THE CENTER FOR
OPEN DATA ENTERPRISE

Discussion Draft:

Briefing Paper on Open Data and Improving Data Quality

Introduction

The U.S. federal government, like many governments around the world, releases *open data* that can be used in a variety of ways. This important public resource has helped the public find better value in education, fair housing, and safer medicines, and has a wide range of other social and economic benefits. Open data also helps government agencies themselves operate more efficiently, share information, and engage the citizens they serve.

Under the U.S. government's Open Data Policy,[1] all federal agencies "must adopt a presumption in favor of openness to the extent permitted by law and subject to privacy, confidentiality, security, or other valid restrictions." Truly open government data is free, public data that anyone can use and republish. Some kinds of data, like data containing personally identifiable information, may not be appropriate for wide, unfettered access, but can still be of use to non-government communities.

While many businesses and other organizations are now working to apply open data, problems in data quality, data management, and data governance can make it difficult to use. To help find scalable solutions, the White House Office of Science and Technology Policy and the Center for Open Data Enterprise are co-hosting the 2016 Open Data Roundtables. These Roundtables bring together participants from federal agencies, academia, the private sector, and nonprofit organizations with technical, policy, and legal expertise. They are designed to:

- Identify open data case studies, learned lessons and best practices across the federal government;
- Strengthen a community of technical, legal, and policy experts in support of open data;
- Support continuity and accelerate the progress of open data work.

This briefing paper presents an overview of the topic for the second Roundtable, which will cover data quality. The goal of this Roundtable is to find strategies and solutions that can help improve data quality across the federal government in efficient and scalable ways.

[1] Sylvia M. Burwell, Steven VanRoekel, Todd Park, Dominic J. Mancini, "M-13-13, Open Data Policy - Managing Information As an Asset", Executive Office of the President, Office of Management and Budget, May 9, 2013, https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf (accessed April 20, 2016).

This paper is meant to be used as a framing document and is not an exhaustive treatment of the subject. It is not an expression of federal policy or the preferred direction of federal policy. It is designed to support participants in their work and should not restrict them from covering other topics that they believe are central to the issue.

Why Data Quality Matters

In order for open data to become a widely used, valuable public resource, government agencies will need to address data quality efficiently, effectively, and systematically. Only high-quality data can be used to "translate…data into meaningful insights that…can [be used] to improve business processes, make smart decisions, and create strategic advantages."[2]

Data quality can be broadly defined as "the suitability of the data for a particular purpose."[3] There is a large academic literature on the many dimensions of data quality, which can include such factors as completeness, coverage, conformity, consistency, accuracy, duplication, or timeliness[4]. Quality has come to be "viewed as a multi-faceted concept," assessed based on "user perspectives, needs and priorities, which vary across groups of users."[5] While data users inside and outside government need data that is of sufficient quality for the intended use, organizations that rely on open data have stressed that government should not hold back important data in the interest of quality improvement. It will take a dialogue between data providers and data users to balance data quality and timely availability.

Federal Data Quality Frameworks

The U.S. government has recognized the value of its data for many years, and a statutory framework has been developed to ensure that quality is an important concern in Executive branch data dissemination. That framework includes the Paperwork Reduction Act of 1980 and 1995,[6] and the Information Quality Act of 2000.[7] Most recently, Congress passed the DATA Act to make data on federal spending more accessible and standardized.[8]

Pursuant to the Information Quality Act, in 2002 the Office of Management and Budget published "Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information

[2] Hongwei Zhu, Stuart E. Madnick, Yang W. Lee, Richard Y. Yang, "Data and Information Quality Research: Its Evolution and Future", in "Computing Handbook: Information Systems and Information Technology", CRC Press, edited by Heikki Topi and Allen Tucker, 2014.

[3] Josh Tauberer, "Open Government Data: The Book", April 2012, https://opengovdata.io/2014/data-quality (accessed March 15, 2016).

[4] See, for example, "The MIT Total Data Quality Management Program", MIT, http://web.mit.edu/tdqm/www/about.shtml
OECD Statistics Directorate, "Quality Framework and Guidelines for OECD Statistical Activities", The Organisation for Ecomomic Co-Operation and Development, January 2011, http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=std/qfs%282011%291&doclanguage=en (accessed March 15, 2016), 7.

