

The background of the entire page is a photograph of the United States Capitol dome in Washington, D.C., viewed from a low angle looking up. The dome is white with a gold-colored top and a statue on top. The sky is blue with some white clouds. The image is partially obscured by geometric shapes: a light blue triangle on the left, a white triangle on the right, and a dark grey triangle at the bottom.

Enterprise Data Inventories

Agencies face challenges and opportunities to increase the value of data assets when implementing data inventories

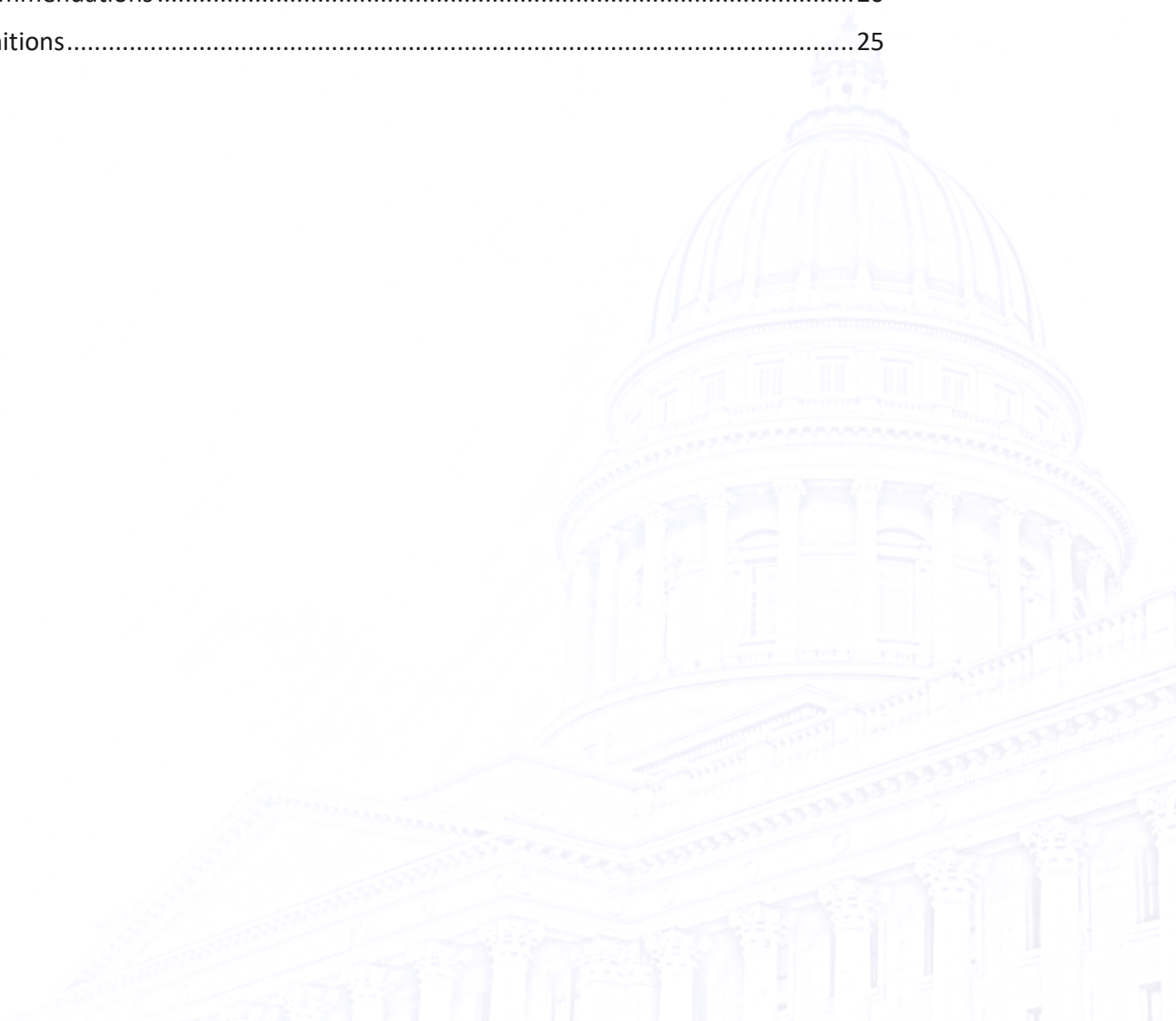
This report was developed by the Chief Data Officer Council's Data Inventory Working Group to support implementing the Federal Data Strategy and the Foundations for Evidence-based Policymaking Act.

Released April 2022



Contents

1. Introduction	1
2. Enterprise Data Inventory Value	4
3. Agency Data Inventory Policy	6
4. Agency Culture	8
5. Technology	12
6. Data Protection and Cyber Security	16
7. Recommendations	20
8. Definitions	25



Acknowledgments

The Federal Chief Data Officers Council Data Inventory developed this report and included input from the CDO Council.

Published: April, 2022

Report Lead: Tod Dabolt, Chief Data Officer, Department of the Interior

Contributors: Anne Levine (FCC), Robin Rappaport (IRS), Arun Acharya (CPSC), Harold Saintelien (Census), Lajuan Bryan-Beveridge (DOC), Chris Alvarez (USDA), Cynthia Parr (ARS), Brian Raub (IRS), Brian Reichenbach (DOS), Hyon Kim (GSA), David Carter (DOI)

Contributing Agencies:

- U.S. Census Bureau
- U.S. Consumer Product Safety Commission
- U.S. Federal Communications Commission
- U.S. Department of Agriculture
- U.S. Department of the Interior
- U.S. Department of State
- U.S. Department of Transportation
- U.S. Department of the Treasury
- U.S. Department of Veterans Affairs
- U.S. General Services Administration
- U.S. Office of Management and Budget
- U.S. Environmental Protection Agency

With thanks to the leadership of the CDO Council Executive Committee

- Ted Kaouk, Chair
- Dan Morgan, Vice Chair
- Maria Roat, OMB E-Gov
- Dominic Mancini, OMB OIRA
- Kshendra Paul, Large Agency Committee
- Kirsten Dalboe, Small Agency Committee
- Matthew Graviss, Data Skills and Workforce Development Working Group
- Nikolaos Ipiotis, Data Sharing Working Group
- Tod Dabolt, Data Inventory Working Group
- Melanie Carter, Operations Working Group

1. Introduction

The *Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act)* calls for a systematic rethinking of how the federal government manages and uses the information it collects, emphasizing strong agency coordination for the strategic use of data. Specifically, Title II of the *Evidence Act* is the *Open, Public, Electronic and Necessary (OPEN) Government Data Act* (an Act within an Act), which lays out certain agency responsibilities, including the requirement for a comprehensive data inventory.

