Note: These minutes summarize the discussion that took place at the Government Data Sharing Community of Practice meeting. The summary does not necessarily represent the views of GAO or the organizations that the discussion participants represent.

Government Data Sharing Community of Practice

Panel Discussion on Technological Challenges to Sharing Data: Meeting Minutes

Presented in Conjunction with the MeriTalk Big Data Exchange (BDX)

February 12, 2014, 9:15-10:15 a.m. ET

The City Club of Washington at Columbia Square,
555 13th Street NW, Washington, DC 20004

http://www.gao.gov/aac/gds_community_of_practice/overview

Background

In January 2013, GAO cohosted a forum with the Council of the Inspectors General on Integrity and Efficiency (CIGIE) and the Recovery Accountability and Transparency Board to explore using data analytics—which involve a variety of techniques to analyze and interpret data—to help identify fraud, waste, and abuse in government. Forum participants included representatives from federal, state, and local government agencies as well as the private sector. Through facilitated discussion, forum participants identified a variety of challenges that hinder their abilities to share and use data. Among other challenges discussed, forum participants discussed technological difficulties, such as interoperable data systems and varying data standards, that hinder their abilities to use data from other agencies. A summary of the key themes from the forum is published at http://www.gao.gov/products/GAO-13-680SP. To continue the dialogue on issues related to coordination and data sharing, GAO formed the Government Data Sharing Community of Practice (CoP).

Topic: Technological Challenges to Sharing Data

Moderator: Naba Barkakati, Chief Technologist, U.S. Government Accountability Office

Panelists:

- Eric Hagopian, Senior Solutions Engineer, Novetta Solutions
- Marcel Jemio, Acting Director, Division of Enterprise Architecture, Bureau of the Fiscal Service; Chair, Fiscal Service Data Stewards
- Marcus Louie, Data Solutions Architect, Socrata

Overview

Steve Lord, Managing Director of GAO's Forensic Audits and Investigative Service (FAIS) team, welcomed attendees to the second CoP event, provided a brief background of the CoP, and introduced the panel discussion topic and moderator, Naba Barkakati. Mr. Barkakati then facilitated discussion among the panelists and solicited questions from the audience. Highlights from the discussion are described below.

Note: These minutes summarize the discussion that took place at the Government Data Sharing Community of Practice meeting. The summary does not necessarily represent the views of GAO or the organizations that the discussion participants represent.

Moderated Discussion

Moderator question: How did you come to be involved in data analytics?

Mr. Jemio started the conversation by describing his background at the U.S. Department of Treasury's Bureau of the Fiscal Service (Fiscal Service), a consolidated agency comprised of the former Bureau of the Public Debt and the former Financial Management Service, which has a mission to promote the financial integrity and operational efficiency of the U.S. government. He described the Fiscal Service as an agency that faces great opportunities to use data analytics to enhance government efficiency.

Mr. Hagopian said he got involved in analytics doing Java and Oracle programming in the private sector. He was integrally involved in developing the U.S. Department of Homeland Security's (DHS) Integrated Common Analytical Viewer (iCAV), a geospatial visualization suite of tools that integrates commercial and government data and helps establish situational and strategic awareness among infrastructure planners and stakeholders. Mr. Hagopian also served as the Analytics Program Manager at U.S. Immigration and Customs Enforcement (ICE). Though he has been involved in data analytics for much of his career, Mr. Hagopian said the goal of using data analytics has always been the same: maximizing the amount of relevant and useful knowledge that can be derived from any given data system. Mr. Hagopian recently became Senior Solutions Engineer at Novetta Solutions, which focuses on data analytics, cyber security, and identity intelligence.

Mr. Louie said that he started using data analytics while working towards his Ph.D. His research routinely required obtaining large amounts of data, cleaning the data to facilitate analysis, and building in entity-resolution capabilities to allow for identifying and linking records from unique entities across multiple sources. Mr. Louie said that analytics then followed him through his career, including positions at a New York City technology startup with a mission to help organizations transform large amounts of raw data into information that can be used for decision making, and now as Data Solutions Architect at Socrata, which has a mission to make data accessible and useful. Mr. Louie said that the key principle in all of his experience in using data is to *first* identify a business problem, and *then* determine which data can assist in solving the problem.

Moderator question: Can you give an example of technological challenges that you have faced in the data-sharing process?