[6] 44 U.S.C. §3510.

[7] Section 515 of Public Law 106-554

[8] Add citation

Disseminated by Federal Agencies" in the Federal Register.[9] That guidance required federal agencies to "issue guidelines ensuring and maximizing the quality, objectivity, utility, and integrity of information (including statistical information) disseminated by the agency..."[10]

The Obama Administration's 2009 Open Government Directive[11] instructed agencies to comply with a new framework for the quality of spending information and asked them to renew their focus on ensuring that their information more broadly complied with Data Quality Act guidelines.[12] This Administration's efforts to promote open data more broadly, from the Open Government Directive to Project Open Data,[13] have consistently included efforts to boost data quality. The Administration has stated that "reliable data allows the public to trust in the information the government provides and for federal and elected officials to use that information to make informed decisions about government programs and projects. It allows federal managers to analyze and better structure government programs to prevent waste, fraud, and abuse."[14]

The federal frameworks in place provide a context for assessing the elements of data quality and what can be done to improve them. This paper describes several of the key factors that determine data quality; describes challenges facing government agencies looking to release high-quality data; and explores approaches to creating, maintaining and publishing quality data.

The Elements of Data Quality

Experts define data quality in many ways, but quality can be described with a relatively short list of common elements. These elements are related but independent. Data can meet some criteria for quality while failing to meet others.

Accuracy: Simply put, accuracy "is the likelihood that the [given] data reflect the truth."[15] Datasets may have hidden accuracy problems that are difficult to detect, or problems that quickly become obvious to anyone trying to use the data. For example, datasets with inaccurate geospatial data show their flaws when the data is mapped and data points show up in clearly inaccurate places. Tables of data that do not add up correctly along rows or columns can also signal problems in data accuracy.

[9] Office of Management and Budget, "Federal Register Notice: Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies", Executive Office of the President: Office of Management and Budget, February 22, 2002, https://www.whitehouse.gov/sites/default/files/omb/fedreg/reproducible2.pdf (accessed March 18, 2016), 8452.

[10] Ibid, 8452.

[11] Peter R. Orszag, "Open Government Directive", The White House, December 8, 2009, https://www.whitehouse.gov/open/documents/open-government-directive (accessed March 18, 2016).

[12] "Open Government Directive", The White House.

[13] Project Open Data, https://project-open-data.cio.gov (accessed March 30, 2016).

[14] Norman Dong, "Memorandum for Agency Chief Financial Officers: Improving Data Quality for USASpending.gov", Executive Office of the President, Office of Management and Budget, June 12, 2013, https://www.whitehouse.gov/sites/default/files/omb/financial/memos/improving-data-quality-for-usaspending-gov.pdf (accessed March 18, 2016), 1.

[15] Ibid.

Precision should also be considered as part of accuracy. Measurements that are precise are as close as possible to the actual size, weight, length of an object, the actual time an event occurred, or other measures.

Metadata: Metadata – data that provides information about data – gives users crucial contextual information. Metadata included as part of a dataset can provide information about the origins of the data, legal restrictions on its use, and other critical factors. It also makes it easier to find important data by searching through relevant criteria.

In addition to metadata, documentation on data sources, limitations, and other factors can give important information and context about a dataset. When datasets are poorly documented, data users may not be able to tell whether they can rely on them and for what purposes. To use the data with confidence, they need documentation that helps them understand the data and how to use it, including any known flaws in the data.

Machine-readability: The Open Data Policy states that open data should be "made available in convenient, modifiable, and open formats that can be retrieved, downloaded, indexed, and searched. Formats should be machine-readable (i.e., data are reasonably structured to allow automated processing)."[16] When data is released in unstructured formats that aren't machine-readable its use is severely restricted. Many federal agencies still release data as PDFs that can't be analyzed until they are converted into digital files, a process that requires both electronic conversion and manual work. Despite the fact that Administration policy calls for data to be released in machine-readable formats, and U.S. Chief Data Scientist Dr. DJ Patil and other experts frequently state that "PDFs aren't open data,"[17] federal agencies still struggle to move beyond PDF documents to more usable formats. PDFs can't be analyzed until they are converted into digital files, a process that requires both electronic conversion and manual work.

