THE CENTER FOR
OPEN DATA ENTERPRISE

Open Data Roundtable on Improving Data Quality: KEY TAKEAWAYS

In 2016, the White House Office of Science and Technology Policy and the Center for Open Data Enterprise co-hosted four Open Data Roundtables to identify case studies, lessons learned, and best practices in open data across the federal government. Open data from government is free, publicly-available data that anyone can use and republish. The Center has summarized these key takeaways from each Roundtable, which brought together experts from inside and outside of government with technical, policy, and legal backgrounds. The Center will publish a full report of Roundtable findings in fall 2016.

BACKGROUND

On April 27th, 2016, the White House and the Center for Open Data Enterprise co-hosted a Roundtable to address a key issue: ***How to improve data quality in efficient and scalable ways.*** Organizations that want to use open government data face a number of obstacles as a result of quality issues with the data. Government agencies and their data users are now working to improve data quality by addressing issues such as timeliness, accuracy, precision, and interoperability.

The Roundtable brought together 75 experts from government, nonprofits, academia, and the private sector to address data quality. Participants were not asked to develop consensus recommendations but to share their own observations and suggestions.

OPPORTUNITIES FOR IMPROVEMENT

Participants identified issues that they have faced in attempting to improve data quality.

- **Limited funding and resources for improving published data** make it difficult to do quality control and quality assurance, which are costly and time-consuming. High-quality data poses a "tragedy of the commons" problem. It is a public good that benefits everyone but is not any specific organization's responsibility.

- **Quality improvements need to be prioritized.** It can be challenging to quantify the value of high-quality data to justify the investment in quality improvement. Better ways to identify high-value datasets are needed, beyond the judgment of individual data providers.

- **Better metadata and metadata standards are needed** to ensure data quality, usability, and interoperability. Many datasets are still released as PDFs rather than in machine-readable formats. Organizations and attributes need to be described with unique identifiers. In addition to these problems with federal data, better ways are needed to aggregate and compare local data. For example:

  - In healthcare delivery and administration of services, data is reported with little standardization regarding patients, providers, insurers, pharmaceutical treatments, and medical devices.
  - In education, it is difficult to ensure one school's data will be comparable to others and to develop reliable statistics.
  - In public safety, the same crime data terms may be described differently depending on the location, or the same term may describe different things across data sources.

- **Human factors need to be addressed in data curation**, which is a human-intensive activity that requires specialized expertise. While data providers may curate their own data, they may not have the full perspective to do so in a way that meets users' needs. Better ways to identify high-value datasets are needed, beyond the judgment of individual data providers. In addition, once data is released, potential users need help finding the datasets that will be most useful to them.

STRATEGIES FOR IMPROVING DATA QUALITY

Participants developed the following best practices and new initiatives for tackling data quality challenges.

BEST PRACTICES

- **Strong data governance.** Agencies can improve data quality throughout the "data lifecycle" by strengthening data governance. Chief Data Officers can lead this effort, particularly if they are empowered to lead the collection, management, and dissemination of data across the agency. It's critical to ensure quality during data collection since errors are difficult or impossible to correct later in the process.

- **Effective feedback systems.** User feedback is an effective tool that can help identify and eliminate data quality issues. Possible strategies include:
  - Developing data quality feedback loops, like a 311 or 611 service, as a common convention for government data providers.
  - Developing stronger feedback channels for data.gov. The portal could develop a public interface for collecting feedback and engage the community to analyze and respond to comments.
  - Institutionalizing Demand-Driven Open Data (DDOD) across government. The DDOD program developed by the Idea Lab at Health and Human Services can leverage user feedback to improve data. Benefits include prioritizing resources to improve quality of the most widely used and valuable datasets.

- **Improved data policies.** Policies such as the Information Quality Act and ISO 8000, which set out quality requirements for open government data, should be reviewed and updated. The OMB Memo M-13-13, which established the current open data policy, could be expanded to cover all government data, not only the Executive Branch, and could be integrated into government modernization plans.

Participants also suggested issuing a second memo clarifying the meaning of “machine-readability,” as some agencies still struggle to move beyond PDFs.

NEW INITIATIVES

- Develop **common tools, platforms, and catalogs** for sharing and improving data. Platforms such as MediaWiki, GitHub, Hackpad can be used to share information about datasets and flag data quality issues. Participants suggested “a Pinterest for data” to assess what data is being used and shared.

- Develop **new data standards.** The federal government can play a lead role in creating standards and taxonomies across all levels of government. Participants also suggested establishing and communicating the best formats to publish data (e.g. appropriate use of raw data files, zipped files, documents, websites) and clarifying that PDFs do not count as structured data.

- Provide **open data guidance** from OSTP and OMB to both federal agencies and state and local government on how to implement their open data programs. Participants suggested:
  - Publishing the Open Data Playbook to show general rules for open data.
  - Develop a coherent government-wide strategy for codifying data quality and standardization, as a service.
  - Hold centralized, sector-specific convenings around data standards, including government as well as industry.
  - Develop a data quality rating system for government agencies according to predetermined quality requirements. Give badges for being compliant with a standard of quality.
  - Give agencies a tool for data validation and assessing metadata quality.
  - Provide authoritative guidelines on how data should be structured, using the 18F guidelines for APIs as a possible model.

- **Utilize open source platforms to use multiple datasets together.** While uniform standards are the ideal, it is possible to use datasets together even when they are not technically interoperable and have different metadata standards. Using the Census Bureau’s CitySDK toolkit as a model, this approach would create a customer-centric, shared federal service providing insights across all sectors. It would also make it possible to use federal data in a local context, and could be tested out with communities through events such as the National Day of Civic Hacking.

ABOUT THE CENTER FOR OPEN DATA ENTERPRISE

The Center for Open Data Enterprise is an independent nonprofit 501(c)(3) organization, based in Washington, DC., whose mission is to maximize the value of open data as a public resource. We thank our Open Data Partner Microsoft and Open Data Supporter Booz Allen Hamilton for supporting the Center’s work on the Open Data Roundtables. We welcome feedback on this report. Please send comments and inquiries to Katherine Garcia, Director of Communications and Outreach, at katherine@odenterprise.org.