CONFERENCE OF EUROPEAN STATISTICIANS WP.5-4

Workshop on Statistical Data Dissemination and Communication 08 June 2017

28-30 June 2017, Geneva, Switzerland

Disseminating Government Data Effectively in the Age of Open Data

Mirvat Sewadeh (US Department of Agriculture) and Jeffrey Sisson (US Census Bureau)

msewadeh@ers.usda.gov

Abstract and Paper

The move towards Open Data presented an unprecedented opportunity to modernize the collection and dissemination of data across the Federal Government. But it also placed federal statistical agencies in a demanding environment that is being reshaped by rapidly changing technology and evolving user expectations. The fragmented approach to the creation of data products and dissemination methods has also created data silos and inconsistencies in standardization that challenge government agencies' ability to modernize dissemination services.

This chapter discusses key challenges facing data dissemination in U.S. Federal Government, and provides examples on how some of the Federal Statistical Agencies were able to overcome these challenges. It also draws on experience of several federal statistical agencies to offer examples of good practice and provide recommendations for developing a modern and effective data dissemination strategy.

Chapter 2: Disseminating Government Data Effectively in the Age of Open Data

Mirvat Sewadeh[1] and Jeffrey Sisson[2]

"*Publicly available statistics from government agencies that are credible, relevant, accurate, and timely are essential for policy makers, individuals, households, businesses, academic institutions, and other organizations to make informed decisions. Even more, the effective operation of a democratic system of government depends on the unhindered flow of statistical information to its citizens***."**

Principles and Practices for a Federal Statistical Agency, Fifth Edition

Abstract

The move towards Open Data presented an unprecedented opportunity to modernize the collection and dissemination of data across the Federal Government. But it also placed federal statistical agencies in a demanding environment that is being reshaped by rapidly changing technology and evolving user expectations. The fragmented approach to the creation of data products and dissemination methods has also created data silos and inconsistencies in standardization that challenge government agencies' ability to modernize dissemination services. This chapter discusses key challenges facing data dissemination in U.S. Federal Government, and provides examples on how some of the Federal Statistical Agencies were able to overcome these challenges. It also draws on experience of several federal statistical agencies to offer examples of good practice and provide recommendations for developing a modern and effective data dissemination strategy.

Keywords: Dissemination, Data, Uncertainty, Strategy

1 Data dissemination in the Federal Government; from Colonial America to Open Data

Dissemination of U.S. official data, defined as "distribution or transmission of information or statistical facts by a public agency to the public" is an old practice that can be traced back to codification of laws in Colonial America. At that time, the rapid growth in the body of laws, and the limited availability of information about these laws, generated confusion and frustration among citizens and legislators and thus compelled lawmakers to start disseminating information about these laws (Tauberer 2014).

But the dissemination of data—narrowly defined as statistical facts- by the U.S. government may have well started with the release of the results of the first census, which took place in August 1790.

Since then, collection and dissemination of data by the U.S. government expanded to cover a large number issues, ranging from economic activities and environmental conditions, to health and education. By the end of 2016, there was a total of 125 agencies whose work involves data

[1] Mirvat Sewadeh is Assistant Director for Communications at The Market and Trade Economics Division the Economic Research Service, USDA

[2] Jeffrey Sisson is Deputy Division Chief, Decennial Information Technology Division Census Bureau

collection or dissemination, of which thirteen are principal statistical agencies that have statistical activities as part of their core mission. The thirteen federal statistical agencies are: the Bureau of Economic Analysis (Department of Commerce); the Bureau of Justice Statistics (Department of Justice); the Bureau of Labor Statistics (Department of Labor); the Bureau of Transportation Statistics (Department of Transportation); the Census Bureau (Department of Commerce); the Economic Research Service (Department of Agriculture); the Energy Information Administration (Department of Energy); the National Agricultural Statistics Service (Department of Agriculture); the National Center for Education Statistics (Department of Education); the National Center for Health Statistics (Department of Health and Human Services); the National Center for Science and Engineering Statistics (National Science Foundation); the Office of Research, Evaluation, and Statistics (Social Security Administration); the Statistics of Income Division (Department of the Treasury); the Microeconomic Surveys Unit, (Board of Directors of the Federal Reserve System); the Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration (Department of Health and Human Services); and the National Animal Health Monitoring System, Animal and Plant Health Inspection Service (Department of Agriculture).

The U.S. Government views statistical data as a public good that provides critical information to public and private decision makers and the public in general. The collection and dissemination of statistical data by Federal agencies are primarily governed by three statutes: The Paperwork Reduction Act, The Information or Data Quality Act, and the Confidential Information Protection and Statistical Efficiency Act.

In addition, several Statistical Policy Directives issued by the Office of Management and Budget (OMB) set minimum requirements for the principle Federal statistical agencies and ensure the quality and coordination of Federal official statistics. Of particular relevance to data dissemination is OMB's Statistical Policy Directive No. 4: Release and Dissemination of Statistical Products Produced by Federal Statistical Agencies; which was issued in March 2008 and sets principles, standards, and guidelines for the presentation and dissemination of statistical information. Among other things, the Directive requires agencies to ensure "equitable and timely access to data that are disseminated to the public" and to release documentation on statistical methodology and statistical frameworks that make it possible for users to verify data validity.