Mr. Louie began the conversation by discussing his experience working with a federal agency that awarded loan guarantees. One side of the agency monitored loan performance, while the other side was responsible for making deals with banks. This agency had two main technological challenges: (1) the monitoring team maintained some standard data that they used to generate reports to track loan performance; however, the reports were generated as HTML tables and then manually extracted into Excel spreadsheets, making analysis cumbersome and time-consuming, and (2) the banking team maintained inconsistent data based on what each individual thought was important for his or her cases, and the data were

maintained in several different ways (e.g., Excel, Access, Google Docs), making it difficult to easily integrate the data to conduct any useful analysis. Mr. Louie and his team worked with the agency to first identify a core set of business questions that data could help answer. They then made changes to the agency's processes to standardize the data that were collected and integrate the data-collection systems from each of the two teams to make it easier to analyze and automate data from both teams to answer the core business questions.

Mr. Hagopian gave an example from his time at ICE. At the time, ICE used over 160 separate systems that did different things with a lot of the same data. Mr. Hagopian and his team did walkthroughs of each system to better understand their purpose and what data they collected. They then built a new system to consolidate and standardize the data from these disparate systems and deployed it across the agency.

Mr. Jemio said that there is an "invisible gorilla" when it comes to big data analytics—there are problem indicators that should be incredibly obvious yet we are missing them. The data exist to identify these problems, but we need to find ways to connect the dots. Mr. Jemio said that we need to employ machines and algorithms to start to translate and find patterns in data that humans cannot easily understand, to help us see the invisible gorilla. Mr. Jemio seconded Mr. Louie's comment that this process, and any new analytics endeavor, should start with identifying core business questions that data can help answer, and *then* finding ways to employ data to answer those questions.

Moderator question: What needs to happen and what can government and industry do to facilitate data sharing?

Mr. Louie highlighted the importance of starting with the end goal in mind by identifying the business questions that need to be answered and then ways to devise solutions to answering these questions using and sharing data. Mr. Louie also commented on the buzzword "big data," and reminded the audience that not all critical data are big, and a lot of analytics can be done with minimal processing, so agencies should keep this in mind when identifying their business questions.[1]

Mr. Jemio said that agencies need to focus on opening their data to other agencies and the public in order to make data transparent, thereby making it easier for everyone to connect the dots. When opening and sharing data, it is also critical to make available any associated metadata—data that describe and provide information about other data. Mr. Jemio described the primary data as the "content," and the associated metadata as the "context." Because agencies and organizations can follow varying data standards, these metadata are needed to accurately interpret and understand the primary data. If we focus on open data and metadata, there can be a dramatic shift in thinking, according to Mr. Jemio.

[1] "Big data" is a phrase used to describe a massive volume of both structured and unstructured data that are so large that they are difficult to process using traditional database and software techniques.

In addition to stressing the importance of open data and metadata, Mr. Jemio also discussed the two kinds of people that tend to be involved in developing data-analytics programs, each with their own strengths and optimal roles: (1) business people who can identify business questions that data can help answer, and (2), technology people who understand data and how data can be applied to answer business questions. His experience shows that business people should lead the development of analytics programs because they have the ability identify the core questions that data can be used to answer. Although not leading the effort, the technology people would play a key role in identifying appropriate data and tools and applying them to address the business questions.

Moderator question: What are some specific things people can do to facilitate data sharing?

Mr. Jemio told the audience that agencies must make sure the data that they release are easy to understand and are of high quality. The quality of the data an agency releases reveals what that agency values regarding data quality and transparency. If an agency publically releases cryptic and difficult-to-understand data, it makes it clear that these things are not priorities, according to Mr. Jemio. He cited the varying levels of quality of data available on Data.gov as an example. While some of the data are of high quality, some of the data are difficult to understand, especially for someone who is not a subject-matter expert. Cryptic data impede the synthesis of multiple data sets, which is key to allowing individuals to connect the dots and see where problems lie.

Mr. Hagopian said that agencies should work to overcome challenges in three key areas to facilitate data sharing:

1. People: Mr. Hagopian has observed that data managers have become more territorial over their data in recent years, perhaps as a result of shrinking budgets. Territorialism and shared and open data are not compatible. The importance of open data should be made clear to data managers by senior agency leadership.

2. Policy: Mr. Hagopian said that in many offices, policies place significant documentation and planning tasks with technology staff whose skills may not be well-aligned with these duties. As a result, documentation and plans may not get at the true business problems. Mr. Hagopian suggested that shifting these responsibilities to business staff would allow technology staff to dedicate more time to technology-related tasks and would allow business staff to better position analytics programs to help the office to address business problems and goals.

3. Technology: Mr. Hagopian suggested that audience members explore the use of NoSQL. NoSQL, also called Not Only SQL, is an approach to data management and database design useful for very large sets of data that seeks to remedy some data-performance issues that common relational databases were not designed to address. NoSQL is especially useful when working with unstructured data that do not fall neatly in the structured fields of a standard database.

