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 Changing the Culture for Open Data

Meeting Minutes

April 22, 2015, 1:00pm-5:00pm

The Loft, 600 F Street NW, Washington, DC 20004

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

Background on GAO's Government Data Sharing Community of Practice

Federal government agencies face challenges in sharing information, using data analytics, and leveraging resources that could assist them in their programmatic and oversight missions. GAO formed its Government Data Sharing Community of Practice to foster an ongoing dialogue about strategies used to overcome challenges that federal, state, and local government agencies face in trying to share data to fulfill their missions. The Community of Practice was an outcome of a forum GAO hosted with the Council of the Inspectors General on Integrity and Efficiency (CIGIE) and the Recovery Accountability and Transparency Board to explore opportunities to use data analytics to identify and prevent fraud, waste, and abuse in federal government programs.[1] GAO's Government Data Sharing Community of Practice is open to all stakeholders, including those from both the public and private sectors.

Changing the Culture for Open Data

Open data is the official policy of the U.S. government, but legal, technological, and cultural barriers all limit its adoption. In this session of the Government Data Sharing Community of Practice, representatives came together from the Center for Open Data Enterprise, Cyrrus Analytics LLC, the Department of the Interior, the National Parks Service, the Recovery Accountability and Transparency Board, the Data Transparency Coalition, GAO, and Socrata. In moderator-led discussions, participants focused on cultural barriers to the adoption of open data, providing success stories where barriers were overcome, as well as areas where change still needs to occur.

---

[1] A summary of the key themes from the forum is published at http://www.gao.gov/products/GAO-13-680SP. Minutes from previous sessions of GAO's Government Data Sharing Community of Practice are available at http://www.gao.gov/aac/gds_community_of_practice/overview#t=1.

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.

<u>Introductory Remarks</u>

<u>Joah Iannotta, Assistant Director, Forensic Audits and Investigative Service, U.S. Government Accountability Office</u>

Dr. Iannotta thanked the Data Transparency Coalition for its efforts in organizing the panel discussion and introduced the topic of discussion for the panel: how to change agency culture to embrace open data.

In 2013, GAO teamed up with the Recovery Accountability and Transparency Board and CIGIE for a forum on data analytics in law enforcement (see GAO-13-680SP). A key theme to emerge from the forum was the challenge of sharing data, caused by: legal constraints; technological limitations; and cultural resistance.

Today's panel discussion was organized to discuss ways to move forward and overcome the third challenge, cultural resistance. Even when sharing of data is legally allowed, people sometimes choose not to (also known as "data hugging"). With passage of the Digital Accountability and Transparency Act of 2014 (DATA Act), there are legal requirements to provide open data. Here, a distinction must be made between "data sharing"—with a specific user in mind—and "open data," where data are accessible to everyone.

While there are some technical issues to work through, the bigger question is—are people in government going to get on board and provide open data?

<u>Hudson Hollister, Executive Director, Data Transparency Coalition</u>

The Data Transparency Coalition is a trade association that advocates for open data and pursues the passage of legislation that requires it. The Coalition advocated for passage of the DATA Act, which requires the federal government to adopt government-wide standards for financial data and publish that data online. On May 9, 2015, the U.S. Department of the Treasury (Treasury) and the Office of Management and Budget (OMB) will issue guidance on data standards. Initially, changes related to the DATA Act will focus on federal spending data. Financial regulation data will also be standardized and shared. But that is only the beginning—data have to be standardized across the federal government. Moreover, while the law can mandate programmatic change, the culture must also change in order for huge transformation to happen.

The Data Transparency Coalition will be hosting a summit focused on the Data Act on June 9 and 10 in Washington, DC.

<u>Stu Rabinowitz, Director of Federal Markets and Channel Partners, Socrata</u>

Socrata helps governments around the world release their data for internal and external use. Mr. Rabinowitz described a worldwide open data "explosion": in London, data help commuters anticipate bus schedules; in New York, diners have access to a heat map of rat sightings through restaurant inspection data; nongovernmental organizations like the World Bank are providing open spending data. At the

federal level, NASA, the White House, the Centers for Medicare & Medicaid Services, and the Centers for Disease Control are sharing data on topics as diverse as disease outbreaks and open payments.

Finally, cultural shifts are happening in open data. For example, Michael Bloomberg provided funding to help mid-tier cities release open data, while John Hopkins University recently established the Center for Government Excellence to promote open data and data-driven government.