2 Open Data Policy; a New Era in Data Dissemination

The term "open data" refers to the idea that data can or should be made available for free use, reuse and distribution. Advocates of "open data" emphasize the value that would be generated by making data available for sharing and redistribution to various users. (Gurin 2014, and Open Data for Development Network 2016).

With the U.S. Federal Government going online in the late 1990s, there was a growing interest, both within the government and among the public, in making government-generated data publically available on the internet in electronic formats. That interest culminated in the 2013 presidential executive order which, along with a memorandum from the Office of Management and Budget, established the U.S. Government's Open Data Policy. The Policy, which is part of a broad effort to promote the principles of Open Government, required agencies to ensure that new information collected or created supports "interoperability between information systems and dissemination of information to the public, without the need for costly retrofitting." To that end, agencies were required to use machine-readable and open formats for information as it is collected or created and use standards in order to promote data interoperability and openness.

The Open Data Policy presented a unique opportunity to modernize the collection and dissemination of data across the Federal Government. But it also placed federal statistical

agencies in a demanding environment that is being reshaped by rapidly changing technology and evolving user expectations. More specifically, three broad trends have shaped the context of data dissemination: [3]

Trend 1 – Increased Demand for Statistical Information: The declining cost of sensors over the past decade, together with expanding storage capabilities, have allowed continuous gathering and storage of increasingly more granular data. At the same time, the shift of the World Wide Web to Web 3.0, which moved the emphasis of reasoning to data, has facilitated data reuse, and reduced the amount of human processing, allowing the release of large volumes of information and data that is currently inaccessible by making it machine-processable (Bates 2011). The increasing availability of data has fueled public interest in accessing, tracking and analyzing data, generating demand for even more data. Whether it is policy makers using data to help formulate policies, scientists and academics interested in data to advance research, or businesses seeking to leverage information about consumers, users' demand for official data, and data in general, has grown exponentially. Access to data released by the U.S. Census Bureau, through one of the agency's main dissemination systems, for example, has grown by more than 25% per year for the last 5 years. At the same time, the growing importance of data as a driver of private and public decision-making has prompted users to look more critically at statistical information and increased their emphasis on accuracy, completeness and consistency of data released by Federal agencies.

Trend 2 – Rising importance of User-Focused Platforms: technological advances have allowed data retrieval systems to become much more capable of identifying and selecting relevant material from large collections of data in response to users' queries. As a result, the dissemination of data has moved to user-focused and interactive platforms where users can easily find, download, manipulate, use, and re-use the data without restrictions. Interactive graphing and mapping tools have also become increasingly common on data dissemination platforms.

Trend 3 – Evolving User Expectations: As a result of technological developments, consumer's expectations are also changing. New expectations require data providers to be ready to deliver and receive digital information anytime, anywhere and on any device. Use of mobile devices to access data, for example, has risen significantly over the last five years, with some Federal agencies reporting up to one third of their users accessing data on mobile devices. And while user needs and preferences vary across user groups and individuals, users generally expect to easily find, download, manipulate, use, and re-use publicly available data without restrictions. Users also expect to be able to search and retrieve data by using plain language searches, rather than understand the structure of datasets or how data are classified. Recent use trends also show that data users are increasingly interested in interactive features, visualizations and mapping of data, which allow presentation of large amount of information and statistical facts in easily accessible and user-friendly formats

3 New era, new challenges

Responding to the growing demand for data amid a rapidly changing technological environment, often requires Federal agencies to establish new systems, re-engineer or abandon, existing ones, set up new processes, and develop modern platforms for delivering data. (Refer to Box 1-Case study from Center for Disease Control).

[3] These trends were identified by authors based on research and discussions with government officials involved in data dissemination in various federal agencies

Box 1: Health Data Interactive: a data tool with decreasing utility

Kate M. Brett[1]
The National Center for Health Statistics

The National Center for Health Statistics (NCHS) provides statistical information that are used to guide actions and policies to improve the health of the American people. In the late 1990s, NCHS began a data dissemination project using an off the shelf software designed to create customizable data tables using pre-tabulated data. At the time, the software offered new functions, including allowing users to rearrange rows and columns, nest additional characteristics within rows or columns, and hide or show items as desired. The success of the application among users, prompted two other groups within the same NCHS division to launch two data dissemination applications using the same software; one focused on health of older adults and the other targeting asthma data.

In 2011, and in an effort to consolidate the three dissemination tools and reduce the cost of maintaining separate applications, NCHS launched Health Data Interactive (HDI) tool. HDI allowed users to find information for many of the most common health risks and outcome measures. All tables were tabulated by data year, age, race/ethnicity, and sex. Where the data allowed, geography, urban/rural status, and income were also included as descriptive variables. Users could create graphs using the date and download the data to Excel or CSV files for further analysis and presentation. But the tool was difficult to learn and, because it rendered tables with potentially millions of data points, it required a lot of resources to maintain. At the same, newer data dissemination systems provided much more user -friendly tools that are easier to maintain.