Mr. Louie discussed another type of cultural challenge common in government: reluctance to share or open data for fear of releasing private information. He encouraged audience members

to evaluate their data and find subsets of information that could be released without risk of releasing private information. He suggested that agencies start small and begin their open-data initiatives by making available very-low-risk data, which helps internal stakeholders become more familiar and comfortable with open data. Such an approach can lead to a "snowball effect," where, as stakeholders become more comfortable, they become more willing to support larger data initiatives. We need to change the perception that open data needs to be all-or-nothing, said Mr. Louie.

Moderator question: How do you see the technological landscape changing as we move forward, and what challenges do you foresee?

Mr. Louie said that he has been observing two significant trends that positively affect how agencies use data. First, data literacy is increasing. Increased data literacy among business staff helps these staff to better understand how data can be used to help achieve strategic goals. This enhanced literacy helps agencies break down organizational silos to allow for broader use of data throughout an organization, rather than confined to specialized teams. Second, the barriers to entry for data analytics are decreasing, allowing more organizations to access data and conduct analytics to make something useful of the information. Lower barriers of entry also make it easier for the public to take open data and use it in important and meaningful ways.

Mr. Hagopian seconded Mr. Louie's observation that barriers to entry are decreasing, allowing more people to access and use data. He said that open-source is the next frontier, and that it will reduce these barriers even more.[2] Although it is becoming less expensive to access and use data, significant challenges remain, including organizations employing various different data standards, resulting in difficulties in interpreting and understand each other's data.

Mr. Jemio agreed with Mr. Hagopian that a lack of uniform data standardization continues to be a significant challenge that hinders data interoperability. Mr. Jemio recalled a work assignment where he was tasked with assessing the interoperability of data systems in one federal agency. He found that the agency had 10 different data systems collecting data in different ways. Sharing information within the agency required constantly translating information between databases, which was cumbersome and expensive. Agencies often face similar challenges working with their own data systems, and these challenges can grow when agencies share data with each other and open their data for public consumption. Mr. Jemio suggested that perhaps the government should participate in global forums on data standards and work to conform to industry data standards. Looking out to the next 10 years, Mr. Jemio expects that the government will open and share data increasingly over time, and that the public will continue to embrace data and demand that it be made available, making powerful synthesis of the growing number of datasets possible.

[2] "Open source" refers to a program in which the source code is available to the general public for use and modification from its original design.

Note: These minutes summarize the discussion that took place at the Government Data Sharing Community of Practice meeting. The summary does not necessarily represent the views of GAO or the organizations that the discussion participants represent.

Audience Questions and Answers

Audience question: Many people assume metadata are large and complicated datasets when they are generally small, simple sets of information that are critical to helping people understand and use the primary data associated with the metadata. Although critical, metadata are not frequently shared, and the importance of metadata is not emphasized. How can we better encourage the use and sharing of metadata?

Mr. Jemio responded that using strategic messaging to executive leadership is important in promoting the use and sharing of metadata. Because executive leaders are usually business people and not technology people, it is important not to overcomplicate the matter. Because of the misconception that metadata are overly complex, one should not start the conversation by trying to explain what metadata are. In promoting the use and sharing of metadata to executive leadership, Mr. Jemio suggested discussing the business problem that metadata could help address, and then explaining at a high level how metadata can assist.

Mr. Hagopian commented that metadata should be integrated as a key internal discussion topic when discussing big data, as they are highly related and both important. These conversations can assist in emphasizing the importance of using and sharing metadata.

Audience question: How do we balance the need to open and share data with the "mosaic effect"—the effect occurs when data from one source alone is not identifiable but when coupled with other available information poses a privacy or security risk?

As an example of the mosaic effect, Mr. Louie discussed a study that was released several years ago that found that it was possible to use information about an individual's place and date of birth, in conjunction with publicly available data, to predict his or her Social Security number. Although the study is a few years old, this and other challenges related to the mosaic effect still exist.

Mr. Jemio said that in his experience, data security is the number-one priority for senior managers in government, and that they are conscious of the risks of the mosaic effect. However, agencies can be overly cautious in trying to mitigate these risks by restricting more data than necessary, which prevents people from doing innovative and impactful things with data. Thorough and comprehensive metadata can assist in mitigating these risks, according to Mr. Jemio. When thorough metadata exist, they allow managers to better evaluate what information can be shared, and which information should not be shared, or should only be shared with certain individuals or entities, because when it is in composite with other information, it can be used to identify individuals.

Mr. Hagopian said that other approaches may assist in safeguarding against the mosaic effect, including processes to strip identifiers and perhaps have certain public data that expires after a certain period.