Panel 1: Stories of Cultural Change

Moderator
- Joah Iannotta, Assistant Director, Forensic Audits and Investigative Service, U.S. Government Accountability Office

Panelists
- Rick DeLappe, Recreation One-Stop Program Manager, National Park Service, U.S. Department of Interior
- Carrie Hug, Director of Accountability, Recovery Accountability and Transparency Board
- Camille Calimlim Touton, Counselor to the Assistant Secretary for Water and Science, Department of the Interior

In opening the roundtable discussion, Dr. Iannotta asked the panelists to share open data success stories and to discuss strategies to overcome cultural resistance. Dr. Iannotta asked that they discuss their respective projects in the context of legislative authority to make changes, but also in the context of support from top agency leadership, stakeholders, and "data geeks."

A question-and-answer session followed individual presentations by each panelist.

Camille Calimlim Touton, Counselor to the Assistant Secretary for Water and Science, Department of the Interior

Camille Calimlim Touton is Counselor to Principal Deputy Assistant Secretary for Water and Science Jennifer Gimbal at the Department of the Interior. Previously, Ms. Touton worked for the House Natural Resources Committee, the authorizing committee for the Department of the Interior.

Ms. Touton described the Open Water Data Initiative, part of the President's Climate Data Initiative, as a success story. The Open Water Data Initiative is aimed at integrating fragmented water information that is already being collected by different agencies at several levels of government into a connected, national water data framework. There are nine different federal departments and 27 federal agencies that have water as part of their mission. For each, data were collected in different formats and for different missions, and as a result weren't always interoperable. In order to integrate disparate data sets, a common platform was needed to make the data uniform. To address this technological barrier to sharing data, the Assistant Secretary for Water and Science at the time, Anne Castle, initiated work on building the platform.

While the Open Water Data Initiative has strong support from the administration and Department of the Interior, agencies need to be shown the value of initiative efforts in supporting their agency missions. To

do that, the initiative uses Lean Startup concepts such as "failing fast"—building a minimum viable product, and then showing what can be accomplished through the product using open data.

An example of value added is the California Drought Visualization website. By integrating data scraped from multiple sources, the website is able to tell a story that hasn't been told before, and shows the value of open data. Data sources included the National Oceanic and Atmospheric Administration (NOAA), the United States Geological Survey (USGS), the United States Bureau of Reclamation, and various state government sources.
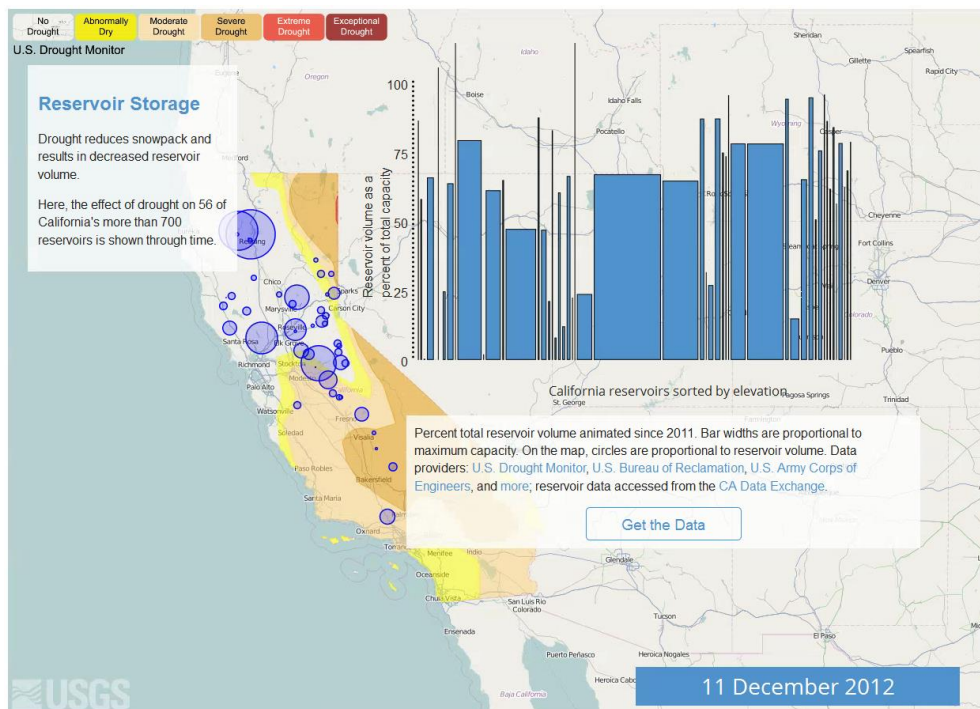
Screenshot from the USGS California Drought Visualization: http://cida.usgs.gov/ca_drought/

The Open Water Data Initiative plans to roll out a visualization of the Lower Colorado River at the end of the year. One end goal of such visualizations is to create an interagency support system at times of drought. Visualizations are also a great example and application of the possibilities with Open Water Data.

