Alation

EVOLVING GOVERNMENT: LEVERAGING A DATA CATALOG FOR A DATA-DRIVEN PUBLIC SECTOR

Most of our nation's Founding Fathers were farmers and merchants — not data scientists and engineers. Yet we can be certain they'd approve data-driven efforts to strengthen our democracy and improve how federal and local governments serve their citizens.

Today's most successful companies are data-driven; today, Public Sector agencies are likewise seeking to better leverage the vast amounts of data they receive, whether it's from traditional sources like census figures or modern inputs like "smart" city trash cans and parking meters.

To achieve this goal, Public Sector agencies must do 3 things:

1. **Manage their increasing volumes of data.** The amount of data collected is increasing in every industry, including within the Public Sector. Factors driving this increase include the digital transformation of government processes and services, IoT devices (like those smart trash cans), social media and online platforms that record public feedback, and increased sharing of data between agencies and agencies.

2. **Empower more people within the organization to access the right data for actionable insights.** Not everyone in government is a data expert. So enabling self-service and leveraging artificial intelligence (AI) and machine learning (ML) for advanced analytics are key for process optimization, which leads to cost optimization.

3. **Ensure the security of private data.** One of the better-known and influential state statutes designed to do this is the California Consumer Privacy Act (CCPA). CCPA further strengthened privacy protections and expanded consumer rights with the California Privacy Rights Act (CPRA), which went into effect on January 1, 2023.

Regardless of the specifics of the privacy statutes pursuant to your organization, implementing a data governance framework will help to ensure data compliance.

Data by the People, Data for the People

The above describes how data is accessed for the public, but there's also a need for data access by the public to foster transparency, accountability, and public participation.

The Open, Public, Electronic and Necessary (OPEN) Government Data Act was enacted in 2019 to make government data more accessible, usable, and open to the public. Data.gov is a repository for Federal, state, local, and tribal government information offering access to more than 250,000 datasets (as of early 2023), from motor vehicle crashes to fruit and vegetable prices.

> ***"When we look at the practices across agencies — how they are structured, how they use data, apply insights, measure the outcomes achieved, and learn from them — less than 10% are truly insights driven. Some industries demonstrate less maturity. In government, for example, less than 5% of agencies exhibit advanced insights-driven practices, and just over a quarter have intermediate practices."***
>
> *Chief Data Officers: Evolve Your Teams to Accelerate Impact from Data Insights.*

Despite the increased collection of data, at all levels of government, people still struggle to find and access the data they need. They waste time searching endlessly through rows and columns and needlessly duplicating efforts in silos. Eventually, as a result, they lose trust in data. Data becomes a problem rather than part of the solution.

This is where a data catalog can help.

Key Benefits of a Data Catalog

An automated data catalog allows people in agencies to easily find data, explore its context, and track the human path of data usage, as well as its transformation via lineage. The data catalog creates a living inventory that continually updates with information on how data is being used, enabling users to know exactly what data they can access and whether it's relevant to their needs — a critical first step toward open and efficient data use.

A data catalog enables Public Sector agencies to:

- Make data accessible to stakeholders and subject matter experts
- Maintain transparency and make data easier to understand
- Make data analysis more productive, spurring the insights that lead to positive changes in public policy and public services

Without a way to find data, understand its meaning, and share that understanding across the organization, the value of data is limited. Data catalogs act as a single-source of reference for all data assets, ultimately empowering smarter decisions and greater insights. In this way, the catalog creates the foundation for a data-driven organization.

Data catalogs allow Public Sector agencies to address these specific challenges:

- **Manual inventories don't scale.** Most government entities today make due with Excel-based manual inventories of data assets. In contrast, a data catalog connects to all of an organization's data sources and BI intelligence tools to automate the registration of all accessible data assets in a single place.
- **Policy decisions are more difficult without a glossary.** With more access to data, more people can create reports and surface insights. But even a simple question can surface numerous answers, depending on how the question is framed. This results in agencies producing many different versions of "truth." A data catalog standardizes definitions of key metrics. This ensures that debates center around the real policy issues — rather than whether the data is trustworthy.

- **Self-service analytics adoption requires curious minds.** Data catalogs help make data more open and interesting, engaging SMEs to share their knowledge with the rest of the organization by endorsing data sets, annotating projects or engaging with other users in one easily referenceable place. This collaborative approach to analytics supports improved insights.

Driving Adoption of Data

Data catalogs are applicable to a wealth of applications. Some of the top reasons for adopting data catalogs in Public Sector agencies are:

- **Moving from data to insights.** With the onset of big data, agencies have sought to empower more people to access all of the newly data made available. Self-service environments were created to allow anyone to access data, create a report, and derive an insight. However, early efforts hit stumbling blocks. While the ability to create and store data had increased exponentially, our ability to derive insights had not.

  Public Sector agencies have learned from these early missteps and are implementing data catalogs to enable a broad set of users to easily find and query data from one single source of reference. A data catalog with machine learning automatically indexes all of an organization's data sources and gathers knowledge on the data. Query time can be cut in half and non-technical users can find the data they need without needing to go to IT for help.

- **The trend toward open data and maximum transparency.** The OPEN Government Data Act requires all federal government information to be machine-readable and accessible through open-source code and pushes agencies to leverage their datasets to support policy development. Even non-federal agencies have taken up OPEN Data initiatives to improve the lives of citizens in various ways, from saving

time when responding to FOIA (Freedom of Information Act) requests, to creating more efficient government, to powering innovative applications and services.