The Chief Data Officer's Council Data Inventory Working Group developed this paper to highlight the value proposition for data inventories and describe challenges agencies may face when implementing and managing comprehensive data inventories. It identifies opportunities agencies can take to overcome some of these challenges and includes a set of recommendations directed at Agencies, OMB, and the CDO Council (CDOC).

A data inventory is foundational to any formal data management program. Data inventories enable stakeholders to efficiently find, access and use data assets. Inventories are also indispensable to managers who need to evaluate the extent to which the organization's data helps meet mission goals or can be shared with other agencies to meet their missions. Policy and program managers not only need to know *what* data the organization collects and maintains but also *how* it aligns with its intended purpose, what condition it is in, how it is stored and accessed. In addition, data inventories can inform records management teams who need to understand the record retention schedule of the organization's data and information; security and privacy teams who need to ensure the organization has the appropriate security and access management procedures in place; and legal teams who need to ensure the organization is using the data appropriately.

The purpose of the comprehensive data inventory is to ensure that agency staff and the public have a clear comprehensive understanding of the data assets in the possession of the agency. While OMB data inventory guidance is forthcoming, the statute requires each agency: "shall, to the maximum extent practicable, develop and maintain a comprehensive data inventory that accounts for all data assets created by, collected by, under the control or direction of, or maintained by the agency." Furthermore, the *Evidence Act* requires that the data inventory metadata schema include, to the maximum extent practicable, the following:

- Description of the data asset (including all variable names and definitions).
- Name or title of the data asset.
- Indication of whether the data asset is eligible for disclosure or if it is exempt (or partially exempt) from public disclosure (or the date by which the agency will make such determination).
- Agency or sub-agency responsible for maintaining the data asset.
- Owner of the data asset.
- Date on which the data asset was most recently updated.
- Any use restrictions on the data asset.
- Location of the data asset.

- Method by which the public may access or request access to the data asset.
- Any metadata necessary to make the inventory useful to the agency and the public.

The differences between a Data Inventory and a Data Catalog.

Although the terms “data catalog” and “data inventory” are often used synonymously, they mean very different things. A data catalog is the mechanism that helps users discover the data assets that are found in the data inventory. The data catalog contains such information as the organizational ownership of data assets, its’ meaning to the organization (business metadata), and where and how to access it. A data inventory can contain more technical and granular metadata such as the definitions of specific data elements, their format, valid values, and their completeness. While the concepts are distinct, they are complementary. A user may search a data catalog to find data sets related to a specific topic of interest and then use information in the data inventory to discover the meaning of data elements within the datasets they discovered through the catalog. A data catalog and data inventory should work in tandem to provide organizations with the best value. An item in the data catalog may reference multiple items in the inventory and vice versa.

Starting with basic business metadata, inventories may also link data elements and data assets to additional information regarding security, privacy, and records. Many enterprise data inventories also capture operational metadata. Operational metadata describes the data lineage, data currency, and the flow of the data across an organization. The technical and operational metadata in an inventory change frequently and require automation while the business information associated with the data catalog tends to be more static.

Once inventoried, the data assets described in a comprehensive data inventory must, as per statute, be submitted for inclusion in the Federal Data Catalogue and must be updated in real-time. Also, as per statute, the data assets themselves must be made available to the public via electronic hyperlink or other means of access, unless there are valid reasons for not doing so, as defined under statute. Valid reasons for data assets being exempt (or partially exempt) from public disclosure, are based largely on (but are not limited to), the following criteria:

- risks and restrictions related to the disclosure of personally identifiable information, including when combined with other information
- security considerations, including when combined with other information
- the cost and benefits to the public of converting the data into a usable format,
- risk of legal liability,
- intellectual property rights,
- confidential business information,
- restrictions based on contract or other binding, written agreement,
- consultation with the holder of a right to such data asset.

These reasons for exemption apply to all individual data assets in isolation, but also if, when combined with other available information, the individual data asset may pose such a risk.

OMB Guidance on the Foundations for Evidence-Based Policymaking Act¹

In September 2020, OMB circulated Draft Phase 2 Implementation Guidance for Inter-Agency Review. OMB Draft Guidance recognized that the comprehensive data inventory of each agency will need to be built out over time, reflecting the priorities of the individual agency. Governing and managing the development and operation of a successful data inventory requires a cross-disciplinary team supported by the agency's data governance body.

“Because developing and maintaining a comprehensive data inventory will require managing a significant breadth and depth of data assets, agencies may want to consider leveraging Capital Planning and Investment Control Coordinators, Chief Information Officers, Chief Freedom of Information Act Officers, Controlled Unclassified Information Senior Agency Officials, Enterprise Architects, Senior Agency Officials for Privacy, and Senior Agency Officials for Records Management, as part of the inventory team.”²

Draft OMB Guidance suggests identifying individual assets from a variety of different sources already institutionalized in most agencies, including, but not limited to the following:

- Information Collection Review (ICR), Inventory comprised of ICR requests submitted to OMB under the Paperwork Reduction Act (as listed on [reginfo.gov](https://www.reginfo.gov)).
- Information System Inventory (IT systems).
- Any data assets that are publicly disseminated on an agency website.
- Any data asset that is associated with a system of records as defined by the Privacy Act (SORN).

¹ Phase 2 Guidance

² Draft OMB Phase 2 Guidance

2. Enterprise Data Inventory Value

An Enterprise Data Inventory enables analysts to find and use the agency's data. Data Inventories also support the management and security of an organization's data assets. An Enterprise Data Inventory is also a key requirement of the 2018 Foundations for Evidence-Based Policymaking Act (FEPA) and subsequent Federal Data Strategy (FDS). Agencies' enterprise data inventories support the cataloging and more efficient use of Federal data. Data inventories increase the value of agencies' data assets by capturing knowledge of what data exists, its meaning, and how it can be assessed. Analysts are more efficient when they have access to this knowledge compiled in a clear, consistent, standard manner with controlled language for metadata. A comprehensive data inventory improves decision-making by helping analysts discover and evaluate all their agencies' data sources beyond those data stored in isolated mission silos.