In pushing forward a project like this, it is definitely necessary to have "data geeks" who understand the data, but you also need people who are really interested in change.

Rick DeLappe, Recreation One-Stop Program Manager, National Park Service, U.S. Department of Interior

Prior to his work as Program Manager for Recreation One-Stop, Rick DeLappe was a Park Ranger in Yellowstone and Zion, and holds a degree in Botany. Recreation One-Stop provides recreation

information and travel planning for federal lands and recreation areas, including a reservation service for over 3,000 campgrounds, tours, and other recreational activities.

In 2002, OMB pushed for the integration of separate reservation systems maintained by the National Park Service, the U.S. Army Corps of Engineers, and other government entities. While rudimentary, recreation.gov had been aggregating such data starting in the late 1990s, and was decided on as the platform to accomplish OMB's goal. A single contract was awarded in 2006 uniting the agencies' reservation services. The program operated quietly, providing the public with an online and call center–supported reservation service for a number of years. In 2011–2012 the "America's Great Outdoors" report recognized recreation.gov as a place to accomplish the aggregation of reservation data.

The recreation.gov program is managed by a representative from each of seven participating agencies: the U.S. Army Corps of Engineers, Forest Service, National Park Service, Bureau of Land Management, Bureau of Reclamation, Fish and Wildlife Service, and National Archives. These representatives meet regularly. When the Open Data Executive Order was released, the agencies enhanced the Recreation Information Database (RIDB) and began to provide data in more formats, including JSON and CSV formats, so that more could be done with the data beyond a single XML download of data. Other improvements made include the addition of a parsing filter in the data, so that one can pull—for example—only Oregon data, or only National Park Service data. Recently, the Recreation One-Stop Program has redirected funding to develop front- and back-end Application Program Interfaces (API) for the site. These new APIs were released in March.

For advocates of open data, the number-one tool at your disposal is to demonstrate the power that open data gives you. On April 11 and 12, the U.S. Department of Agriculture (USDA) hosted a "hackathon" with the Department of the Interior to expose newly developed APIs to the public. Eleven different products that came out of open data were presented at the event; it was eye-opening to see what could be accomplished in only 48 hours.

Fixing data errors is also of high importance. One developer pulled together a world map of One-Stop data and found campgrounds incorrectly geocoded in the Atlantic Ocean, amongst other examples. To support the reservation and travel planning services on recreation.gov, the program recognized that getting accurate and up-to-date data would continue to be an ongoing challenge. The Federal Interagency Council on Outdoor Recreation made a push to have agencies provide data; however, agency data warehouses and content-management systems are not quite able to supply all the data needed. Collecting and validating data is a collateral duty for most agency personnel and it may often be prioritized behind other critical agency functions. The next step for recreation.gov is to find data gaps, and fix them.

Carrie Hug, Director of Accountability, Recovery Accountability and Transparency Board

Carrie Hug is the Director of Accountability at the Recovery Accountability and Transparency Board (Recovery Board). Initially, the Recovery Board collected and displayed award and expenditure data on American Recovery and Reinvestment Act funds on recovery.gov. After January 2013, the Recovery Board also posted information on Hurricane Sandy funds. Previously, Ms. Hug served as the Branch Chief for Financial Standards and Grants within the Office of Federal Financial Management at OMB. OMB owned two large federal spending data-collection systems—Federal Assistance Awards Data System (FAADS) and USASpending.gov, operated and managed by the Census Bureau and General Services Administration, respectively.

Resistance to transparency within the federal government predates the uniformed, systematic collection of data as we know it. One of the first bills on transparency was the Freedom of Information Act in 1967, which was actually an amendment to a bill from 20 years earlier. There have been several legislative catalysts since then, including:

- the 1982 Consolidated Federal Funds Report Act (CFFRA)—resulted in FAADS;

- the 2006 Federal Funding Accountability and Transparency Act (FFATA)—resulted in USASpending.gov; and

- the 2009 American Recovery and Reinvestment Act (ARRA)—resulted in recovery.gov.

Although each system provided the public with open data, a certain amount of change and resistance came with the deployment of FAADS, USASpending.gov, and recovery.gov. Change meant the need to allocate or reallocate resources and potentially establish new processes, and a perceived or real increase in agency workload. The implementation of recovery.gov resulted in an increase in work for federal agencies, as well as grant, contract, and loan recipients, because all of these parties were responsible for reporting and assuring data quality. This was a switch from USASpending.gov where data is provided by the agencies.

Maintaining quality is imperative in order to maintain confidence in the data. Good data quality is a measure of success. In the first quarter of ARRA reporting, thousands of recipients successfully submitted data about their awards; however, there were recipients who did not report at all. The Recovery Board determined that more than 4,300 award reports were missing. Additionally, recipients entered data in error; for example, the number of jobs "created" equaled the state's population; or the full award amount had been received, jobs had been created, and the project was completed without the expenditure of any funds.