Timeliness: Data timeliness depends on two factors: the time it takes for data to become available after the event or phenomenon it describes (also called latency), and how frequently datasets need to be updated to remain relevant to data users.[18]

Timeliness is especially important for data about areas that change rapidly, such as data relevant to federal elections. The United States Senate, for example, has a notably slow process for filing campaign finance reports. Unlike candidates for the House of Representatives and the White House, who file electronically, Senate candidates file paper documents that need to be entered into the system manually. This process results in significant time delays, making it difficult to conduct relevant pre-election analysis of the data. [19]

[16] Burwell et al, Memo 13-13.

[17] Akanksha Jayanthi, "12 highlights from US Chief Data Scientist DJ Patil's #WHData Twitter chat", Becker's Health IT & CIO Review, August 20, 2015, http://www.beckershospitalreview.com/healthcare-information-technology/12-highlights-from-us-chief-data-scientist-dj-patil-s-whdata-twitter-chat.html (accessed April 2, 2016).

[18] OECD, 9.

[19] OpenSecrets, "e-Filing Senate Campaign Reports", Center for Responsive Politics, https://www.opensecrets.org/action/issues/efiling-senate-campaign-finance-reports (accessed March 21, 2016).

Granularity: A spreadsheet with many data points broken out by multiple subcomponents could be said to be of high granularity, while the same dataset summarized at a high level, presenting the data in aggregate form, may be less useful because it is less detailed. Developers, computer scientists, and statisticians often desire data in the most granular form it is available. An earlier Briefing Paper and Roundtable explored ways to open granular data while protecting privacy in datasets that contain personally identifiable information.

The need for granularity varies by the type of data and its use. Under the Justice Department's Foreign Agents Registration Act (FARA), for example, lobbyists and others representing foreign interests in Washington, DC report far more granular information on their activities than those representing domestic clients. They are required to disclose copies of contracts, materials that they distribute on behalf of foreign clients, details on their political contributions, and more.[20]

Interoperability: Interoperability is an increasingly important element of data quality and one that has a number of dimensions. Different information technology systems and software applications are interoperable when they can communicate, exchange data, and use the information that has been exchanged. Interoperability makes it possible to combine and leverage datasets across various entities and data sources, including government agencies, businesses, and nonprofit organizations.

To be interoperable, datasets must be coherent or "mutually consistent" internally as well as across time, location, and organizations or geographies.[21] To achieve this goal, datasets should apply consistent definitions, standard naming conventions, and unique identifiers to the data they contain. When consistency is not possible, discrepancies should be clearly documented.

The Office of the National Coordinator for Health Information Technology (ONC) is now focused on improving data interoperability. The ONC endeavors to ensure that the health care providers adopt architectures that make it possible for diverse electronic health record (EHR) systems to work compatibly in a true information network[22].

Standardization is often needed to make multiple datasets interoperable. It is easier to compare and combine datasets if they use standard, consistent definitions, identifiers, and data formats. In contrast, when federal datasets define basic terms in different ways, it makes it difficult to analyze data within an agency or between agencies. For example, "datasets across [the United States Department of Agriculture] have different codes and standards as a result of the authorizing legislation for different data collections."[23] There are currently eight different statutory definitions of 'rural area', a lack of coherence that makes analysis of these datasets difficult. Similarly, basic terms like "revenue" are not always defined consistently across all federal financial data.

[20] Ben Freeman, Lydia Dennett, "Loopholes, Filing Failures, and Lax Enforcement: How the Foreign Agents Registration Act Falls Short", The Project on Government Oversight, December 16, 2014, http://pogoarchives.org/m/fara/pogo-fara-report-20141216.pdf (accessed March 21, 2016), 8.

[21] OECD, 10.

[22] https://www.healthit.gov/providers-professionals/interoperability-training-courses

[23] Joel Gurin, Audrey Ariss, Katherine Garcia, Laura Manley, "Using Open Data to Protect the Food Supply: A Roundtable with the U.S. Department of Agriculture," GovLab, January 2015, (accessed March 31, 2016),18.

