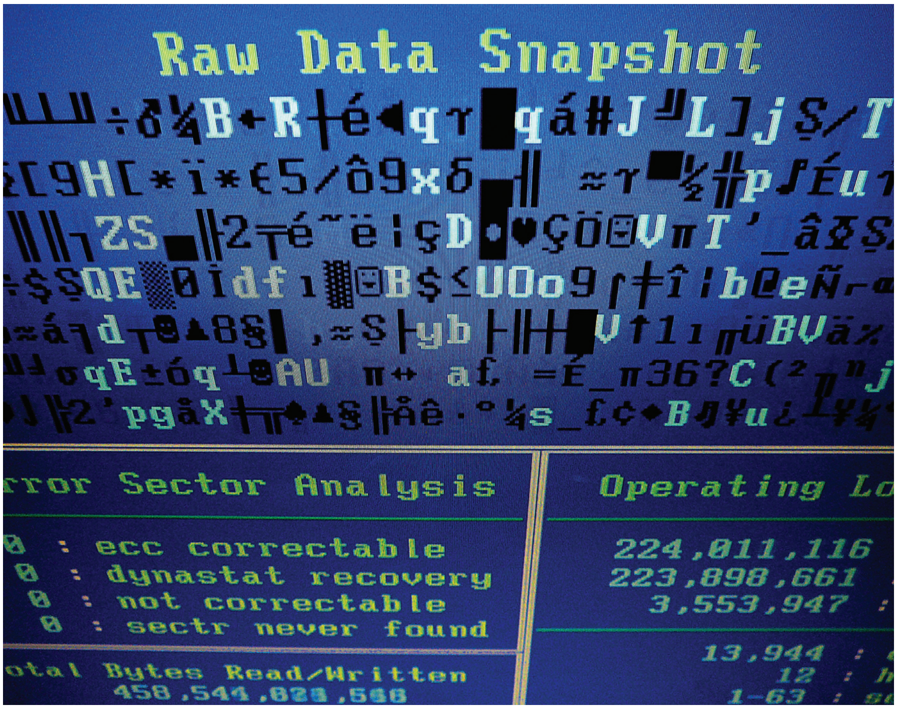
DOI:10.1145/1859204.1859220

Francine Berman

Viewpoint
We Need a Research Data Census

The increasing volume of research data highlights the need for reliable, cost-effective data storage and preservation at the national scale.



PHOTOGRAPH BY PAUL DINEEN

THIS PAST YEAR was a census year in the U.S. We responded to arguably the most long-lived and broad-based gathering of domiciliary information about the American public anywhere. U.S. Census data, collected every decade, provides a detailed picture of how many of us there are, where we live, and how we're distributed by age, gender, household, ethnic diversity, and other characteristics.

The Census (http://2010.census.gov/2010census/index.php) provides an evidence-based snapshot of America. This important information is publicly available and used in a variety of ways—to guide in the planning of senior centers, schools, bridges, and emergency services, to make assessments informed by societal trends and attributes, and to make predictions about future social and economic needs. The Census is particularly valuable as a planning tool in the building of physical infrastructure, as the distribution and characteristics of the population drive the development of hospitals, public works projects, and other essential facilities and services.

Given the role and importance of the Census in the physical world, it is useful to ask what provides an analogous evidence-based and publicly available snapshot of the "inhabitants" of the Digital World—our digital data.

What do we know about our data? How much is there? Where does it reside? What are its characteristics? Good "top-down" methodological estimates of these questions have come from the reports on the increasing deluge of digital information developed by the IDC (http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf), by Bohn and Short (http://hmi.ucsd.edu/pdf/HMI_2009_ConsumerReport_Dec9_2009.pdf), and (some time ago) by Lyman and Varian (http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf). These provide intriguing, analytically derived bounds of the Digital World.

However, to make economic decisions that can drive the cost-effective development and deployment of the cyberinfrastructure needed to support long-lived digital data, we need more resolution. This is particularly important in the research arena, where federal R&D agencies apportion funding between the competing priorities of conducting basic research, and creating and supporting the cyberinfrastructure that enables that research. Just as the U.S. Census drives planning for infrastructure in the physical world, a Research Data Census would inform cost-effective planning for stewardship of federally funded, shared cyberinfrastructure in the Digital World.

A Census for Research Data

The 10 questions on the 2010 U.S. Census form are well defined and provide basic information. There are many things not addressed in the Census—educational level of the population, for example. Similarly, an effective Research Data Census should provide basic information about the research data generated from federal funding. It should help us design, develop, and identify appropriately sized and outfitted storage, repositories, and services in the Digital World. It should provide a quantitative snapshot of the research data landscape at a given point in time, exposing key characteristics, such as:

▸ **Number and size distribution of federally funded research data sets.** How many research data sets generated by federal funding are less than a terabyte (that is, host-able on a researcher's hard drive), between 1 and 100 terabytes (perhaps host-able at a university repository), between 100 terabytes and a petabyte (perhaps requiring a larger-scale shared archive), more than a petabyte? What is their distribution?