As a result, in 20I6, the agency decided to discontinue the HDI tool and explore more advanced data dissemination applications. The decision to terminate HDI was justified in that the tool's interface was not intuitive, the effort to learn how to use the tool did not translate into learning other NCHS tools, and agency leadership wanted to support emerging technology. While it is easy to keep systems simply because they are not broken, the best systems should continually provide more utility over time, stay current in data policy requirements, understand their users' needs, and integrate into the agency data and IT architecture rather than exist as stand-alone tools.

.

[1] Kate M. Brett is Kate Brett is Research Scientist Officer at the National Center for Health Statistics/Centers for Disease Control and Prevention

Changes to data dissemination systems raise many challenges for Federal agencies engaged in data dissemination activities:

Challenge 1- Data Silos: Chief among these challenges is the prevalence of "data system silos" across government agencies and within individual statistical agencies, which has led to the use of different approaches to the development of data products and dissemination systems. Some data silos date back to the times when official data were published only in print, and data integration was not possible for the end user. More recently, data system silos emerged as agencies, or individual programs within an agency, made "isolated" decisions on how to best create, package and disseminate certain data. As a result, many federal agencies, or program areas within a single agency, developed unique data rules, data sets, processes, and procedures to serve specific needs with little, if any, coordination to maximize data congruency. In many cases, the fragmentation of data systems also led to the development of access tools that supported only a few, or even a single, data set.

The prevalence of data silos has many implications for data dissemination efforts across the Federal Government. It often leads to inconsistent metadata (discussed in more detail below), variations in data definitions or aggregation levels across datasets, and differing processes and methods for expressing data quality. Some examples of these differences include:

- **Data definition differences** in what is included in the data collected for a given topic or inconsistency in how data are aggregated. An example is defining income by one agency to include only wages earned, while another agency includes capital gains, supplemental income from Government programs and other types of income in their definition.
- **Variation in definitions of geographical definitions** which can be seen in the way data are tagged to a geographic classification (e.g. rural vs. urban), or in the way physical boundaries are defined. For instance, The U.S. Department of Agriculture definition of "rural" areas, is based on programmatic needs such as the school lunch program, while that of the Census Bureau uses a geographical measure.
- **Data Quality**: Agencies and programs within agencies often differ in both the measure of error they publish (margins of error, coefficients of variance, etc) and the methodology they use to calculate the measures. These differences can be confusing to end users, especially those that aren't as familiar with survey methodologies.

Box 2: U.S. Census Bureau, Moving Beyond Data Silos

Jeffrey Sisson

The Census Bureau is one of the largest providers of demographic and economic census and survey data in the Federal Statistical system. The Bureau completes over 60 censuses or surveys as part of its portfolio or on behalf of other agencies. The existence of this number of surveys often leads to the mergence data silos, with significant implications for processing and dissemination of data for the Census Bureau. The Census Bureau currently has over 50 different dissemination tools or methods for end users to get access to data, ranging from large, sophisticated tools like American FactFinder (the Census Bureau's major dissemination system, https://factfinder.census.gov) and the Bureau's API, to simple tools built to disseminate one data set in spreadsheet formats.

As the Census Bureau began efforts to ease the complexity of getting data for end users, the existence of data silos has complicated these efforts. The first American FactFinder was designed to bring together the largest, and most significant and popular datasets released by the Bureau s and provide a single access point to users. However, the differences in all of these data sets due to the variations in definitions of variables and inconsistency in metadata, meant that once the various data sets are linked to the FactFinder interface, users had to select a single dataset to ease the navigation experience.

Current efforts by the Census Bureau continue to improve users' experience. The second version of American FactFinder gave users the ability to search by plain language topics or by geographies. It also offered different navigational paths to satisfy different user groups and there needs (discussed in more detail below). However, the data silos still existed internally and there was significant behind-the-scenes processing and mapping of data to "hide" those silos from end users. The Census Bureau has recently started efforts to break down these silos as much as possible in conjunction with the development of a new data platform. While this is a promising effort, it will take many years to reach the point where data and systems can be truly integrated and the majority of the data silos are gone. The effort is also mainly focused on the integration of data that are being collected moving forward. It would require a significant effort and resources to integrate historic data, which is unlikely to occur under current budget constraints.

Challenge 2 – Lack of Metadata Standards: Another challenge impacting data dissemination efforts in the Federal government, is lack of standards governing metadata (i.e. statistical and other information describing the data). Metadata serves many important functions including helping users locate the data, defining the content of the data and allowing software and machines to store, exchange and process information. Therefore, it is critical for sharing, querying, and understanding statistical data, especially in the context of the semantic web. But for metadata to be effective, it has to be structured and consistent across its various sources in the Federal government. The inconsistency of metadata hinders Interoperability and Communication Technology (ICT) systems coherence within federal agencies, and ultimately undermines effective dissemination of data to users. (Refer to text box 3-Case study from