Another pervasive problem is the lack of standard identifiers for many entities in federal datasets. The same company can show up under many different names – for example, “IBM” and “International Business Machines” – in different datasets or within a single dataset. Lack of standard nomenclature can make it difficult to track entities or individuals across datasets and systems. This issue has been raised in the context of federal spending[24] which is now tracked using a number of incompatible identifiers (including DUNS Numbers and CAGE Codes).[25]

In contrast, OMB’s Office of Information and Regulatory Affairs defines a variety of standards and best practices for use in population and establishment surveys conducted by the federal government. There are federal standards for the collection of race and ethnicity data[26] that allow compatibility of numerators and denominators across large scale statistical surveys. Similarly, the business sector relies on the federal standards in the North American Industry Classification System (NAICS) [27] and the Standard Occupational Classifications (SOCs)[28].

Approaches to Quality Improvement

Digitization and Automation of Data Collection

Manual data collection and entry is a common source of data quality problems. Modern systems allow for electronic filing, automated device information capture, and automated metadata production that can help reduce the rate of human error.[29] New technological processes also make it possible to convert PDFs to digital information with a minimum of human involvement, potentially easing the shift from those documents to machine readable data.

The best way to ensure data quality is to design data collection processes with an eye towards ensuring quality from the start. As these systems are implemented, they may help prevent data quality problems associated with accuracy, precision, timeliness, and usability by producing data in more useful formats that can be automatically certified, checked for quality, and disseminated.

A study analyzing “the various data quality issues related to the design, implementation and operation of a specific data initiative, the U.S. Army's Medical Command (MEDCOM) Medical Operational Data System (MODS) project” highlighted some of the benefits associated with more automated processes.[30] MEDCOM MODS is not an open data project, but the lessons related to data quality are applicable. The case study noted issues with the accuracy of geospatial data, and argued that “the adoption of automated device capture of geospatial data for MODS will actually dramatically improve its quality

[24] Center for Open Data Enterprise, “Tracking Federal Spending: An Open Data Roundtable with the U.S. Department of the Treasury”, Center for Open Data Enterprise, DRAFT IN PROGRESS, (accessed March 31, 2016), 8.

[25] FSD.gov, “What is a CAGE code?”, General Services Administration, Last updated November 30, 2015, https://www.fsd.gov/fsd-gov/answer.do?sysparm_number=KB0011119 (accessed March 31, 2016).

[26] https://www.whitehouse.gov/omb/fedreg_1997standards

[27] http://www.census.gov/eos/www/naics

[28] http://www.bls.gov/soc

[29] Dave Becker, Trish Dunn King, Bill McMullen, Lisa Deifer Lalis, David Bloom, Ali Obaidi, Donna Fickett, “Big Data Quality Case Study: Preliminary Findings, U.S. Army MEDCOM MODS”, The MITRE Corporation, September 2013, http://www.mitre.org/sites/default/files/publications/13-2568.pdf (accessed March 18, 2016),12.

[30] The MITRE Corporation.

while at the same time making it much easier and more efficient by removing manual entry steps and eliminating human generated data entry errors."[31] Automating instruments also allows for the collection of more precise geospatial data.[32]

Embracing e-filing for data collection can help agencies increase accuracy and timeliness while avoiding some of the problems associated with inflexible formats like PDF. Recently, supporting "a legislative change from paper-based or PDF systems to an e-filing system for patent applications" was suggested as one way for the U.S. Patent and Trademark Office to tackle data quality issues.[33] E-filing is among recommendations suggested to improve SEC financial information and the FARA system mentioned earlier in this paper. Congress has also recognized the importance of accessible data from the form 990, which all nonprofit organizations file with the IRS, and recently proposed legislation calls for it to be filed and released electronically.[34]

That said, modernizing alone will likely not eliminate all data quality problems. Standardized formats, frequent quality checks, and robust governance structures have also been suggested as hedges against human-related and other types of errors.[35]

Standardization

As described above, standard definitions and identifiers play an important role in ensuring data quality by making data more interoperable.

Agencies are considering how to make their definitions more consistent internally as well as across government.[36] The DATA Act, which will overhaul the way federal spending is tracked and reported to the public, requires establishing consistent definitions across a wide range of datasets and systems. The goal of this process is to "create both functional and IT definitions and standards that allow for consistency across communities, so that data elements reported from different agencies correspond in definition and in format." [37]

Additionally, efforts such as the National Information Exchange Model (NIEM) exist to help facilitate standardization around specific types of information across agencies and levels of government. NIEM

[31] Ibid, 12.