Prior to the first reporting period, OMB held town-hall events to educate stakeholders on ARRA guidance. After reporting began, Vice President Biden and OMB worked directly with agencies, and agencies worked with their recipients, to correct data. Throughout ARRA reporting, the Recovery Board promoted data quality using a variety of approaches, such as: conducting training sessions, facilitating agency forums and one-on-one meetings with high-risk agencies, and meeting with state officials. Awards not reported by recipients were posted on a "wall of shame" on recovery.gov. The data-collection system was enhanced to add edits and provide an automated data change (ADC) capability. The ADC function allowed corrections to be initiated at multiple levels including at the subrecipient/recipient level, and required approvals from the agencies and senior Recovery Board management. Initiatives to improve data quality were taken seriously and acted upon because of visible support from senior management within the Office of the Vice President, OMB, and the Recovery Board.

A lesson learned, whether trying to reconcile data between federal, state, or local governments, or consolidating financial statements from all of the federal agencies, is that without data standardization, translating, matching, and comparing disparate data can be time consuming and very expensive.

Question and Answer Session

Joah Iannotta: *Ensuring data quality, completeness, and reliability requires a lot of human work and human thought. In your experience, what has helped to convince people to do that work?*

Camille Calimlim Touton: Quality Assurance / Quality Control (QA/QC) standards at USGS in water data are highly respected. As a result, conforming to completeness and data-quality standards wasn't as much of a challenge; integrating data sets was the biggest challenge we faced. We had to recognize the inherent limitations of certain data sets. Some data sets are just built differently, for different purposes— for example, if recording tools were just totally different between agencies. The same issue cropped up with state data; states have their own way of collecting data for various reasons.

Rick DeLappe: One of our biggest challenges is that data originate at the field level—at individual national parks and forests, where staff enters those data, sometimes manually. Naturally, you often hear "do more with less," but when you rely on a manual process, it's a lot of work to keep data updated. This is our biggest challenge, and leadership recognizes that. Another challenge is educating the field staff and other agencies in using automated tools like APIs.

Carrie Hug: Tracking the quality of the data is one of our biggest challenges. You need to interact with your stakeholders. If we saw the same issue over and over, it was usually for two reasons: a misunderstanding of the guidance or confusion on how to use the data-collection system. Reaching out to our users and providing continuing education was essential to improved data. People do not want to give you bad data. It happens for a variety of reasons: maybe the user is trying to push data through outdated software, or they do not have the right personnel because they cannot hire subject-matter experts that can assure data quality, due to budget restraints. You need to consider how to work with users' impediments.

Joah Iannotta: It seems there are two key things to focus on. One, do outreach and education, get smart people in the room. Two, try to make open data easy to accomplish and maintain when you can, and consider whether there is a tech solution to the problem.

Audience Question for Carrie Hug: *Do you see exposing fraud, waste, and abuse as a resistance issue, and if so, what do we do to overcome it? Is there agency resistance to exposing fraud, waste, and abuse?*

Carrie Hug: We [Recovery Board] like open data because it could help us expose fraud. However, if you expose fraud, waste, and abuse, and you publicly post it, fraudsters will find smarter ways to commit fraud. So you will need smart people to continually develop ways to fight fraud. Agencies want to do the right thing and prevent fraud, but sometimes they do not know how. Agencies may have to work harder on front-end fraud prevention, such as installing stronger internal controls and doing more due diligence before awarding contracts. There is some resistance at the agency level to opening data to the public, but our goal was to post the data we received without changing or manipulating it.

Joah Iannotta: *In developing open data initiatives, bringing together diverse data sets may show a limitation in a particular program. How do you convince stakeholders they will not experience a backlash as a result of their participation?*

Camille Calimlim Touton: We recognize that each agency collects agency-specific data to fulfill its mission. For example the Western States Water Council is a partner on the ongoing Lower Colorado

initiative. We are obtaining Water Data Exchange (WaDE) data through the council. Our message is that we are not trying to criticize the quality of the data; we are just sorting through it and improving it. As a result, we are not saying someone is doing it better or worse. For the Open Water Data Initiative, USGS is the standard for QA/QC; since they are a partner, it brings credibility to the initiative.

Rick DeLappe: When you expose poor-quality data, it's a reminder to clean things up. It also brings together a community of people who are willing to comment about problems in the data. By opening your data, if the public doesn't like the way your system works, they're free to take the data and build something else that might better suit their needs.