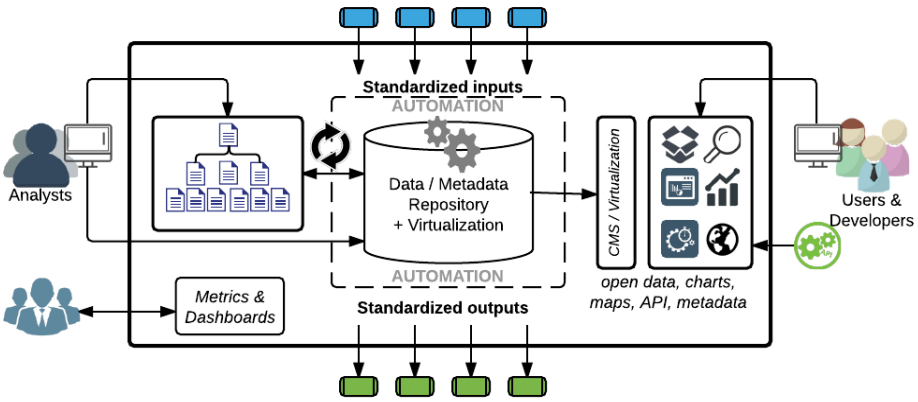
Box 3: Using Metadata to Modernize Operations at the Bureau of Justice Statistics (BJS)
Timothy Kearley[1]

The mission of the Bureau of Justice Statistics (BJS) is to inform criminal justice policy and practice by disseminating accurate, timely, and relevant information on crime and the administration of justice. BJS maintains over three dozen major statistical series designed to cover every stage of the American criminal and civil justice system. BJS disseminates the statistical information produced in the form of publications, data analysis tools, data tables, archived micro data, and related documentation.
BJS must improve public access to an increasing volume of data and a growing public demand for information in a variety of new formats. This must be done while adhering to strict data quality and security protocols that ensure that we are seen as a trusted, authoritative source of official information. As a result, BJS is gradually shifting away from the management of files and documents toward the development of centralized databases that store agency data, metadata, and knowledge at a granular level. One of the foundational components of this modernization effort has been to establish a consistent and comprehensive approach to managing our metadata.

Establishing a Metadata-Driven Architecture

The cross-cutting nature of metadata creates interdependencies across an organization. For example, the same metadata used to ensure coherence and consistency when information is exchanged between BJS and an external facility may also be used within metadata-driven dissemination tools to allow end users to reliably search, navigate, and interpret that information on our website. The same XML-based metadata used to describe a BJS questionnaire can be accessed to automatically produce online forms used to administer that questionnaire.

Figure 1 – Metadata-Driven Enterprise Architecture



Source: Bureau of Justice Statistics (BJS)

Department of Justice).

Box 3 - Continued

A few of the key challenges driving the transition to a metadata-driven architecture at BJS are presented below along with related solutions that leverage metadata:

Challenge: Need to securely store and manage a large volume of data with variant attributes and security requirements.

> **Solutions**: Transition to a single database of record for all official statistics and maintain an accurate, connected metadata repository that includes information about data security requirements. BJS can then quickly respond to data calls, change numbers in one place and have those changes apply to many downstream access points, and apply new data security controls across all collections at once.

Challenge: Need to quickly produce and disseminate an array of products and data with limited staff.

> **Solutions**: Establish tools and processes that use metadata to automate transformations and the production of reports, tables and products. Publish products to the web on demand from the original underlying source data. Within online data access tools, use metadata about population sizes and aggregation levels to automatically suppress access when confidentiality protection thresholds are surpassed by user selections.

Challenge: Need to make it easier for internal analysts, external users, and machines to find, understand, and access information in multiple formats.

> **Solutions**: Tag online content using a taxonomy drawn directly from the collection metadata to allow more accurate search and discovery for users and other systems. Use metadata and data virtualization tools to generate data cubes and longitudinal datasets that enhance data mining and visualization capabilities for analysts.

Many organizational changes are required to establish and sustain use of a modernized, metadata-driven architecture. The work required cannot be taken on all at once and must be sequenced to account for project interdependencies. BJS is adopting a "service-oriented" approach that makes use of existing technologies and capabilities whenever possible as modernization efforts proceed.

Some of the key changes being undertaken include the following:

- Adopt established metadata standards and best practices
- Assign and define key roles (e.g. data stewards and curators)
- Establish standard naming conventions and business rules
- Develop and maintain a reference metadata bank at the agency level that defines organizational units, products, variables, concepts, classifications, and controlled vocabularies
- Implement new technologies for data and metadata management
- Establish methods and systems for assigning unique identifiers
- Establish a collaborative space for sharing and accessing all this information
- Set new policies for metadata management and data processing
- Train staff to gradually build up in house capability and capacity
- Automate data processing and publishing processes

By establishing a metadata-driven architecture and tools, automation will be used to increase efficiency. Access to information for both our internal and external data users will be enhanced by ensuring that data is easier to find and can be presented in whatever manner is most useful to a broad audience.

[1] Timothy Kealey is Chief of Technology and Data Management at the Bureau of Justice Statistics

Challenge 3 - Understanding User Needs: As discussed earlier, expectations of data users are constantly evolving with changes in technology. Users expect to get data and information tailored to their needs and to be able to access that information seamlessly on any device at any time. Federal agencies providing statistical and other data have to overcome the challenges of both understanding customer needs and responding to the ongoing changes in these needs and expectations. For many agencies, this requires a shift in "culture" to meet users' needs' with relation to the type of data release and dissemination methods. For many agencies this shift could mean implementing new procedures and processes to capture feedback from users and incorporate this input into decision on data product changes.