An Enterprise Data Inventory facilitates the use, storage, and sharing of government data amongst various stakeholders. It supports Diversity, Equity, Inclusion, and Accessibility by providing equal access to Federal Open Data. It enables fulfillment of the requirement to "... make evidence-based decisions guided by the best available data, ..." ³ reiterated in the January 27, 2021, Presidential Memorandum, Restoring Trust in Government Through Scientific Integrity and Evidence-Based Policymaking. It also supports the following OMB guidance (issued June 30, 2021):

... build and maintain trust in government and ensure that decisions best serve the American people....The Evidence Act provides a statutory framework to advance this vision ... it calls on agencies to strategically plan and organize evidence-building, data management, and data access functions to ensure an integrated and direct connection to evidence needs.... collection, curation, governance, protection, and transparency of data are also essential for evidence-building The need to collaborate extends within and across agencies as well; evidence-based government requires cross-agency work including data sharing in support of addressing Learning Agenda and evaluation activities data ... identified to address data linkages in which records from two or more datasets that refer to the same entity are joined.... ⁴

An Enterprise Data Inventory allows agencies to actualize compilation and consolidation realizing the full value of Federal data. As Chief Data Officers (CDOs) endeavor to inventory agency data assets, "according to the effort is the reward" applies. This process improves the discoverability, usability, and governance of Federal data. It supports the dimensions of Data Quality (Timeliness, Relevance, Accuracy, Accessibility, Interpretability, and Coherence) identified by Gordon Brackstone in "Managing Data Quality in a Statistical Agency" in the Federal Committee on Statistical Methodology (FCSM). ⁵

³ 2021 Presidential Memorandum, Restoring Trust in Government Through Scientific Integrity and Evidence-Based Policymaking

⁴ OMB guidance (issued June 30, 2021) M-21-27

⁵ Gordon Brackstone in "Managing Data Quality in a Statistical Agency" in Federal Committee on Statistical Methodology (FCSM)

Metadata provides the key to unlocking the meaning of data and providing context. Analytics professionals typically spend 80% of the time allocated for data preparation (including understanding the data to identify the appropriate data for use) leaving only 20% for actual analysis of the data (Forbes, 2016)⁶. The metadata in the Enterprise Data Inventory empowers analytics professionals to spend more time on analytics and less on data wrangling resulting in improved decision making and better governance.

The metadata used to describe the Federal data assets needs to be clear, consistent, and structured in a standard manner with controlled language. It should be searchable and include mission context, tagging, relationships, data quality, and usage. The platform should be easily accessible to all possible users including the public (as allowed), data stewards, IT professionals, and the workforce within and across agencies.

A Federal government-wide Enterprise Data Catalog implementation fed by Agencies Enterprise Data Inventories benefits include improved:

- data-driven decisions,
- data sharing,
- data usage,
- data analytics,
- data and metadata management,
- data and metadata quality,
- data and metadata standards,
- data provenance,
- data lineage,
- data and metadata governance,
- data stewardship,
- data and metadata curation,
- data and metadata validation,
- data discovery,
- data transparency,
- data access,
- data protection and
- data linkage.

⁶ <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=11a050a96f63>

3. Agency Data Inventory Policy

A well-crafted, measurable policy is among the most effective means of both initiating the practice of data inventory creation and improving and sustaining such a practice. Depending on what best suits the environment, the data inventory policy may be as complex as a fully detailed policy document or as succinct as a policy statement.

Challenges Identified by the Workgroup

The specific policy challenges that an agency may face will vary according to the specific characteristics of its environment.

1. Issuance of Agency Policy on Data Initiatives

U.S. Code Title 44 Chapter 35 Section 3511 - Data inventory and Federal data catalogue requires agencies to maintain a comprehensive data inventory which “provides a clear and comprehensive understanding of the data assets in the possession of the agency.” Section 3511 also directs the OMB Director to establish guidance “for agencies to develop and maintain comprehensive data inventories.” Agencies’ data inventory policy must align with U.S. Code and OMB guidance on Data Inventories. Section 3511 does not differentiate the types of data assets Agencies’ comprehensive data inventories include. Agencies who have multiple types of data including geospatial data assets and statistical data assets should include these data in their comprehensive data inventory, except for those assets that are related to national security. The Geospatial Data Act, U.S. Code Title 43 Chapter 46 Section 2807 requires a subset of agencies to include metadata for their geospatial data assets in the GeoPlatform. This requirement does not negate the requirement for agencies to include this metadata in their comprehensive data inventories.

Agencies that already have an established, sustainable policy on data Inventories, may only need to update it to fit new requirements. Since all policy documents should have a specified review cycle, such updates would fit well into the scheduled review cycle.

2. Scoping a Policy on Data Inventories

Agencies that are starting in formalizing data inventory development and management should consider an iterative bootstrapping approach. In this approach, a skeleton policy is drafted and concurrently tested by deploying it on a small set of (or single) prioritized data assets. The policy is then improved based on the experience in each iteration. A final, formal enterprise policy is issued based on improvements to the original through several iterations.

The benefit of this approach is that the version released as enterprise-wide policy will have a high level of effectiveness and survivability. However, this approach will take time. Agencies should take care to prioritize and select the data asset(s) for each iteration based on mission drivers.

3. *Metadata Standards*

Agency data inventory policy will need to reference metadata standards to maximize the utility of the metadata these inventories manage. Agencies with multiple types of data may need to reference multiple metadata standards in their policy while maintaining the minimum set of metadata elements required by OMB. A best practice is to reference a metadata standard or standards in the policy rather than incorporate the specific standard into the policy. Metadata standards are likely to change more frequently than the overall data inventory policy. There are multiple metadata standards in use today and referenced or prescribed in public policy. Useful data catalogs and data inventories require a high degree of metadata interoperability both between metadata standards as well as within a given standard.

For example, sometimes data itself can be lost when translating across existing standards. For example, the Federal Geographic Data Committee (FGDC) recommends agencies use ISO standard 19115 for documenting geospatial data. ISO 19115 is a more comprehensive and descriptive standard than the DCAT-US (formerly referred to as Project Open Data) V1.1. DCAT-US is the metadata standard required by Data.gov for all Federal Open Data. Agencies that document their metadata following ISO 19115 and translate that metadata to DCAT-US 1.1 lose the specified meaning of key fields in the ISO 19115 standard. As a result, user experience suffers with downstream applications like GeoPlatform whose search functionality and data distribution rely on compromised metadata. Interoperability is also compromised when different organizations interpret a field within the same standard differently. For example, a metadata standard may include a field for “distribution” and another field for “online distribution”. Some metadata records might link to the data asset download page through the distribution field and others through the online distribution field. The result is catalog users (human and machine) cannot easily establish code to access data consistently from a catalog.

Metadata profiles are one way to improve the consistency of metadata records to improve search and data interoperability. A metadata profile should be established for an agency’s catalog as well as for the federal catalog. A metadata profile encodes the business rules as well as data domains for a specific standard. Using a standard profile will decrease the ambiguity of how users populate metadata as well as improve the efficiency of software consuming metadata from catalogs and inventories. Some agencies overcome the challenge of managing multiple metadata formats by designing a data inventory that can cleanly import and export multiple metadata data standards. These agencies often use a well-governed set of metadata translation functions to map metadata from one standard to another.