Audience Question: *Mr. Hollister referenced OMB standards arising from the DATA Act. How significant of a development will the DATA Act standards be compared to efforts in the past?*

Rick DeLappe: In terms of recreation data, partner agencies have started implementing some standards already in anticipation of the OMB guidelines. Agencies are also working with startups to develop the open data out there (OpenTrails is one example) to conform to standards. These private organizations have an idea of how people want to use that data, which is helpful.

Camille Calimlim Touton: Prior to the standards release, the Department of the Interior is already having internal discussions to be ready for it.

Carrie Hug: Developing standards is the right direction to go. We saw this with the ARRA where the same data elements displayed on recovery.gov were adopted and used on state websites for consistent award information at the federal and state levels. However, where there is a lack of standardization, analytics becomes more complex. For example, for ARRA reporting, there were data elements (received/invoiced and expensed) that meant something different in the contracts' and grants' communities. In contracting, items are invoiced and expensed at the same time (one data field), whereas in the grants' world items are received and expensed separately (two data fields). The same data element, "invoiced," was being treated differently.

We often think that data are more standardized than they are because we understand the definition of the field. However, formatting, coding, or mapping critical data elements or a combination of data elements to the wrong data source or value affects data quality as well. For example, the address of an award recipient in Illinois was recorded as Israel because the abbreviation for both is IL. Standardization will not occur overnight, but the DATA Act should move it in the right direction.

Audience Question: *How important is leadership buy-in to a project versus revamping the federal data technology procurement process? It seems that Federal Acquisition Regulation (FAR) and other constraints exclude working with companies that could do the best and most efficient job.*

Rick DeLappe: In relation to recreation.gov, a Request for Proposal (RFP) for a follow-up contract was just developed. We were able to start consulting with the newly created U.S. Digital Service, which helped develop how the contract written. Previously, contracts were not written well. We would award a contract, and then 12 to 18 months later receive a product. We could either live with inevitable bugs or overhaul the product. Our current contract instead creates a process where the contractor builds

something, which we evaluate and provide feedback on. The contractor will then provide improvements until our needs are met.

Carrie Hug: We need acquisition reform, but leadership is more important. Strong sponsorship is essential to making groundbreaking changes timely.

Joah Iannotta: *Who was your biggest cheerleader to keep your respective project going? What skill set did they bring to the table?*

Camille Calimlim Touton: Several cheerleaders were important to the project, including the data geeks; the USGS Center for Integrated Data Analytics put together the California Drought Visualization Project in 2 weeks through a "hackathon." You need cheerleaders at each level, and 100 percent backing from leadership.

Rick DeLappe: The end user is the most important cheerleader. The public innovators and entrepreneurs suggest changes and revisions. When management sees the passion those people have, leadership listens.

Carrie Hug: There were lots of different people involved, including leadership and customers, who had concerns and called to suggest improvements. There were people who actually believed we could do it, and that was key. You need people who stand up for you; GAO was also a cheerleader.

Panel Two: Future Challenges for Cultural Change

Moderator
- Hudson Hollister, Executive Director, Data Transparency Coalition

Panelists
- Rich Beutel, Principal, Cyrrus Analytics LLC
- Joel Gurin, President, Center for Open Data Enterprise
- Jerry Johnston, Ph.D., Geospatial Information Officer, Department of the Interior

Mr. Hollister described the format of the second panel as an open question-and-answer discussion of places where cultural change is necessary, but hasn't happened yet.

Jerry Johnston, Ph.D., Geospatial Information Officer, Department of the Interior

Mr. Johnston described himself as having two roles at the Department of the Interior. One, he is a promoter and enabler of standards and policies for mapping communities across the department, where he is working on putting mapping infrastructure in place. Second, he facilitates interagency coordination and support. The Department of the Interior has been chair of the Federal Geographic Data Committee for over 20 years. The committee has historically been composed of civilian agencies, but now the Department of Defense and other military agencies are getting involved. Mr. Johnston's primary committee responsibility is encouraging cross-agency collaboration on standards and sharing of technology enterprises.

Joel Gurin, President, Center for Open Data Enterprise

Mr. Gurin is a former federal employee who founded the Center for Open Data Enterprise, which is currently a small team of four employees and two interns. The Center created the Open Data 500, a study of businesses that use open government data, which is housed at Govlab at New York University.

During creation of Open Data 500, the center found that federal data providers and data consumers were talking past each other, or were not talking at all. As a result, the center began to host "open data round tables" to facilitate dialogue between providers and users. Past participants have included the Department of Commerce, USDA, the Patent and Trademark Office, the Department of Transportation, the Department of Energy, the Department of Veterans Affairs, and others.