▸ **Type and area distribution of federally funded research data sets.** What percentage of the U.S. federally funded research data is text, video, audio, and so forth? How much digital research data is generated within specific research areas (as categorized by NSF Directorates, NIH institutes, and other groups)?

▸ **Needs for preservation.** How much federally funded digital data must be retained by policy or regulation (HIPAA, OMB A-110, and so forth) for up to 1 year, 1–3 years, 3–5 years, 5–10 years, more than 10 years?

▸ **Common services and tools.** What categories of services and tools (gene sequence analysis, data visualization, mosaicing, and so forth) are used in conjunction with federally funded research data sets?

Note that basic questions along these lines will *not* provide a complete picture of our data. They do not differentiate between derived data and source data, for example; nor do they provide comprehensive information about the necessary data systems and environments required to support data.

A Research Data Census *will* provide some specifics critical to cost-effective planning for stewardship of federally funded research data, however, and it will allow us to infer some key requirements for data cyberinfrastructure. In particular, a Research Data Census could help inform:

- **Useful estimates of the storage capacity required for data stewardship, and a lower bound on data that must be preserved for future timeframes.** Data required by regulation or policy to be preserved is a lower bound on valued preservation-worthy research data—additional data sets will need to be preserved for research progress (for example, National Virtual Observatory data sets).
- **The types of data services most important for research efforts.** Knowing the most common types of useful services and tools can help drive academic and commercial efforts.
- **Estimates of the size, training, and skill sets that will be needed for today's and tomorrow's data work force.**

Getting It Done

A Data Census sounds like a big job and it is, however there is potential to use existing mechanisms to help gather the needed information efficiently. We already provide annual and final reports to federal R&D agencies to describe the results of sponsored research. One could imagine a straightforward addition to annual reporting vehicles and/or sites such as grants.gov to collect this information (preferably electronically). Although U.S. Census information is gathered every 10 years, the Research Data Census would require frequent updating in order to provide useful information for planning purposes about our dynamically changing data landscape. The right periodicity for reporting is a topic for discussion, but an annual update probably provides the best resolution for the purpose of tracking trends.

Note also that there is real complexity in doing an effective Data Census: much of our data is generated from collaborative research, which can cross institutional, agency, and national boundaries. The Data Census reporting mechanisms must take this into account to produce relatively accurate counts. Data sets are often replicated for preservation purposes—do we count the data in all copies (all of which require storage), or do we count only the non-replicated data? (It is interesting to note that the U.S. Census has a related problem and covers it as question 10: "Does person 1 sometimes live or stay somewhere else?" If yes, check all that apply....). As with any survey, careful design is critical in order to ensure the results are accurate and useful as the basis for making predictions and tracking trends.

An effective Research Data Census should provide a quantitative snapshot of the research data landscape at a given point in time.

Using the Research Data Census to Create Effective Data Stewardship

An important outcome of the Research Data Census would be evidence-based information on the amount of data in the research community that must be preserved over time. This would help in understanding and meeting our needs for archival services and community repositories.

Such information can help cut data management and preservation problems down to size. Knowing that data valued by a particular community is typically of a certain type, a certain size, and/or needed over a certain timeframe, can help the community plan for the effective stewardship of that data. For example, accurate estimates of the digital data emanating from the Large Hadron Collider at CERN have been instrumental in the development of a data analysis and management plan for the High Energy Physics community.

It is likely that some of the capacity needed for stewardship of research data will come from university libraries reinventing themselves to address 21st century information needs; some of the capacity may come from the commercial sector, which has responded to emerging needs for digital storage and preservation through the development of commercial services. In some cases, the federal government will take on the stewardship responsibilities for research data (for example, the NIST Science Reference Data). It is clear that the size, privacy, longevity, preservation, access, and other requirements for research data preclude a "one-size-fits-all" approach to creation of supporting data cyberinfrastructure. It is also true that no one sector will be able to take on the responsibility for stewardship of all research data. A national strategic partnership spanning distinct sectors and stakeholder communities is needed to effectively address the capacity, infrastructure, preservation, and privacy issues associated with the growing deluge of research data. The development of a Research Data Census can provide critical information for more effectively developing this partnership.

No Time Like the Present

The 2010 requirement for a data management plan at the National Science Foundation (http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928&org=NSF) joins existing requirements for data sharing and management at NIH and elsewhere. Such requirements expand community awareness about responsible digital data stewardship and will exacerbate the emerging need for reliable, cost-effective data storage and preservation at the national scale.

A Research Data Census will provide a foundation for estimating the data cyberinfrastructure required for strategic stewardship. It can lay the groundwork today for access to our most valuable digital research assets tomorrow, and the new discoveries and innovation they drive.

Francine Berman (bermaf@rpi.edu) is Vice President for Research at Rensselaer Polytechnic Institute, the former director of the San Diego Supercomputer Center, and the co-chair of the Blue Ribbon Task Force for Sustainable Digital Preservation and Access (http://brtf.sdsc.edu).