Challenge 4: Communicating Uncertainty. Data errors can occur during collection or processing of data, arise from errors in sampling or coverage, or from nonresponse, among other sources. In addition, there is an inherent uncertainty that occurs solely as a result of using a sample from a population, instead of conducting a census (complete enumeration) of the population in question. U.S. Federal agencies often report official statistics as point estimates, without providing measures of error. As a result, users of the data may incorrectly view these statistics as error free or may incorrectly underestimate error magnitudes (Manski 2015). Communicating uncertainty about official data is particularly important given the impact that data released by the government can have on markets, policy-making and the public in general. Yet, different users vary in their understanding of measures of uncertainty and, therefore, it can be difficult for federal agencies to effectively educate end users on how to interpret or appropriately use the data. Novice users tend to be less informed about uncertainty in data and more likely to ignore any information provided about the measures of error around an estimate. However, more advanced users relying on data to make critical decisions need to understand and make informed decisions on the use of the data and any associated measures of uncertainty. (Refer to

Box 4: Communicating Uncertainty in Official Statistics

Michael D. Levi[1]

The Bureau of Labor Statistics (BLS) of the U.S. Department of Labor is the principal Federal agency responsible for measuring labor market activity, working conditions, and price changes in the U.S. economy. Maintaining the credibility of BLS' published data is essential to the Bureau's mission and critical to data users. However, survey results, come with some uncertainty and data users need to understand both the strengths and limitations of the statistical information they use. Although BLS has long published standard errors for many of our surveys, this information was somewhat obscure, often hidden in technical documentation, dense tables, or available only upon request.

In recent years, and in an effort enhance transparency and contribute to the statistical literacy of data users, BLS launched an initiative to better explain and display measures of survey reliability. One outcome of this initiative was a series of posts in the BLS Commissioner's Corner blog. The blogs use two BLS examples to explain, in non-technical terms, what is meant by sampling error and how to understand confidence intervals.

Another outcome of this initiative was visual displays of survey estimates and their accompanying confidence intervals. The visualization effort began with a session for the BLS Data Users Advisory Committee which represents a diverse group of users and provides advice to BLS on matters related to the analysis, dissemination, and use of the Bureau's data products.

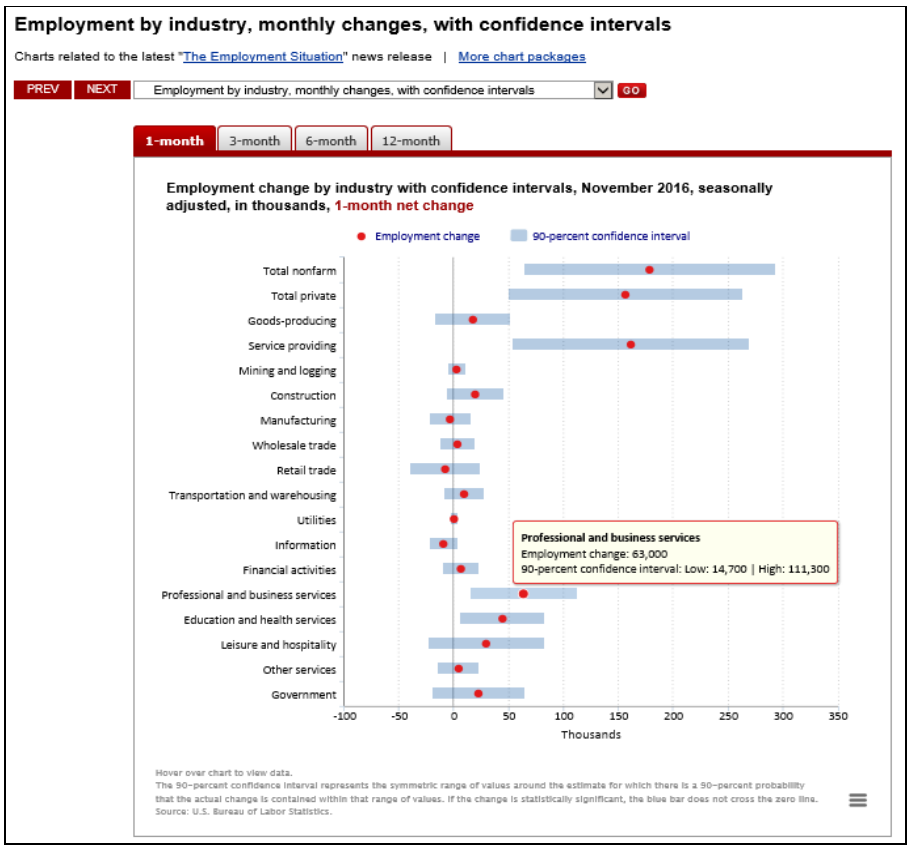
Box 4: BLS responding to user needs and communicating uncertainty in data)

Box 4- continued

The Current Employment Statistics (CES) program presented several possible charts employing a variety of visual techniques to convey the 90% confidence interval surrounding estimates of nonfarm payroll data, one of the most prominent data sets produced by BLS, and asked for feedback. The committee selected on approach and recommended that the program reach out to the public for additional comments. CES refined the preferred approach into two charts and conducted an online survey that included an explanation of the survey, clarifying text about each chart, and eight questions about the charts. The survey, which targeted existing CES data users and had a goal of 600 responses, received almost three times as many responses.