The mission of the center is to maximize availability of open data and maximize the usefulness of that data by bringing together users and providers to create "demand-driven open data." A lot of the work on open data has been from a "supply-side" approach. It is important to get data "out there." but to make data effective, it is also important to talk to potential users; sometimes agencies see open data policies as an unfunded mandate, without additional staff or budget. The center estimates that about 20 percent of the public data holds 80 percent of the value. As a result, the government needs to prioritize sharing of this data, in a way that meets the needs of data consumers. There is also a massive amount of data in legacy systems, and releasing that data will require a commitment to an agile development approach. The center has found "change agents" in agencies that it is working with to spur agile development.

Rich Beutel, Principal, Cyrrus Analytics LLC

Before founding Cyrrus Analytics, Mr. Beutel worked alongside Hudson Hollister at the House Committee on Oversight and Government Reform, where he helped developed the Federal Acquisition Reform Act and specialized in government acquisition policy. Mr. Beutel is a government acquisition policy specialist. He is currently working with OMB on Federal Information Technology Acquisition Reform Act (FITARA) implementation. Government has traditionally done a poor job in acquiring and deploying technology, and needs to move to an agile and iterative acquisition process. The government must also provide new resources in a way that is affordable to end users.

Question and Answer Session

Hudson Hollister: *If you could snap your fingers and change culture, whose culture would you change?*

Jerry Johnston: I have been involved in publishing open data for a long time, dating back to my involvement with the Geospatial One-Stop Portal Initiative, which predated data.gov. We produced the first national database of metadata about 15 years ago. Because geospatial was a niche technology, government agencies in that niche spent 20 years talking amongst themselves. We haven't traditionally asked downstream consumers what they want; in government, there were statutory and mission-related reasons to publish data, while any end-user benefit was seen as a side benefit. That has changed over the

last several years because geospatial is no longer a niche interest; as a result, expectations are much different.

Currently, 80,000 geospatial data sets are available. Of that data, the federal geographic data community is asking: What are the high-value, "nationally significant" geospatial data sets? Roughly 200 data sets were identified as most important. For that data, we are trying to go beyond metadata and are moving to establishing metrics for managing the data sets, such as data content standards and data models. We are also building infrastructure to publish data sets. As a next step, we need to have government geospatial data providers come into the open data community in discussions, and use those discussions as an opportunity to learn what is valuable to consumers.

Joel Gurin: I will answer a different question: what kinds of thought patterns need to change? We need to move from process-driven to impact-driven thinking.

Process requirements can't be ignored, but process is not the core mission; government should focus resources on areas where they will have the biggest impact. Agencies need to think outside the box. As an example, NOAA management launched an effort to take the "ocean" of NOAA data and put it in the cloud. To create the most impact, NOAA brought in private partners to identify data that are most important and make them available to users. The cultural change to impact-driven thinking can drive creative and effective thinking. It can answer the question: How do you get the data into the right hands?

Rich Beutel: Evaluation of program success needs to be metrics-based. Aversion to open data occurs because "sunshine" can kill programs; if a program can't be justified based on metrics, then the program is at risk.

Currently, the government has a compliance culture. Agencies focus on compliance with process, not actual outcome. As an example, the House Committee on Oversight and Government Reform led an effort to ferret out waste fraud and abuse in a $2.9 billion, taxpayer-funded program meant to provide employment opportunities to the severely disabled. Funding went to companies where 75 percent of labor hours were performed by people who are severely disabled. Under the program, people self-certified as severely disabled, and funds were going to 10 large companies. In contrast to this self-certification model, the Social Security Administration validates the disability of people within their database. They have a process in place whereby judges ascertain disability.

For the program we were investigating, we thought there should be significant correlation between people who are on Social Security Administration disability and people who self-certify as disabled through the program. We estimated that about 80 percent overlap should occur between the two populations. The agency responsible for managing the employment program requested data from the Social Security Administration, but the Social Security Administration denied its request, saying that the administration couldn't provide it for legal reasons. The House Oversight Committee was able to obtain comparable data through other means, and found that approximately 7 percent of individuals self-certifying as disabled were actually on SSA disability, which is an audit red flag.

Audience Question: *How did the committee get to the 7 percent figure?*

Rich Beutel: Because of the influx of companies with self-certified disabled people, legitimate companies with employees who were actually disabled were being crowded out of the program. These employees and companies voluntarily provided the data.

Audience Question: *A recent article by Ben Casselman at fivethirtyeight.com argues that a lot of barriers to big data are legal, not cultural.[2] Do barriers to open data need to be looked at from a policy and legal perspective?*

Joel Gurin: Legal constraints affect agency action. There are lots of guidelines that sometimes limit the ability to collect and clean data. Patent and Trademark Office data are one example: the office publishes incomplete data, but it operates under strict guidelines about what it can consider official patent data. As a result, companies that use the data spend a lot of time cleaning the data. One company has 15 percent of employee hours dedicated to cleaning the agency data. Data quality is the elephant in the room. In open data, there may be platforms that provide access to data, but if those data are flawed, they may not be useful. The data-quality problem is the most important and difficult problem to solve. It's going to take some creative public–private collaboration to overcome; third-party contributions are necessary given the rules agencies operate under.

