POLICY FORUM

DATA POLICY

Government data, commercial cloud: Will public access suffer?

New data handling models may affect use for research

By **Mariel Borowitz**

Many government agencies have adopted open data policies, making their data freely available to all users online. At the same time, many agencies are experiencing substantial increases in the volume of data that they generate or collect. Simply upgrading existing systems is often not sufficient to ensure that large volumes of data remain accessible and can be analyzed efficiently; instead, agencies are transitioning to cloud infrastructures. Whereas some agencies seek to develop new in-house cloud platforms, many others are turning to commercial cloud providers. However, the ways in which agencies are partnering with these commercial entities vary considerably, as does the distribution of costs among agencies, cloud providers, and users. In some cases, users may need to pay to work with government data that were previously freely available. Although the underlying data may remain free, agencies may allow commercial providers to charge fees to users to download or analyze data using their commercial platform. Such challenges to the implementation of open data policies must be addressed to ensure that the current benefits of these policies are not lost and to realize the opportunities for researchers and society presented by big data and cloud computing.

These challenges are faced by governments around the world. For example, the European Union is developing cloud-based systems to facilitate access to data from its Copernicus environmental monitoring program, and state governments in India are turning to commercial cloud providers to facilitate citizen services (*1*, *2*). Many of these efforts are in their early stages, and reasonable people disagree about the best way to develop and structure data distribution systems in the cloud. Those promoting commercial providers point to their well-developed platforms and ecosystems and substantial human and physical resources dedicated to maintaining and improving their cloud products. Agencies hope to leverage these resources to make data available quickly and affordably. To illustrate a variety of approaches being undertaken, consider developments in three U.S. government agencies with pilot programs in development or under way.

The National Oceanic and Atmospheric Administration (NOAA) houses more than 100 petabytes (PB) of environmental data and generates more than 30 PB a year from satellites, radars, computer models, and other sources. Because of limitations of traditional data portals, only about 10% of those data are accessible via NOAA websites (*3*). To address this concern, in 2015, NOAA implemented the Big Data Project, under which commercial cloud providers host NOAA data at no cost to the agency and make it available to the public. Although these providers cannot charge for the actual data, they are allowed to charge end users for distribution of the data and for processing and applications (*4*).

The National Aeronautics and Space Administration (NASA) also maintains a large archive of environmental data, with more than 20 PB stored as of 2017, and it projects that the archive will grow to nearly 250 PB by 2025. NASA is exploring the use of commercial cloud providers to store and distribute these data. Unlike the NOAA program, however, in the initial phase NASA is paying the commercial provider to store and distribute data, with the provider charging NASA the costs of distribution and analysis, rather than charging the end user (*5*).

The 2018 National Institutes of Health (NIH) Strategic Plan for Data Science notes that genomics is one of the largest data-generating fields and that the amount of data stored at NIH's National Center for Biotechnology Information has been growing exponentially for many years, with no signs of slowing. NIH intends to leverage private sector cloud providers to ensure continued data access for users while minimizing infrastructure and maintenance costs. The distribution of costs among NIH, cloud providers, and researchers has not been determined (*6*).

To understand why this new technology is changing the distribution of costs, it is necessary to understand the economics of open data and how the new models compare to existing government open data systems.

ECONOMICS OF OPEN DATA

A key economic attribute of data is that they are a nonrival good; when one person uses data, those data are still available for others to use. In the case of perfectly nonrival goods, the marginal cost of providing the good to an additional user is zero—the costs are all associated with the fixed effort of creating the data, and they remain the same whether the data are used 10 times or 10,000. Because societal benefits from those data will be maximized when everyone who wants to access and use the data does so, the best way to achieve this is to make those data freely available to all.

This logic has been central to the broad adoption of open data policies among government agencies. Ethical arguments have also played a role, asserting that taxpayers have already paid to collect the data and shouldn't have to pay again, and that open data policies are key to government transparency and citizen engagement (*7*). Experience suggests that these policies are working, having increased the distribution of U.S. Geological Survey Landsat satellite images (*8*), increased scientific publications using Earth-monitoring sensors (*9*), and increased the pace with which genomics advanced (*10*).

This economic argument relies on the assumption that the marginal cost of providing data to an additional user is zero. Under the current system, in which agencies store data in government-owned servers and make it available via a web portal for users to download, this assumption is fairly accurate. Although the cost is not truly zero, it is very low, and agencies do not typically calculate or charge users a fee associated with their individual download (however, many agencies place restrictions on the total download volume within a period of time to ensure that costs remain reasonable; users wishing to access very large volumes of data often work with agencies to do so in a more timely or cost-effective manner). By contrast, commercial cloud providers, as part of their business model, calculate the cost of each marginal download (the egress cost). When the government works with a commercial provider, this cost is typically passed on either to the agency or to the end user.

Cloud computing also raises another issue not addressed under the current system. Users have to not only access the data, but also be able to view or analyze them. In the current system, data can be downloaded to a personal or institutional computer and analyzed there. This involves essentially no marginal

Sam Nunn School of International Affairs, Georgia Institute of Technology, Atlanta, GA 30332, USA. Email: mariel.borowitz@inta.gatech.edu

Published by AAAS

cost—researchers use computers and software that they already own and use for other purposes. In a cloud computing model, users will increasingly find it efficient to do analysis in the cloud. This would avoid the issue of egress costs. However, unlike the current model, users face a marginal fee to carry out each study or analysis as they use cloud computing capability for that specific purpose. Although computing and analysis costs are not generally considered part of open data policies, the practical effect of this change on the accessibility of data are important.