The majority of online respondents, 82%, found the charts useful in identifying statistically significant changes more quickly and getting a better sense of how the variability of estimates differs among industry sectors. Responses from users that self-identified as basic, intermediate, and super users; were not substantially and were overwhelmingly in favor of publishing the chart with the data.

Based on the results of the survey, CES further refined the visualization and, in December, 2016, BLS began publishing an interactive chart showing 1-month, 3-month. 6-month, and 12-month changes in employment by industry with confidence intervals as a standard accompaniment to its monthly Employment Situation news release. (see image below)



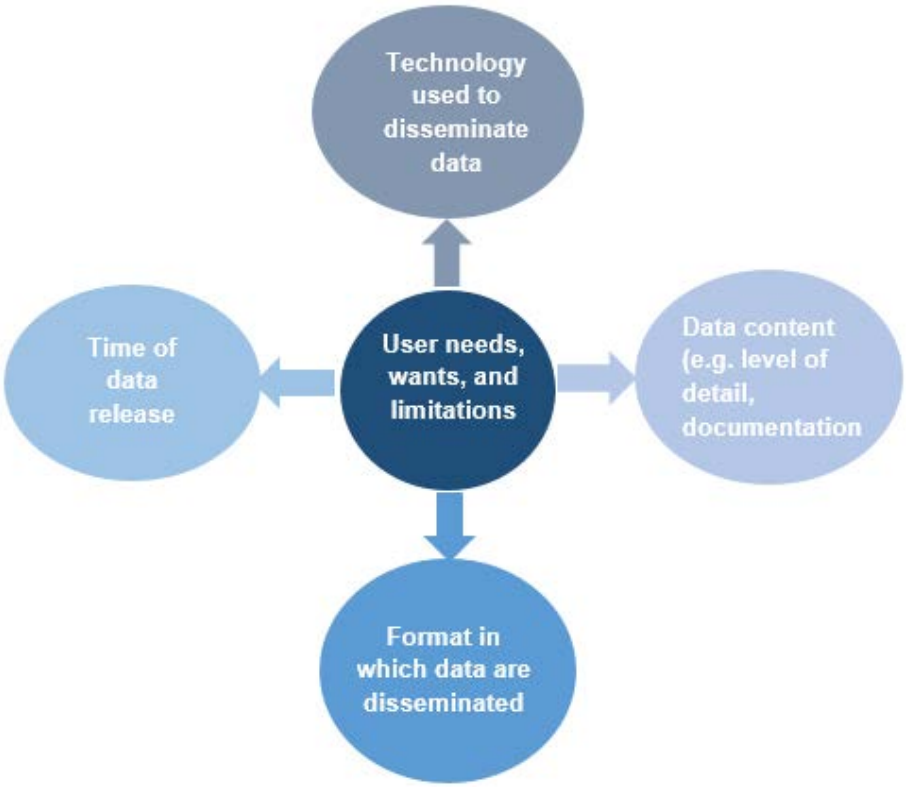
https://www.bls.gov/charts/employment-situation/otm-employment-change-by-industry-confidence-intervals.htm

[1] Michael D. Levi is Levi is Associate Commissioner for Publications and Special Studies at the Bureau of Labor Statistics

4 TOWARDS A MODERN AND EFFECTIVE DATA DISSEMINATION STRATEGY

An effective data dissemination strategy should be designed to provide users with the data they need in the right format and in a timely fashion, while ensuring accuracy and integrity of the data. Put differently, the focus of a successful data dissemination paradigm should be the users' needs, wants and constraints with relation to the content of the data, the format in which data are released, timeliness of delivery, and the technology used to deliver the data.

Figure 1: A User-Focused Data Dissemination Strategy



Source: Authors

In developing a user-focused approach to data dissemination, it is important to recognize that user needs and expectations are not only constantly changing in response to technological advancements, they also vary significantly across user groups. To respond to various user groups, federal agencies engaged in data dissemination can use different approaches to communicate data to reach different users (Speyer and Pagels 2014).

To help understand and address the needs of various data users, it would be useful to divide users into three major groups based on their needs and intended uses of the data:

- **General public**: The largest number of data users are made up of users who are not subject matter experts (e.g. students working on term paper or casual users interested in obtaining some data for one-time use). Members of this group tend to be interested in single point estimates or mega trends over time. They are generally less experienced data users and struggle with complex dissemination. This group needs simplified paths with plain language guidance to get them to a limited set of data for their topic of interest. Punchy data and visualizations showing key trends are also very effective in communicating data to this group. A good example is a recent effort by the Economic Research Service (ERS) to modernize delivery of some of its data. In March

2015, the ERS began releasing forecasts on U.S. agricultural trade, the fruit and vegetables markets in a series of interactive charts and maps. The agency used a new visualization software to make technical information and data—traditionally available in long reports releases in PDF format—available in a more modern, user-friendly and easily accessible format.