4. *Roles and Responsibilities*

The success of data inventory and metadata management will depend on several specific responsibilities being executed by individuals in designated roles. The enterprise policy must specify these roles and responsibilities. The policy should consider the roles and responsibilities needed to manage the data assets across the entire data lifecycle. For example, what is the role and responsibilities related to the data inventory for those employees engaged in paperwork reduction act responsibilities?

4. Agency Culture

The challenges agencies face in establishing and maintaining comprehensive data inventories will vary based on an agency's culture. Implementing the processes necessary to maintain an enterprise data inventory will require moving through organizational change. CDOs will need to clearly communicate why the change is happening, establish a vision, point out the specific benefits an inventory provides to different stakeholders, and guide the organization through the change. Once staff across the organization understand their specific roles associated with the data inventory, CDOs will need to build formal and informal mechanisms to reinforce the changes and prevent backsliding. These could include incentives to reward staff and organizations who are following the new procedures as well as management controls to monitor progress. Lastly, organizations should routinely measure the progress and performance of their processes, making improvements as needed.

Centralized Management is a Culture Shift

Building a comprehensive inventory requires significant planning and collaboration. A centralized effort with strong leadership and significant cooperation from numerous stakeholders across the organization is required to ensure that each asset is identified, categorized, fully described, and has a point of contact to inform the inventory. Previous efforts to comprehensively inventory government data assets (M-13-13), in absence of a Chief Data Officer (or other centralized data management team) meant the inventory wasn't explicitly anybody's responsibility making accountability a challenge. Agencies couldn't sustain leadership buy-in, as Open Data's value proposition was mostly external facing and marginally connected to the agency's mission objectives.

As a result, many agency inventories on data.gov are incomplete. U.S. Code Title 44 Section 3520 corrects this deficiency and requires CDO's to carry out the data inventory requirements on behalf of the agency. CDO's need to reframe the value of enterprise data inventories. The greatest return is improving agencies' data-driven decisions through Open Data managed within an enterprise data inventory. In doing so, the Public Open Data catalog becomes a valuable by-product of the internal agency data management processes.

Even with a carefully planned centralized effort; however, cooperation can be difficult to mobilize. Most Federal agencies are siloed and have different business requirements or cultures both across and sometimes even within silos, so creating a central repository in and of itself can represent a dramatic culture shift. Even internal business units with similar missions and or shared data may not collaborate and resolving data discrepancies can represent a massive change in business process.

Getting "Buy-in"

Getting everyone on board to develop a data inventory is no small task. Building "buy-in" can be a lengthy process, encumbered by significant staff and management resistance. Many organizations and their staff did not benefit from prior data inventory efforts, and staff with direct experience of prior efforts are particularly skeptical. Those who will be required to dedicate time to curating the inventory understandably want to know how much of their time will be needed and if the benefit will exceed

their contribution; some will wonder how this inventory effort will be different from prior inventory efforts.

Why is “Buy-In” important?

Inventories require subject matter experts to describe and categorize data assets. While data managers can set up databases, organize the information and build out the IT infrastructure, it is the input from data stewards and subject matter experts (SMEs) that will make the inventory useful. While inventories enable stakeholders’ access to and use of data assets by making those data assets more discoverable, it only happens when data assets have sufficient metadata in the enterprise inventory. For many SMEs, contribution to the inventory is “extracurricular work” on top of their already overloaded schedules. Despite the requirements from the *Evidence Act* and Federal Data Strategy, participants want to understand how the effort will benefit them if they are going to be spending time away from their regular work to contribute to it.

Building ‘Buy-In’ How to Overcome Cultural Obstacles

1. Tailored Messaging

Helping staff understand what’s in it for them is a crucial aspect of gaining full participation, but not overselling the effort is also important. In communicating the benefits of an inventory, CDOs and their teams must manage expectations of what is possible. Some organizations will have enough funding to build out highly functional user interfaces for their inventory, while others will be relegated to a spreadsheet in the short (or long) term. It is essential to communicate with each stakeholder in a way that ensures they are all prepared to hear your message and it is a message that aligns with their goals.

Advocating for the value of an enterprise view of the organization’s data can mean developing a different message for each outreach encounter. Designing the outreach for inventory stakeholders requires the inventory team to understand the culture in which the different stakeholders operate. For many organizations, there is no one-size-fits-all approach. Each business unit will need to hear the message their culture says they are ready to hear.

2. Demonstrate Value with Quick Wins

Some agencies have begun to produce some quick wins to demonstrate how the inventory can be useful.

- In one case, a CDO leveraged the agency’s response to an Executive Order to build interest and momentum in identifying and inventorying data assets necessary to inform the agency’s response to the requirements.
- Another CDO used a Cybersecurity & Infrastructure Security Agency (CISA) cyber data call to collect and inventory the agencies’ sensitive data assets.

Other agencies recognize that people will only get involved in a useful way if the Inventory team “seeds” the effort. This means data custodians aren’t asked to start with a blank page. In

many instances, contributing to the inventory will be a task above and beyond their normal work. Asking someone to validate and curate metadata is vastly different from asking them to list and describe their data. Providing the subject matter experts with an organized, and partially pre-populated inventory will demonstrate what is needed while keeping the burden minimized for contributors.

3. *Data Literacy Training*

Increasing an organization's overall data maturity and the workforce's data literacy helped some agencies improve their data inventories. Staff and managers who actively use data are more likely to see the value of a data inventory and contribute to its success. Many federal employees don't have enough awareness of our current mandates and emerging responsibilities concerning our data. It may be hard for some to understand what a "data asset" is, and it is up to the Inventory Team leadership to clearly define this. It is also helpful for SMEs to understand the connection between the value and utility of data tracking and why it's so important and valuable for SMEs to contribute.

4. *Changing Business Process*

Some agencies have focused on starting to capture metadata as early in the data's lifecycle as possible. Data architecture and documentation are as important to systems as the IT architecture. If the inventory process becomes part of the data creation process, then it won't be necessary to collect all the metadata at the point of data dissemination.

5. *Creative Approaches*

While the topical area of data inventory may have an immediate appeal to a few positions within an organization, most roles across the enterprise will seldom think of data unless a need arises. Even then, the level of understanding can be minimal at best. That is why it is important to incentivize data literacy and participation.