IMPACTS AND OPPORTUNITIES

The transition to cloud computing is forcing agencies to rethink how they implement open data policies, and decisions about cost allocation are not simple. For example, at the Workshop on Maximizing the Scientific Return of NASA Data, held in October 2018, multiple U.S. and international agencies discussed these transitions, including technical, organizational, and policy challenges. But with the multitude of issues to discuss, such as the relative benefits of in-house versus commercial systems, the best way to solicit user feedback, and the importance of user training, differences in cost structures and their impacts received little attention.

Generally, agencies remain committed to ensuring that the data remain free—they aren't commercializing their data or generating revenue from data distribution. However, government policy does allow agencies to recover costs associated with dissemination of data (*11*). Agencies have not traditionally calculated the marginal cost of user downloads nor paid costs associated with individual data analyses. Agencies may feel that it is fairer if users cover these user-specific fees, or agencies may simply not have adequate funds to cover these costs.

Although the exact impact of policies requiring users to pay to work with data depends on specific implementation and pricing structures, history suggests that much of the benefit associated with open data may be lost. To avoid decreasing the benefits to society from government data, agencies must maintain truly free access. To do so, agencies will need to budget for egress and analysis costs, which they have not had responsibility for in the past, and legislators, in turn, will need to understand the purpose of, and approve, funds for these uses. Many argue that the total costs of commercial cloud data provision will be lower than the total costs of developing in-house systems, because agencies will not need to procure and maintain their own servers and data portals. Others point out that even if this is the case when the system is first developed, there is no guarantee that commercial entities will not raise prices in the future after agencies have committed to one provider. Regardless of total cost, changes in the structure of costs will still need to be understood by, and justified to, decision-makers.

There will also be many specific issues to address in implementation. For example, some reasonable limits would need to be placed on the costs covered by the agency. Similar to policies already in place at some agencies, a user who wants to download petabytes of data to their private or university system or who wishes to use massive amounts of cloud computing capability, for example, would need to work with the agency to discuss a workable solution, likely involving some cost to the user.

A researcher examines NOAA fisheries sonar data.

PHOTO: NOAA'S NATIONAL OCEAN SERVICE/FLICKR

To the extent that agencies have implemented programs that involve different distributions of fees or different limitations in data egress or analysis, metrics should be collected to allow analysis of how these programs affect data use. When possible, policy implementation should adopt an experimental design approach. These empirical data can help to validate theoretical and historical understandings of the impact of variations in data policies and can help to improve our quantitative understanding of these impacts. Agencies should also consider non-economic impacts of shifting data to the commercial cloud, such as those on security and government transparency.

Changes to cost structures may not be the only challenge to arise from the transition to commercial cloud computing. For example, because U.S. agencies are barred by the 1976 Copyright Act from copyrighting works created by the federal government, commercial entities can repackage government data and distribute it under their own licenses, restricting access to particular groups or uses. In the past, the existence of government portals has ensured that citizens have at least one option for free and open access, but if a commercial system fully replaces the government portal, this may no longer be the case. Without carefully thought-out contract language, commercial providers with a monopoly on government data distribution could potentially put in place licenses that would substantially decrease reuse of the data. The pilot programs discussed above have addressed this issue in existing agreements, but continued awareness of this and similar issues is important as these programs evolve and new programs begin.

Building dedicated, in-house clouds may avoid some challenges related to commercial pricing structures, data security, and long-term availability, but it raises others. Companies like Amazon, Google, and Microsoft have vast workforces dedicated to developing and advancing cloud technology. They have infrastructure in multiple locations around the world to ensure redundancy in multiple physical locations. It seems unlikely that a government agency with limited resources can compete in terms of capabilities and price. Using public-private partnerships to develop in-house cloud-computing systems may lower costs, but once again raises challenges in cost structures and data access. Political realities also mean that agencies must consider industrial policy and the value of working with domestic companies.

The transition to cloud environments for the distribution and analysis of government data is under way, but in most cases, long-term decisions have not been made. Now is the best opportunity to design systems that help us to understand and maximize the value of open government data. ■

REFERENCES

1. European Space Agency, "Accessing Copernicus Data Made Easier," 14 December 2017; www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Accessing_Copernicus_data_made_easier.
2. "State governments fast embracing AWS Cloud in India: Teresa Carlson," *Economic Times (India)*, 20 September 2018.
3. R. Showstack, *Eos (Wash. D.C.)* **95**, 95 (2014).
4. National Oceanic and Atmospheric Administration, "Big Data Project: Frequently Asked Questions"; www.noaa.gov/big-data-project-frequently-asked-questions.
5. National Aeronautics and Space Administration, "EOSDIS. Inf. Serv. Cloud Evolution," NASA, September 2018; https://earthdata.nasa.gov/about/eosdis-cloud-evolution.
6. National Institutes of Health, "NIH Strategic Plan for Data Science," NIH, June 2018; https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf.
7. M. Borowitz, *Open Space: The Global Effort for Open Access to Environmental Satellite Data* (MIT Press, 2017).
8. M. A. Wulder *et al.*, *Remote Sens. Environ.* **122**, 2 (2012).
9. C. Wilson, *ICES J. Mar. Sci.* **68**, 677 (2011).
10. J. Kaye *et al.*, *Nat. Rev. Genet.* **10**, 331 (2009).
11. United States, Office of Management and Budget, Circular A-130, Managing Information as a Strategic Resource (2016).

10.1126/science.aat5474

Science

Government data, commercial cloud: Will public access suffer?

Mariel Borowitz

Science **363** (6427), 588-589.
DOI: 10.1126/science.aat5474