- **Media and policy makers**: The second largest group of users comprises members of the media and policy makers in general. This group has more specialized knowledge of the topic in question and tend to be interested in statistics that have relevance to specific policy issue or "story". For this group, data visualizations and briefs describing key findings can be very effective. A good example of data dissemination approaches targeted at policy makers and media is the Charts of Notes series which is released daily by USDA's Economic Research Service (ERS) and provides charts drawn from the agency's reports and data products containing notable facts and figures about the U.S. agricultural sector. Media and policy makers are more adept at navigating complex search and download tools, but still prefer simpler methods of getting to the larger quantity of data they need. The media, especially, needs simplified methods as they are often working on stories under tight deadlines and don't have time to search and analyze data. Media and policy makers are also very interested in map-based data that can be easily incorporated into stories or briefs.

- **Data analysts and academics**: This is perhaps the smallest group of users in terms of numbers, but the one that requires the most complete data sets and has the most need for information about data collection and processing methods. They tend to use query tools, application programming interfaces (APIs) and require documentation regarding methodology and other technical information. To meet the needs of this group, many Federal agencies have developed one or more APIs to deliver data in machine readable format to data analysts, academics and others who use the data in analytical models or scientific studies.

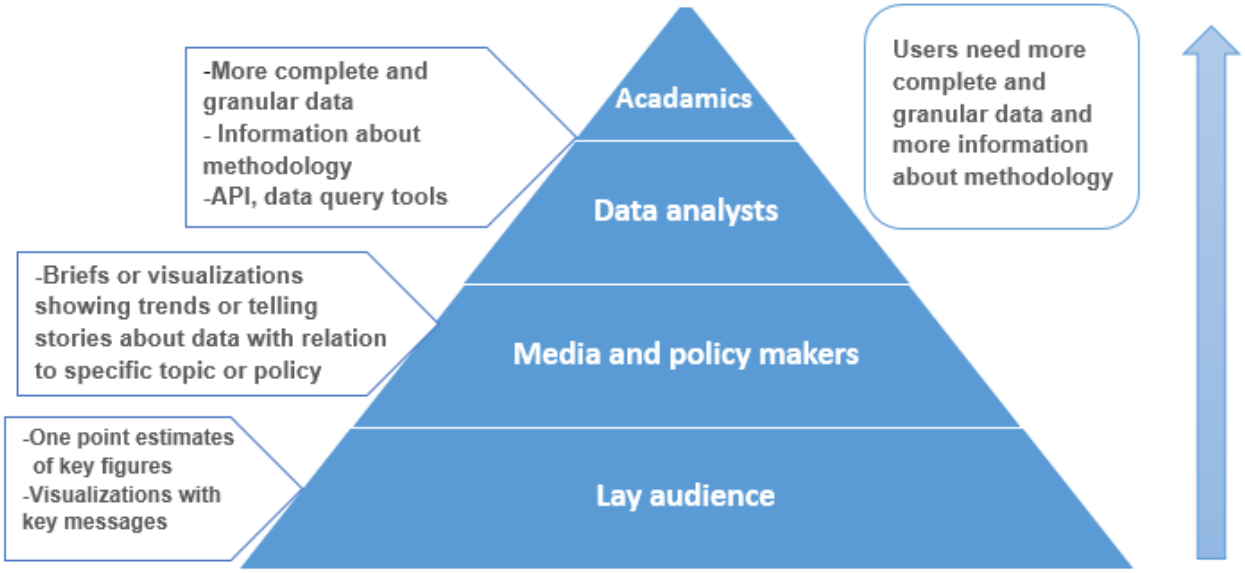
In order to meet the ever changing needs of data users and satisfy user groups with different needs and levels of understanding, agencies will need to develop dissemination platforms that are agile and can adapt quickly. Technical advances can help drive this process, but are hindered by the existence of data silos mentioned earlier. As agencies begin to break down the data silos, they can create integrated and standardized data and metadata repositories that will be the basis for data access moving forward. These integrated repositories will allow agencies to create "lightweight" data access tools that can be customized to user groups, access methods (mobile vs. desktop) or data displays (tabular, map-based, visualizations, etc.).

A common data and metadata repository would also allow agencies to publish data once and create tools that access the single instance of data and metadata. The paradigm of "publish once, use many times" is critical in moving agencies towards an agile, user-centric data dissemination approach and ensuring data quality. Having a single instance allows agencies to ensure that data changes and corrections are handled efficiently and not be concerned with updating multiple instances of the data. It would also allow the areas creating the data for publication to have a single delivery point using a standard format. These factors are all critical in enabling federal agencies to develop a more modern dissemination platform under the current budget and resource constraints.