One potential approach is to focus on a topic that is pertinent to the business that the agency or office conducts. Instead of attempting to garner enterprise participation through invites titled "Data Inventory: Mandated by Federal Code", try and make the theme topical and relevant to the enterprise, such as "Data on Diversity and Inclusion: Why It Matters". It is much easier to get people to engage and continue to participate if they can see an immediate benefit to such.

Additionally, you may consider data-driven competition. Many offices enjoy friendly competition amongst each other. If you track the amounts of data assets being uploaded, you can create a leaderboard. It is also a great idea to have leadership talk about the importance of populating a data inventory and recognize offices doing a great job. This could also be extended to an interagency level.

Another group to consider engaging is the public. What would they like to see and how do they want to see it? Have an existing data set already available? You can hold a visualization competition. Want to derive insights from multiple data assets but would like some fresh eyes?

Host a public competition and feature the best few projects. Do you think you got it all entered, but can't help thinking you may have missed something? You can have a competition for the public to list data that may be missing. These competitions don't even have to be public. You can host them internally or with other agencies or important stakeholders.

Reassurance for CDOs It is easy to hear success stories from other agencies and feel frustrated that your agency isn't close to that kind of success. It's also easy to think that success was easy for those agencies. The reality is that it's likely that there was a lot of drudgery (and possibly some setbacks) that went into that success. It's important to not be afraid to fail or have setbacks because they will happen; however, be prepared to be iterative so you can overcome these setbacks.

Similarly, it's important to remember that many organizations find the inventory process slow and tedious, and "slow and steady" isn't hugely satisfying. It means going through long periods curating metadata or planning out how to set up a certain tool before seeing the fruits of your labor. Building an inventory requires an effort that is not that dissimilar to the preparation steps that occur when processing data for analysis; the preparation can take as much as 80% or even 90% of your time just to get the data ready for the actual analysis. Similarly, metadata curation and organization, when done properly can take a long time.

The *Evidence Act* says throughout that the mandates are to be implemented "to the maximum extent practicable." This language recognizes that different agencies are all starting in different places and have different resources available to them, and thus will also have different goals. It's important not to compare yourself too much with what other agencies are doing. You must move ahead with what works for your organization, regardless of what resources you have. The culture of your organization may not allow you to take a path that the textbook or generally accepted best practices approach, but that's okay. Many agencies are following a different path as well. Organizations are like bodies, and everybody is different. You can't treat them all the same.

5. Technology

Management of data as a strategic asset has never been more necessary. This coupled with the exponential growth of data creates unique technology challenges. Fortunately, market innovation is emerging to meet those needs. Leading efforts in the space of data fabrics opportunities offer cost control, avoidance, and increased value added. To the extent possible, federal data inventory initiatives would benefit from decomposing innovation opportunities in the market and identifying foundational practices that build on a greater value stream to overall increased data maturity.



Figure 1: Increasing the value of data inventories through technology processes

Traditional approaches to data inventories often focused on accounting for the data across an organization and describing those data. Extracting additional value from these data assets requires making the data understandable to humans and machines by using governed taxonomies as well as capturing the relationships between data assets and data elements through semantic enrichment using formal ontologies and knowledge graphics. Agencies may need to invest in several software products over time to fully realize the value of their data assets.

Summary of Challenges

The breadth and variety of different data management scenarios including structured and unstructured data with a wide array of data management systems lend to data inventories obstacles. The following highlights some of those challenges.

1. Metadata

Metadata or the data that describes data affords foundational meaning within a data inventory by organizing and describing data. Three types of metadata include technical, operational, and business metadata. Their distinction helps understand associated challenges.

a. Technical Metadata

Technical metadata regards column name or some form of label that identifies the data element and describes the storage characteristic or datatype (i.e., integer, character, date, time, etc.). This is essential for understanding the dataset, relating to other datasets, and inferring technical meaning among disparate sets of data. As an example, two datasets with

well-defined technical metadata might each reference a column by something like “last name” with a datatype of character. This attribution clarity affords basic descriptions, lending to universal utilization, greater insights, and understanding.

Organizations could rely on manual documentation. However, this is only as reliable as the administration process, which is often either nonexistent or not maintained. Alternatively, assimilating technical metadata would require some level of machine provisioning to automatically archive or assimilate technical metadata embedded within datasets. This approach benefits from distinguishing related challenges between structured and unstructured data.

Database management systems (DBMS) typically support structured data bound with technical metadata. However, each platform vendor typically has distinct methods and practices of labeling and describing technical metadata. Additionally, each platform has different connection methods, drivers, and resides across various authenticated networks.

Unstructured data is becoming increasingly ubiquitous and regards data that is more free form and may lack clear labeling and descriptions of data elements. It may originate from the Internet of Things (IoT) devices (i.e., sensors, phones, GPS devices), PDFs, or flat files. This lack of clarity and structure creates additional challenges to simply understand what is in each dataset. Before any meaning can be derived or inferred a process is required to provide some level of data element distinct or structure and labeling.

Whether structured or unstructured, automated processes would need to access datasets where they reside without negative impacts such as data loss or corruption. This requires a fair level of technical complexity, some of which may be negotiated by tools, but others require organizational processes. As an example, connecting to each dataset requires appropriate authentication to each network, negotiating firewalls, properly developed drivers and connection methods by platform technology type and version, along with a tool or process code to effectively extract technical metadata.

b. Business Metadata

Business metadata is the plain language description of data from a mission perspective. The composition of business metadata is a time-intensive effort that requires subject matter experts to adequately describe and catalog data elements (i.e., databases, tables, and columns) with business descriptions. This process is often overlooked when datasets originate, and these metadata are rarely maintained. Moreover, these metadata are often not as machine-readable as they could be, are not well standardized across organizations, and there is only limited adoption of semantic-based approaches.

Geospatial data supports business metadata in a standard Government-wide format as determined by the Federal Geographic Data Committee (FGDC) and is frequently maintained with a vendor-specific tool (ESRI ArcGIS). The ArcGIS format of Business Metadata currently conforms to the Content Standard for Digital Geospatial Metadata (CSDGM) standard as well as

the ISO-19115-3 standard. ESRI implements both standards, but with additional items and vendor-specific XML tags which makes converting to CSDGM and ISO standard XML tags and structure challenging. The ESRI tool has limitations in working with the newer ISO standards and does not provide the full range of capabilities identified by that standard. Other tools such as the GeoNetwork tool can handle the ISO standard fully since it functions within the native ISO XML format and structure. While the metadata can be exported to a machine-readable format and imported into an enterprise inventory tool, there are challenges related to translating the geospatial standard into a format that can be read automatically by the inventory tool. The creation of such a translation template/tool is not trivial nor inexpensive. However, without such capabilities, there could be enormous manual entry requirements that also introduce metadata quality concerns and synchronization issues with the source metadata systems.