Jerry Johnston: Sometimes there are good reasons to restrict data access. For example, the Census Bureau does not allow exposure of data for privacy reasons. However, the bureau also has comprehensive GPS location data. Those data could be easily anonymized and shared—it is information you could theoretically obtain driving down the street. However, the agency cannot legally share the data. As a result, Google and other private companies spend millions of dollars to recreate what the agency already has.

Hudson Hollister: *In the face of legal constraints, what makes you optimistic?*

Jerry Johnston: The pace of technology has made it easier to share data. Before, we made the end user do the work to make the data useful. Technology is now cheap, easily understood, and is an enabler to sharing. Data quality is still an issue, but putting data out there is the best way to find out what needs to be fixed. For example, the Environmental Protection Agency's Toxics Release Inventory data were published openly, and there were many data points geocoded in Outer Mongolia. Sunshine can be a disinfectant of bad data.

Joel Gurin: First, the quality and commitment of people in the federal government, who have a real desire for the data to be used, is a reason for optimism. There's a cultural change that has been key to that—you need a few people who are change agents.

Second, roundtables we hold generate excitement and conversations between people who didn't know they wanted to talk to each other. Participants realize they have common interests and concerns, and discussion results in action, like identifying legacy systems' data that no one is responsible for and that needs a steward.

---

[2] "Big Government is Getting in the Way of Big Data," http://fivethirtyeight.com/features/big-government-is-getting-in-the-way-of-big-data/.

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.

Rich Beutel: There are a number of reasons for optimism, including the Presidential Innovation Fellowship, the United States Digital Services, and 18F. It is important to bring smart young people into the government with a different paradigm. We should amend Title 5 to bring in talented people faster and raise skill sets in agencies.

Agency Chief Information Officers should have authority over their own budgets. That will bring in higher-quality Chief Information Officers, who can help push transparency and data openness.

Hudson Hollister: In regard to the hiring process, that is why 18F is bringing people in under 2-year fellowships—because the federal hiring process is just too slow.

Audience Question: *Some scientists feel that data are "their baby," and are afraid to share them. Do you see concrete ways to carve out a safe harbor to have discussions like the one we're having today? The Data Quality Act states that the federal government can't release data unless data quality is certified. Does that hold back the release of data?*

Rich Beutel: The Federal Procurement Data System–Next Generation (FPDS-NG) is an example of low-quality open data. The system purports to report government procurement. But the data is so bad, private firms like Deltek have set up their own systems that provide more accurate data. This impacts the government's ability to do its job. In cases like this, discussions need to occur, and silos need to be broken down.

Joel Gurin: How much is the Data Quality Act front and center in people's minds, or is the problem really "data hugging"? People don't seem to be complaining about the act. But we definitely need a cultural and legal framework that: (a) makes it ok to release data that are less than perfect and (b) fosters public–private partnerships to quickly improve data quality thereafter.

Jerry Johnston: In working to move geospatial data to data.gov, because of the Data Quality Act, OMB's Office of Information and Regulatory Affairs did not allow state and local data sets to be combined with federal data. The Department of the Interior worked closely with OMB, and eventually that policy changed.

On a related note, GAO recently made a suggestion to Congress to look at Title 13 as a potential barrier—Title 13 deals with how the census collects and shares data. We should talk about barriers like this in public forums so that they get on everybody's radar.

Hudson Hollister: *There is a synergy between the legislative change from the top and the cultural change that bubbles up from the bottom. Have you ever seen a legislative mandate that touches off cultural change?*

Joel Gurin: I have high hopes for the DATA Act to touch off change. There are a number of examples where Congress has a clear opportunity to greatly enhance agencies' ability to provide open data. For example, at the U.S. Patent and Trademark Office, mandating electronic filing of patents would remove barriers to maintaining accurate data.

Another example is the Occupational Information Network (O*NET), which is used by employers and job hunters to match people to a job. Its job descriptions are out of date. O*NET is updated on a three-year cycle. That is restrictive because of the pace of new occupations being created. Thirty laws require O*NET to be used in federal programs, and O*NET could benefit from new approaches.

A final example is USDA's Common Land Unit Data. The Farm Act (2008) resulted in the data being released to the public. The Farm Act was revised, and the data were pulled back 3 weeks later. Now, anyone who wishes to use the data can only get records that are years old. These data could easily be opened up again by legislation, but at this point the agency's hands are tied.