The breakdown of these data and metadata silos needs to begin within each agency, but also needs to spread and be coordinated between agencies. The federal government has recognized

this need through publishing guidance on open data and interoperability between agencies, but significantly more work needs to be done. The demand for integrated data will continue to grow as the push to make data-driven decisions increase and Federal agencies, especially the principle statistical agencies, are in unique position to provide the foundation for such integration. The creation of an agile dissemination platform will give agencies the ability to provide end-user focused data access, but it does not ensure that this will happen.

Figure 2: Data User Segments and Needs



Source: Authors

At the same time, to better understand users' needs and wants, agencies engaged in data dissemination should seek continuous feedback from users. One way to obtain such feedback is by adopting a Customer Experience Management (CEM) system to ensure that customer interactions are captured and assessed systematically and that information obtained are used to further refine data products and dissemination platforms. Agencies should use web analytics systemically to collect, measure and analyze usage of data. Many Federal agencies also use online surveys to gather information from users about their online experience.

Agencies also need to continue to find creative ways to engage with their data users earlier in the development of new products and systems. Gathering user feedback early in the development process will allow government agencies to test ideas without spending much time and resources developing a tool that will not meet customer needs. Some agencies have created groups of users to provide early feedback on ideas for new data products or early testing of new tools. Such groups allow an agency to get early feedback on innovative ideas without investing significant resources in development of a new product or system that may not meet user needs.

Finally, *ensuring data accuracy is perhaps the single most important component of any data collection and dissemination program*. No matter how easily users can access data or how modern the mode of the dissemination platform is, if users don't have confidence in the integrity and accuracy of the data being released, they will not use it. As mentioned previously, data errors can occur at different stages during the collection, analysis or processing of data. Ensuring data quality requires developing a system for monitoring various sources of errors and implementing strategies for reducing errors during estimation, measurement or processing of

data. Clearly, many of these strategies are implemented long before the data are disseminated to the public. But any dissemination system should include three quality assurance components; a process or system for data validation, guidelines on communicating uncertainty about data to users, and a process for informing users about errors identified in data after it was released.

The objective of data validation during the dissemination stage is simply to profile the data for missing data points, inconsistencies, or outliers. The verification process could be carried out by an individual who is familiar with the data. Alternatively, agencies can use a simple automated process that identifies errors. The Economic Research Service, for examples, implements a tool that identifies and flags data outliers where data points that are one or two standard deviations from the mean are flagged automatically.

5 CONCLUSION

Demand for and use of U.S. Government data will continue to grow in the foreseen future, driven by technological change and rising interest in more and increasingly granular data. This placing increasing pressure on government agencies to expand access to official data, while ensuring its timeliness and accuracy. At the same time, user demands and expectations are constantly evolving, largely driven technological advances and the availability of increasingly granular data.

For Federal agencies involved disseminating official data, the Open Data policy has served as a catalyst for modernizing their data collection systems and dissemination platforms to meet users' demands and expectations. As shown in the examples drawn from some Federal Statistical agencies and discussed in the chapter, the scale and scope of these modernization efforts varied depending on agency priorities and objectives.

More work remains however. As federal agencies continue their modernization efforts or embark on new ones, it is important to improve coordination among government agencies to break down data silos and minimize the fragmentation of the Federal data system, while also integrating new technology to ease the transition. New search and navigation techniques can mitigate existing data silos and get end users to desired results more quickly. At the same time, new mapping and visualization tools can facilitate integration and interpretation of disparate data for the end user.

And in order to meet the diverse and evolving needs of data users, agencies should aim to develop agile data dissemination platforms that can adapt more quickly than they have historically. Federal agencies would also benefit from gathering continuous feedback from users and disseminating data in different formats.

REFRENCES

Bates, M. 2011. *Understanding Information Retrieval Systems: Management, Types, and Standards* December CRC Press, Boca Raton, Florida.

Gurin, J. 2014 *Open Data Now, The Secret to Hot Startups, Smart Investing, Savvy Marketing, and Fast Innovation*-McGraw-Hill.

Manski, C.F. 2015. "Communicating Uncertainty in Official Economic Statistics: An Appraisal Fifty Years after Morgenstern," *Journal of Economic Literature* 53(3): 1–23

Open Data for Development Network (2016). Annual Report. "Open Data for Development; Building an Inclusive Data Revolution." Retrieved from http://od4d.com/wp-content/uploads/2016/06/OD4D_annual_report_2015.pdf

Paperwork Reduction Act of 1995 (Pub. L. No. 96-511, 94 Stat. 2812, codified at 44 U.S.C. §§ 3501–3521)

Speyer, P. Pagels, B. Ho-Park, N. (2014) "*Communicating Data for Impact*." Forum One

Tauberer J. 2014. *Open Government Data*. Second Edition. Retrieved from https://opengovdata.io/2014/ancient-origins-open-access-to-law

The Confidential Information Protection and Statistical Efficiency Act of 2002 ((Pub.L 107–347, 116 Stat. 2899, 44 U.S.C. § 101).

The Federal Data Quality Act 2001 (Pub. L. No. 106–554—C 114 STAT. 2763A–153)

U.S. Department of Commerce, Bureau of the Census Historical Statistics of the United States. 1970. Colonial Times to 1970 Part 1, September 1975, Government Printing Office