Agencies have adopted several metadata standards, such as ISO14008, ISO9115, FGDC data standards, Federal Data.gov standards, etc. Many catalog tools do not provide sufficient support for these standards out of the box, either for importing or exporting data.

Some agencies have a federated approach to inventorying data and oftentimes use different inventory tools. Integrating data across multiple tools into enterprise inventories often requires significant manual effort. Opportunities exist to develop more common standards or interoperability across tools.

c. Operational Metadata

Operational metadata captures the administrative information necessary to manage a data asset and includes information such as when it was created; the file type; the purpose of the data; information needed for archival, integration, and update of schedules; and access rights and entitlement restrictions. The administrative metadata related to data governance and stewardship is also included under Operational Metadata. This is descriptive information used to understand the roles of the individuals involved in governing the data. It identifies governance bodies and their scope, process, participants, structure, and responsibilities; and is used to manage change to all types of metadata. In addition, operational metadata is used for process improvements to enhance productivity and improve data quality. Process metadata, a subcategory of operational metadata, addresses process steps for production and maintenance, as well as for data quality measurement and analysis. Business rules; names of relevant systems, jobs, and programs; as well as governance and regulatory roles, and other control requirements are examples of process metadata.

This type of metadata is important for supporting data and information-related processes associated with records management, the Freedom of Information Act (FOIA), privacy, and data security. They can be used to create audit trails, which help with data-driven regulatory requirements related to data security and access, the status of any FOIA requests, and whether the data are currently available online or are archived. Operational or administrative metadata reduces overhead in data administration by providing adequate documentation relating to the current state of the data.

Opportunities

The explosion of data throughout the government and private sector is driving innovation in the marketplace. Innovation efforts often described within the “data fabric” originate in the data inventory space.

6. Data Protection and Cyber Security

Information security analysts and data analysts benefit from an Enterprise Data Inventory. Knowing what types of data, where it resides and the flows of that data across and beyond organizational boundaries provide critical information needed to protect the data as well as use the data.⁷

However, agencies face a variety of challenges when trying to implement data-centric security management approaches such as those necessitated by Zero Trust Architectures. A recent NIST publication, DATA CLASSIFICATION PRACTICES Facilitating Data-Centric Security Management by Karen Scarfone and Murugiah Souppaya summarize these challenges:

Data-centric security management necessarily depends on organizations knowing what data they have, what its characteristics are, and what security and privacy requirements it needs to meet so the necessary protections can be achieved. Standardized mechanisms for communicating data characteristics and protection requirements are needed to make data-centric security management feasible at scale.⁸

The limited nature of existing data classifications standards outside of the government and military means that most organizations do not use classifications that are consistent with those of their partners and suppliers. Organizations perform countless transactions with others for which data classification and protection are relevant, and the lack of industry standards impairs organizations' ability to enforce data controls.

Handling requirements:

- The lack of common definitions for and understanding of classifiers can result in information being classified and labeled inconsistently. Reliance on end-users to identify and classify the data they create and receive is particularly error-prone and incomplete.
- Data is everywhere: on devices (e.g., laptops, desktops, mobile devices), in applications running in both on-premises and outsourced environments, and in the cloud. This distributed nature of data complicates the process of establishing and maintaining data inventories.
- Data classifications and data handling requirements often change during the data lifecycle, for example safeguarding the confidentiality of data at first, then subsequently releasing that data to the public. Another example is data being safeguarded and retained for a certain period, then being destroyed to prevent further access.

Existing NIST standards and guidance regarding data classification and labeling, such as Federal Information Processing Standard (FIPS) 199 and NIST Special Publication (SP) 800-60 [3], address

⁷ It should be noted that the OPEN Government Data Act provisions do not apply to National Security Systems. 44 U.S.C. §3511(a)(2)(B)

⁸ <https://csrc.nist.gov/publications/detail/white-paper/2021/07/22/data-classification-practices-data-centric-security-management/final>

federal government-specific requirements, but not the many other requirements to which federal agencies and other organizations are subject.”⁹

In addition to determining the appropriate data classification scheme, determining the appropriate level to classify data assets is an additional challenge. For example, organizations often apply FIPS 199 information types at the system level which is too broad of a grouping of data assets for data analysts who are trying to discover and use a specific category of data. Classifying data at the system level can result in overclassifying data assets. In addition, some types of applications have a very large number of information types assigned to them, for example, an email system or shared network drive. These broad classification levels complicate determining the actual risk associated with the system and precisely knowing which specific types of data are available for re-use. The current practice assigns the degree of cyber risk for an entire system based on the most restrictive information type associated with the system. Data analysts and security analysts would benefit from understanding the meaning of individual data attributes and their sensitivity within a given system. This level of data tagging improves the effectiveness of data search and discovery, data reuse, and data protection.

As part of the Cyber Security Executive Order of May 2021, agencies were requested to assess the number of sensitive data assets held per each information system. Knowing the amount of sensitive data within a system helps organizations understand the magnitude of risk associated with the loss of that data. Most data inventories today do not capture the number of records associated with a specific set of data attributes or categorization schemes.

The draft Federal Zero Trust Strategy wasn’t explicit on how existing federal data categorization regimes should or could be used to label data assets. Existing federal data categories include:

- Controlled Unclassified Information (CUI)¹⁰
- NIST SP 800-60, Guide for Mapping Types of Information and Information Systems to Security Categories
- High-Value Asset designations and related frameworks like Primary Mission Essential Functions
- Freedom of Information Act exemptions

There are 125 CUI categories. As with any categorization scheme, assigning data to the proper CUI category requires subject matter expertise and management controls to limit improper categorization. The handling requirements associated with various CUI categories can be very specific - making automation difficult. Data labeled as CUI is “sensitive”; however, there is no scale to determine the degree to which the data is sensitive. From an information security standpoint, CUI data is all considered sensitive even as the risk of disclosure varies across categories. For example, consider a data asset with 100,000 unique identities, social security numbers, and bank accounts are as much CUI as is a single legal brief in a civil case.

CUI may make logical sense for categorizing data at the system as well as data asset level until you consider the inherent temporal and contextual complexity required to properly categorize the data.

⁹ Data Classification Practices

¹⁰ See E.O. 13556 Controlled Unclassified Information, (Nov 4, 2010)

The same piece of data could be CUI or not depending on the context and time frame. For example, a system may not contain any CUI; however, when an analyst combines that data with another data set – the resulting product could be CUI. Consider a list of vendor addresses in system A while System B has a list of vendors that bid on a procurement. A data analyst combines data from the two to determine the relationship between a company’s address, the bidder, and proximity to DOI lands. The resulting data product is now CUI – proprietary information.