Jerry Johnston: Another example was the Pesticides Regulation program at the Environmental Protection Agency. USDA wouldn't give relevant data to the Environmental Protection Agency, even though the government had invested millions of dollars in the data set. Sharing the data could have benefited the nation.

The DATA Act is driving change in the Department of the Interior, where a team is currently working on implementation planning. The team is looking at questions like: How do we define a program? What is the definition of an obligation? That work has been positive.

Another important piece of legislation on the horizon is the Geospatial Data Act (2015), which will coordinate funding amongst federal agencies and state partners.

Audience Question: *The Cyber bill contains a provision with broad language about restricting the release of computer-generated information. Will that hinder open data?*

Jerry Johnston: I haven't read the legislation. In general, we are going to need to find a "sweet spot"—protections are needed for security, and that is a tradeoff with accessibility that we might want to accept.

Joel Gurin: I haven't read the legislation. There are individual privacy issues in cyber security, national security, and health data. Health "Datapoloozas" are conferences that address the use of open data in health care. Initially, these tended to be about data management. But at the latest one, people were asking: how do we take zillions of patient records, de-identify them, and then use that incredibly valuable data set to see treatment effects on different people. There is no consensus as to whether de-identification can be done reliably or not, and there is a tradeoff of value versus privacy.

Jerry Johnston: Another example is drones and privacy. There's potential for public-lands data collection using drones, and the Department of the Interior has approximately 300 drones. But the current legal landscape still needs to develop to set conditions around how to utilize drones without violating privacy. In some areas, technology has gotten way in front of privacy policy issues.

Audience Question: *In regard to the discussion above about combining local, state, and federal mapping data: mapping data is expensive. Is there any way for the Department of the Interior to coordinate with states to collect the data to prevent duplication of efforts?*

Jerry Johnston: The discussion earlier was about challenges in aggregating the data, not duplication of data. A GAO report on duplication in mapping data found that duplication was not a big problem.

However, we could be better at reusing IT on various projects. We also need a shared geospatial platform where partners can publish data, license IT, and work together more effectively.

Audience Question: *Are federal agencies making human resource investments to get value from data? Are enough hours spent on analysis and data management?*

Joel Gurin: Efforts in the federal government are a mixed bag. Good examples of progress are 18F, and the fact that more agencies have a Chief Data Officer. The problem is that the Chief Data Officer is often given no staff and no budget. With the coming change in the presidential administration, the challenge is to institutionalize progress and make gains a durable part of federal government.

Jerry Johnston: In geospatial, there is no office responsible for data management. There is a group of people across agencies that take on responsibilities, but it is not their "day job."

Audience Question: *We have seen issues where agencies don't want to share data, even within components. How do you deal with agencies that won't share data?*

Jerry Johnston: Inside different components of the Department of the Interior and partnering agencies, people are now more open to sharing data. From a technological and policy standpoint, sharing has gotten easier, but there are still stove pipes. For example, publication credit is a big issue. USGS is a science organization where people get promotions based on publications. As a result, individuals may not want to share their data. Academic institutions and journal publishers have made progress on this typing of sharing barrier, but it is still an issue.

Joel Gurin: We have to get better at identifying and activating the demand side of the equation. Agencies often see more risk in releasing data than holding it back. However, where there is more demand for data, and more ability to obtain it through the Freedom of Information Act (FOIA), there is more risk in holding data back—because agencies want to be seen as open.

When interacting at roundtables, you can see that agency data people are stove-piped; they are working on the same thing, but have never met. More collaboration within and across agencies, and between private and public entities, is needed.

Question: *Is there always a compelling argument to publish, regardless of data quality?*

Jerry Johnston: With the exception of specific statutory / regulatory requirements for data quality (i.e., toxicology data at the Department of Health and Human Services), always publish. We do need to do deep documentation of fitness for use as we publish data. However, very few people read this fitness for use documentation. We need to accept that people may misuse the data—that's something people will do outside of government, but it doesn't mean we shouldn't publish the data.

Joel Gurin: It is essential to have feedback loops so that if users find an error, they know a point of contact at the agency they can notify, so that the error can be fixed.

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.

Final Comments

Joel Gurin: We should emphasize and encourage people to attend roundtables. The Center for Open Data Enterprise is an independent nonprofit that partners with the federal government free of charge. Please see Opendataenterprise.org.

Jerry Johnston: More forums like this one are important. Some of the points we've discussed about understanding customer needs are where work needs to be done. But the government community is interested in taking the next step.

Hudson Hollister: I'm optimistic—there is a developing community interested in open data as a discipline, not just as a particular agency interest.