[32] Ibid, 16.

[33] "Report of Findings from an Open Data Roundtable with the U.S. Patent and Trademark Office", GovLab, 14.

[34] "S.2750 – Charities Helping Americans Regularly Throughout the Year Act", Sen. John Thune and Sen. Ron Wyden, April 6, 2016, https://www.congress.gov/bill/114th-congress/senate-bill/2750, (accessed April 14, 2016).

[35] Martin Doyle, "Is Open Data At Risk From Poor Data Quality?", Business 2 Community, September 20, 2014, http://www.business2community.com/big-data/open-data-risk-poor-data-quality-01010535#V5cfdmiXX8w50PIF.97 (accessed March 21, 2016).

[36] Center for Open Data Enterprise, "Improving Safety Data: A Roundtable with the U.S. Department of Transportation", Center for Open Data Enterprise, 2015, https://s3.amazonaws.com/odenterprise/DoT+Roundtable+Report.pdf (accessed March 31, 2016), 7.

[37] Federal Spending Transparency, "Frequently Asked Questions: What is the meaning of 'data standards?'", http://fedspendingtransparency.github.io/faq (accessed March 21, 2016).

"connects communities of people who share a common need to exchange information in order to advance their mission" through a community-driven standards based model.[38]

Federal agencies are moving to establish standard identifiers to make it easier to combine datasets. For example, the U.S. Patent and Trademark Office's (USPTO) PatentsView uses disambiguation techniques to identify inventors and assign them a unique ID.[39] More generally, the development of a new "unique entity identifier" system, such as the Legal Entity Identifier,[40] could help correct this problem across government. Along with consistent definitions, settling on a common, open entity identifier has been identified as a key to improving federal spending data, and options to transition to such a system are being examined.[41]

Standards must be flexible and updated to ensure that they remain strong over time. For example a standardized list of company names cannot simply reflect the world on the day that it is adopted. Instead it would have to be updated consistently as new companies are launched, and others evolve, acquire, merge and divest.

Crowdsourcing and User Feedback

The simple act of releasing data publicly can have positive effects on data quality.[42] Just preparing data for publication can reveal "problems and issues with...processes that were never highlighted before."[43] Once the data has been released, ongoing public feedback, in the form of crowdsourcing, is essential to ensuring that the data is relevant and useful.

Crowdsourcing can be a strategy for improving data by inviting users to help correct inaccuracies or add new information. For this strategy to be effective, however, it will need to include more than simply releasing data into the world and waiting for data users to help improve it. In fact, British Prime Minister David Cameron's assertion that open data will result in an "army of effective armchair auditors" has been slow to emerge.[44]

Google's Mapmaker program is a prime private sector example of building data quality through crowdsourcing.[45] The program allows average Google Maps users to share information about places they are familiar with, finding errors and boosting the quality of Google Map data. Users who regularly

[38] "About NIEM", The National Information Exchange Model, https://www.niem.gov/aboutniem/Pages/niem.aspx (accessed March 31, 2016).

[39] "PatentsView: Methods and Sources", US Patent and Trademark Office, http://www.patentsview.org/web, (accessed April 20, 2016).

[40] https://financialresearch.gov/data/legal-entity-identifier

[41] Center for Open Data Enterprise, "Tracking Federal Spending"

[42] Open Data Institute, "Environment Agency: Going Open" Open Data Institute, http://theodi.org/ea-going-open-benefits-for-ea (acccessed March 18, 2016), Benefits for EA.

[43] Andreas Addison, "OpenGov Voices: The next chapter of open data", The Sunlight Foundation Blog, February 29, 2016, http://sunlightfoundation.com/blog/2016/02/29/opengov-voices-the-next-chapter-of-open-data (accessed March 18, 2016).

[44] Becky Hogge "Open Data: Six Stories About Impact in the UK", Omidyar Network, November 2015, http://theodi.org/ea-going-open-benefits-for-ea (accessed March 18, 2016), 4.

[45] https://www.google.com/mapmaker/about

submit accurate information are given increased moderation and editing power.[46] This two tiered system allows average users to build expertise, rewards power users, encourages engagement, and results in timely, accurate data.