Security analysts and data analysts also benefit from understanding the flow of data across organizations and systems. Information Sharing Agreements (ISA) document the interconnection between an agency system and an external system. Registering the details of which data and the flow of that data between systems will provide more actionable intelligence to data analysts and security analysts.¹¹ Most data inventories, as well as information assurance repositories, do not capture the ISA information in formats easily queried and analyzed. Data analysts benefit from improved documentation and understanding of data flows managed through a data inventory. Capturing data provenance through data inventory processes also facilitates better data governance. Before data is registered into enterprise data inventory, data security must also be addressed.¹²

Enterprise data inventories should capture data asset metadata at multiple levels. Agencies should consider capturing metadata in their enterprise inventories at the most granular level, the data element, and then aggregating knowledge upwards to the data asset as well as the relationships between those data assets, IT systems, applications, and the information products they derive. Linking enterprise data inventory processes with data profiling processes should help organizations populate their data inventories with the necessary information required for data and security analysts. Data profiling enables organizations to better understand the nature, characteristics, and volume of their data holdings. Data Inventories should capture the definitions of each data attribute for each data asset. Data custodians should link these individual data attribute definitions to the underlying semantic concepts the data represents.

Linking definitions to their underlying semantic meaning is a requirement of machine readability and will minimize miss-categorization of data as well as increase the interoperability of the data. A better understanding of the underlying data elements improves search, discovery, and data use while also improving overall data asset and system categorization. Various commercial off-the-shelf software products can assist an organization’s data profiling efforts using semi-automated approaches. Capturing and benefiting from semantic approaches requires a segment of the data management workforce who understands semantic concepts and standards and can guide the implementation.

Linking data profiling results to data inventories provides data analysts the information needed to access the fitness of use of specific data while also helping security programs better assess and mitigate the risks associated with sensitive data. Optimally the data profiling processes can be semi-automated to maintain a near up-to-date record count associated with an organization’s sensitive

¹¹ <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-47r1.pdf>

¹² DAMA-DMBOK: Data Management Body of Knowledge, 2nd Edition by DAMA International, dama.org

data. Understanding and capturing the full array of sensitive data characteristics will facilitate improved algorithms needed to detect the movement of sensitive data across the enterprise.

CDOs should collaborate with their Department's Senior Information Security Officer and any other relevant security officials to include a CDO review step during the development and review of the Agency's Information Sharing Agreements (ISA) and capture the relevant information from ISA into data inventories.

7. Recommendations

Recommendations to OMB

1. *Issue Phase 2 Guidance*

2. *Continue providing agencies flexibility to set Inventory priorities*

Agency data inventory priorities will vary according to their mission priorities. Aligning data inventories to the missions they support will re-enforce their value and garner the support needed for long-term sustainment.

3. *Establish metadata standards governance*

Ideally, the federal enterprise data catalog should use a single, cohesive metadata standard with a defined standards profile. The Data Catalog Vocabulary (DCAT) v.2 is one such metadata standard and is the logical starting point. The World Wide Web Consortium (W3C), an open international standards organization, maintains DCAT. DCAT has widespread software industry adoption, and the standard meets a broad spectrum of data documentation requirements. The W3C is continually adding extensions to DCAT v2 which provide coverage of geospatial and other unique needs. To make the standard more applicable to federal agencies and their constituents, data inventories should adopt a standardized profile to complement the DCAT v2 standard. Profiles are a way to unify vocabularies used in the standard as well as define constraints. The use of a common federal profile would improve data asset search, discovery, and appropriate use.¹³

4. *Reward and encourage inventory progress*

5. *Reinforce the use of inventories rather than 1 time or recurring data calls*

The Administration can strengthen data inventories by cross-referencing data inventories in other policies and required data calls. For example, if the Administration establishes a new policy related to climate data and requests agencies submit a list of data assets that supports that topic, instead of asking agencies to submit a list of climate-relevant data assets they should require agencies to update their data inventories with climate data, and then submit a list of links to all the relevant assets in their data inventories.

6. *Coordinate a review with NARA of the CUI Program to better automate data classifications*

OMB and NARA should evaluate the effectiveness of the current Controlled Unclassified Information categories to determine if the number of categories could be reduced to improve automation and agency implementation.

¹³ <https://w3c.github.io/dxwg/profiles/#introduction>

Additional research should be funded to help identify common patterns within a collection of data attributes that could be used to automate and refine data asset and system classifications.

Recommendations to the CDO Council (CDOC)

1. *Coordinate the minimal update to the federal metadata standard*

Project Open Data schema is the current metadata standard and must be updated to be in sync with the international standard it's derived from and to capture additional required elements from the Open Government Data Act. The standard should adhere to the FAIR Principles¹⁴. A wealth of educational materials and best practices are available through <https://www.go-fair.org/resources/more-on-fair/>. To improve search and discoverability the metadata standard should utilize a set of controlled vocabularies for key concepts such as data classifications, organizations, public law, and regulations, location, and budget codes. Metadata must be machine-readable and capture the meaning of the data asset at the most granular levels. The standard should require the use of Universally Unique Identifiers (UUID) for data assets and metadata. UUIDS will improve search, discovery and allow agencies to measure data asset usage.¹⁵ OMB should consider establishing an all-government approach to UUID.¹⁶ (see UK and EU policies and approach)¹⁷

2. *Learn what it will take for Federal agencies to transition.*

3. *Identify the resource needs to update the Web properties of data.gov, Geoplatform, Standard Application Portal to implement the standard.*

4. *Recommend a set of metrics to track inventory progress.*

5. *Recommend a more permanent metadata standards governance structure.*

6. *Establish a formal mechanism to collaborate with NIST in the development of a sensitive data classification scheme.*

The CDOC should establish a formal mechanism to collaborate with NIST to develop a sensitive data classification scheme. Agencies should participate in the NIST NCCoE's¹⁸ project to examine different approaches to data categorization and the implementation of protections based on those categorizations.

Improved and consistent data classifications schemes will benefit cyber security and improve data asset discovery. Any new data classification scheme should include crosswalks to link to existing classification schemes such as CUI and FIPS 199 information categories. All data classification schemes should be published as open data assets to improve interoperability. CDOC in partnership with CISA should better define the appropriate level to tag data assets and

¹⁴ <https://www.nature.com/articles/sdata201618>

¹⁵ <https://fairtoolkit.pistoiaalliance.org/use-cases/adoption-and-impact-of-an-identifier-policy-astrazeneca/>

¹⁶ <https://www.gov.uk/government/publications/linked-identifier-schemes-best-practice-guide/linked-identifier-schemes-best-practice-guide>

¹⁷ <https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic>

¹⁸ <https://www.nccoe.nist.gov/projects/building-blocks/data-classification>

establish a formal ontology that links common concepts such as data asset, IT system, application, data product in a manner most useful for implementing zero trust.