Crowdsourcing can be an effective quality mechanism if fully considered and robustly implemented. The United States Agency for International Development (USAID) was able to significantly boost the quality of some selected data through a carefully considered combination of crowdsourcing and engagement with expert stakeholders both inside and outside government.[47] In fact, this process produced more accurate results than the automated process the agency applied to the bulk of the dataset in question.[48] Ongoing feedback loops can also enable users to spot and correct errors in the data.

By putting processes in place to learn about their data users and how they apply published data, governments can fine-tune their data releases and make the most relevant datasets more accessible and usable. Project Open Data, the online hub for the Administration's open data program, outlines a number of events and communications avenues to engage stakeholders and gather their feedback.[49] The administration has also called for "agencies to engage with data users to prioritize release of open government data… [through] customer-friendly feedback mechanisms on opening new datasets and improving existing datasets."[50]

The Department of Health and Human Services has created a possible model with its Demand Driven Open Data (DDOD) Program. DDOD is "a framework of tools and methodologies to provide a systematic, ongoing and transparent mechanism for you to tell public data owners what's most valuable."[51] In other words, HHS, via DDOD, is engaging stakeholders in an attempt to ensure that it is considering its audience and releasing relevant, high quality data.

Private Sector Data and Collaboration

The private sector may help improve government data quality in two ways. First, private sector data can be combined with public sources to create higher quality outputs – if both sides can figure out how to do this without releasing proprietary information that must be protected. And second, private sector companies may collaborate with government to develop new approaches to quality improvement.

There are several potential opportunities for government and private sector organizations to share data to boost quality. For example, it has been suggested that the Department of Labor needs "a better

[46] https://www.google.com/mapmaker/about/regionalleads

[47] Shadrock Roberts, Stephanie Grosser, D. Ben Swartley, "Crowdsourcing to Geocode Development Credit Authority Data: A Case Study", United States Agency for International Development, https://www.usaid.gov/sites/default/files/documents/2151/USAIDCrowdsourcingCaseStudy.pdf (accessed March 18, 2016).

[48] Ibid, 1.

[49] "Open Data Engagement, Project Open Data, https://project-open-data.cio.gov (accessed March 30, 2016).

[50] "The Open Government Partnership: Third Open Government National Action Plan for the United States of America", October 27, 2015, https://www.whitehouse.gov/sites/default/files/microsites/ostp/final_us_open_government_national_action_plan_3_0.pdf (accessed March 18, 2016), 11.

[51] "Demand Driven Open Data for the U.S. Department of Health and Human Services", last modified July 5, 2015, http://ddod.healthdata.gov/wiki/Main_Page (accessed March 29, 2016)

taxonomy and hierarchies of skills, including distinctions between professional and soft skills and between skills and specific tasks."[52] Private sector companies are a potentially significant source of new, timely labor-related data. For example, LinkedIn has a database of over 45,000 skills that could be leveraged to improve federal data quality.[53]

Agencies are also beginning to use collaborative efforts to improve their data and data-delivery platforms. USPTO's PatentsView website was developed in collaboration with the companies Twin Arch Technologies and Periscopic, as well as the American Institutes for Research, the University of California at Berkeley, and the USDA. The site provides patent data in a way that makes it easy to find and has improved the quality of the data in the process. PatentsView aims to make it easier for researchers to focus on making productive use of the data by eliminating "wasteful and redundant cleaning, converting and matching…by many individual researchers."[54]

Governance Structures

It is common for organizations dealing with data to propose guidelines or implement governance structures in an attempt to ensure that their data maintains certain standards of quality. Examples range from the OECD's Quality Framework and Guidelines for OECD Statistical Activities[55], to the EPA's Open Government Data Quality Plan,[56] written to comply with OMB guidance, to the W3C's "Data Quality Vocabulary".[57] A number of governance structures from the federal context were detailed in the introductory section of this paper.