Recommendations to CDOC Workgroups

1. *Continue to emphasize data literacy, especially in positions key to the inventory such as data stewards and data producers.*
2. *Identify inventory requirements to facilitate data sharing.*
3. *Continue to foster dialog across other councils for inclusive data inventory requirements.*
4. *Revise and renew Data Inventory Work Group Charter and objectives.*

Recommendations to Agencies

1. *CDOs need to stress the internal mission value of data inventories*
Inventories benefit agencies' missions and are more than open data catalogs. Agencies should manage their inventories for internal value and achieve data inventory compliance as a by-product of sound data management. CDOs should demonstrate the value of a data inventory with a series of quick wins. CDOs should align data inventory efforts to produce immediate value to the organization by iteratively collecting and curating metadata around subjects that are relevant and valued by the mission and stakeholders.
2. *CDOs should continue to increase data literacy across the organization.*
Organizations with greater degrees of data literacy appreciate and value the access to and use of data and as a result value a comprehensive data inventory. Improving the data literacy of those positions key to the inventory such as data stewards and data producers may be a fruitful place to start.
3. *Agencies need to establish incentive structures for good data asset stewardship*
Data stewards are essential to maintaining a useful data inventory; however, many stewards lack any reward for well-documented data assets and may not understand their stewardship responsibilities.
4. *CDOs should collaborate with their CIO to establish a technology vision inclusive of data inventories, data cleansing, data enablement, and data asset usage.*
CDOs should collaborate with their CIO and CTO to establish a high-probability foundational technology roadmap that aligns and supports the vision. This roadmap should:
 - Identify technology investments and the supporting people and processes necessary to succeed at foundational elements.
 - Consider who and what capabilities are needed to support the technology as well as the processes.

- Consider the organizational data management maturity when investing in new data management technology. Be cautious of purchasing capabilities that are beyond the organization's data management maturity.
- As organizations strive to better *leverage data as an asset*, understanding and aligning with commercial data management innovations may yield considerable opportunities for cost avoidance and lay the foundation for significantly increased insights. The technology and processes for curating a data inventory do not exist in a vacuum, rather they must integrate into the agencies existing and planned IT portfolio.
- Organizations should invest in easy-to-use metadata creation and management tools that support data validation and workflows. CDOs should integrate these tools into the agency's data lifecycle management processes.
- Agencies need greater investment in metadata automation and integration to reduce the burden of collecting and maintaining metadata

5. *CDOs should align data inventory processes and management controls with the full data lifecycle and not just start and stop at data publishing.*

Too often metadata creation doesn't occur until an agency disseminates a data asset. Unfortunately, this creates an added burden for whoever needs to publish the data asset. A better practice is to start capturing and documenting metadata at the beginning of the data lifecycle. This spreads the burden out and improves data quality by capturing relevant business information directly from the subject matter experts.

6. *CDOs should engage internal and external data user communities to understand and address data demand signals*

7. *CDOs should consider linking and connecting additional attributes needed to support other Information Management functions within the agency and work with the applicable agency officials. Examples include:*

- Records management: *What is the record retention schedule for these data?*
- Data Management/Stewardship: *Who is responsible for maintaining the data asset?*
- Privacy: *Does the data asset contain PII? Is the data asset covered by a System of Record Notice (SORN)?*
- Security: *Do the data contain sensitive information? Are the security measures in place for appropriate access management?*
- Legal: *Are there copyright or licensing restrictions on the data? What are the publication rights on the data?*
- Open Data: *Is the data asset available to the public in an open format? How often is the data asset searched and accessed? Which stakeholders use the data asset?*

- Information Collection: *Which if any Information Collection Request is associated with the data asset? What law or regulation requires the collection of the data contained in the data asset?*

8. Definitions

Data: recorded information, regardless of form or the media on which the data is recorded. (Source: [44 USC § 3502\[16\]](#))

Data Asset: a collection of data elements or data sets that may be grouped together. (Source: [44 USC § 3502\[17\]](#))

Data Governance: the exercise of authority, control, and shared decision-making (planning, monitoring, and enforcement) over the management of data assets. (Source: DAMA Dictionary of Data Management, Data Management Association [[DAMA International](#)])

Data Lineage: A description of the pathway from the data source to their current location and the alterations made to the data along that pathway (Brackett 2011)

Data Provenance: Provenance applied to the organization's data resource. (Brackett 2011)

Data Stewardship:

1. The formal, specifically assigned, and entrusted accountability for business (non-technical) responsibilities ensuring effective control and use of data and information resources.
2. The formal accountability for business responsibilities ensuring effective control and use of data assets. (DAMA-DMBOK Guide, 1st edition, pg. 39)

Evidence: information produced as a result of statistical activities conducted for a statistical purpose. (Source: [44 USC § 3561\[6\]](#))

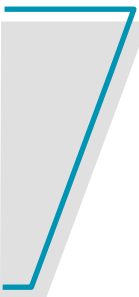
Machine Readable: when used with respect to data, means data in a format that can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost. (Source: [44 USC § 3502\[18\]](#))

Master Data: Master Data is the consistent and uniform set of identifiers and extended attributes that describes the core entities of the enterprise including customers, prospects, citizens, suppliers, sites, hierarchies, and chart of accounts. (Source: DAMA Dictionary of Data Management, Data Management Association [[DAMA International](#)])

Metadata: structural or descriptive information about data such as content, format, source, rights, accuracy, provenance, frequency, periodicity, granularity, publisher or responsible party, contact information, method of collection, and other descriptions. (Source: [44 USC § 3502\[19\]](#))

Open Government Data Asset: a public data asset that is machine-readable; available (or could be made available) in an open format; not encumbered by restrictions, other than intellectual property rights, including under titles 17 and 35, that would impede the use or reuse of such asset; and based on an underlying open standard that is maintained by a standards organization. (Source: [44 USC § 3502\[20\]](#))

Reference Data: Any data used to categorize other data, or for relating data to information beyond the boundaries of the enterprise. Earley, S. (2011). *The Dama dictionary of data management*. Technics Publications.



If you have questions or would like more information about the case studies, contact cdocstaff@gsa.gov.



www.cdo.gov | cdocstaff@gsa.gov.