Data Lifecycle Management

Federal agencies are realizing the importance of "data lifecycle management" – an approach to the integrated management of data at all stages. The U.S. Geological Survey (USGS) describes the goals: "When we start thinking of our data as corporate assets with value beyond our immediate need, the idea of managing data through a whole lifecycle becomes more relevant. All of the questions of documentation, storage, quality assurance, and ownership then need to be answered for each stage of the data lifecycle, starting with the recognition of a need and ending with archiving or updating the information." The USGS describes six stages of data lifecycle management: Plan, Acquire, Process, Analyze, Preserve, and Publish/Share.[58]

Other public and private sector methods for Lifecycle Management may be applicable to open data. These include the Task, Collect, Process, Exploit, Disseminate (TCPED) model often used in the

[52] Center for Open Data Enterprise, "An Open Data Roundtable with the U.S. Department of Labor", Center for Open Data Enterprise, DRAFT IN PROGRESS, (accessed April 8, 2016), 7.

[53] Ibid.

[54] "PatentsView: About", US Patent and Trademark Office, http://www.patentsview.org/web, (accessed April 14, 2016).

[55] OECD.

[56] "U.S. Environmental Protection Agency Open Government Data Quality Plan 1.0" U.S. Environmental Protection Agency, May 18, 2010, https://www.epa.gov/sites/production/files/documents/opengov_data_quality_plan.pdf, (accessed March 31, 2016).

[57] W3C, "Data on the Web Best Practices: Data Quality Vocabulary", W3C, Updated on December 17, 2015, https://www.w3.org/TR/vocab-dqv (accessed March 31, 2016).

[58] http://www.usgs.gov/datamanagement/why-dm/lifecycleoverview.php

intelligence community. This process has adapted to the internet age, shifting from a linear model to one which can rapidly engage all five aspects and ensure data is shared in an accurate, timely manner across the 16 intelligence agencies. A partnership between Carnegie Mellon University and Booz Allen Hamilton recently resulted in the Data Management Maturity Model (DMM) designed to enable "organizations to improve data management practices across the full spectrum of their business model."[59]

Questions for Further Consideration

Many elements of data quality need to be considered, and balanced, to find the most pragmatic ways to best meet users' needs. Some of the overarching questions to be considered include the following:

- How can a balance be struck between timely data and the need to ensure certain levels of quality before release?
- Agencies hold large numbers of datasets that are of varying levels of quality. How should they prioritize data for quality improvement?
- What approaches to quality improvement can be shared among agencies?
- How can government best get feedback to help develop the open data ecosystem, improve data quality through crowdsourcing, or explore collaborations with companies and organizations outside government?
- What role should data quality play in managing a world of constant data creation?
- How can agencies better design data collection procedures to create high quality data from the start?

Acknowledgments

The Center for Open Data Enterprise thanks Open Data Partner Microsoft and Open Data Supporter Booz Allen Hamilton for their support of the 2016 Open Data Roundtable Series and their input into the preparation of this briefing paper. The lead researcher and author for this paper is Matthew Rumsey, Research Fellow at the Center for Open Data Enterprise. The Center 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.

Suggestions for Further Reading

Hongwei Zhu, Stuart E. Madnick, Yang W. Lee, Richard Y. Yang, "Data and Information Quality Research: Its Evolution and Future", in "Computing Handbook: Information Systems and Information Technology", CRC Press, edited by Heikki Topi and Allen Tucker, 2014.

[59] "DMM Frequently Asked Questions (FAQs)", CMMI Institute, http://cmmiinstitute.com/resources/dmm-frequently-asked-questions-faqs (accessed April 14, 2016).

OECD Statistics Directorate, “Quality Framework and Guidelines for OECD Statistical Activities”, The Organisation for Ecomomic Co-Operation and Development, January 2011, http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=std/qfs%282011%291&doclanguage=en

Dave Becker, Trish Dunn King, Bill McMullen, Lisa Deifer Lalis, David Bloom, Ali Obaidi, Donna Fickett, “Big Data Quality Case Study: Preliminary Findings, U.S. Army MEDCOM MODS”, The MITRE Corporation, September 2013, http://www.mitre.org/sites/default/files/publications/13-2568.pdf

Shadrock Roberts, Stephanie Grosser, D. Ben Swartley, “Crowdsourcing to Geocode Development Credit Authority Data: A Case Study”, United States Agency for International Development, https://www.usaid.gov/sites/default/files/documents/2151/USAIDCrowdsourcingCaseStudy.pdf

Office of Management and Budget, “Statistical Programs and Standards,” https://www.whitehouse.gov/omb/inforeg_statpolicy