But while OPEN data initiatives offer many benefits, there are still barriers to successful implementation, including:

- Managing data as an asset and inventorying what data is available
- Making data available in a timely manner will still ensuring consistency, protecting sensitive information, and making the data easy to understand for citizens and internal users alike
- Making the data available to citizens in a way that is transparent and effective

Data catalogs can remedy this complexity by creating one point for accessing and managing data, enabling agencies to enforce consistency without having to migrate to new systems or invest in new architecture. Two other compelling reasons public agencies favor data catalogs include:

- **The potential for AI and ML to improve efficiency and effectiveness.** AI and ML can accelerate a range of services, including welfare payments, immigration decisions, and fraud detection; they can also be used for key activities like planning new infrastructure projects, answering citizen queries, triaging healthcare cases, and establishing drone paths.

  However, AI and ML also present risks. Once automation begins, the algorithms that power AI and ML are difficult to manage and monitor. Without a way to see what data the algorithms are pulling, algorithms are a black box, creating the potential for misuse and bias. For example, as mathematician Cathy O'Neil describes in her book, *Weapons of Math Destruction*, algorithms scanning resumes (which were intended to find the best-qualified job candidates in a non-discriminatory way) instead inadvertently recreated racial and gender bias in hiring decisions.

For AI and ML to be effective in the Public Sector, there must be transparency into the data the algorithms are using — and checks in place that allow humans to step in and make key decisions when necessary. Data catalogs can help alleviate the risks of negative occurrences like "algorithmic bias" by tracking the data sets — understanding which sets of data are getting fed into an algorithm, giving opportunity for regulators and analysts to be able to question any bias, look at the output of an algorithm, dig into details, and ensure quality output.

- **Data governance capabilities to ensure compliance.** "Data governance isn't just a maintenance task," notes the Forrester report *Predictions 2023: Data and Analytics*, "it's the foundation of business insights and data products." The report warns that a dedicated data governance team is critical, and notes that during the 12 months after the report was published (November 2022), "36% of agencies plan to bring on a head of data governance, and 32% of agencies will bring on a chief data officer."

  A data catalog provides a centralized and organized view of data assets, metadata, and governance-related information. This helps build a strong foundation for effective data governance practices to promote transparency, collaboration, and compliance — while also fostering better data management and data-driven decision-making.

Case Study: How a Data Catalog Formed the Foundation for Data Governance at the State of Tennessee

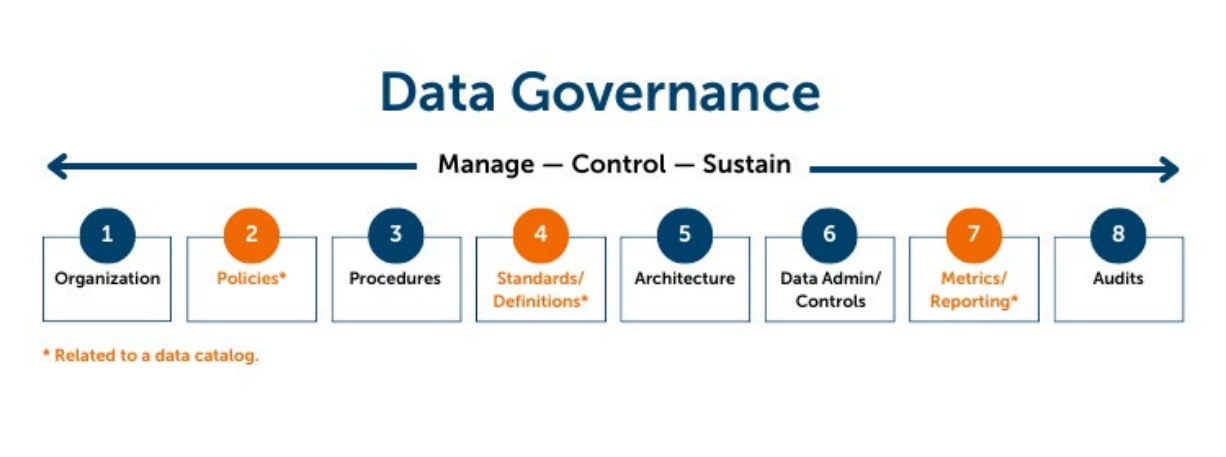
In his 2021 State of the State address, Tennessee governor Bill Lee declared the State's mission "to see the needs of our neighbors around us — every single one of them — and commit to serving them."

For Kim Wienzierl and Amber O'Connell, the State's chief data officer and director of data governance, respectively, that meant making better decisions, based on data, that impact lives of Tennesseans.

The State of Tennessee has 38 agencies, which Wienzierl compared to a business with 38 different product lines. They introduced a data catalog to:

- Instill uniformity in the data governance process
- Improve data quality
- Prevent audit findings related to data quality or data governance
- Share data across those 38 different agencies, meaning they need to understand what data is key or critical and get it into a governance framework
- Accelerate data-driven decision-making

In late 2020 they implemented a data governance framework with 8 elements:



This framework included a methodology and templates for instituting data governance across all the agencies. It was assumed that much of the State was doing informal data governance, but the framework was designed to formalize and complete the process, and built for reuse and completeness. Wienzierl said her team uses a data catalog for policies, standards and definitions, and metrics and reporting.

Why Did Tennessee Turn to a Data Catalog?

Amber O'Connell became the first person to serve as the director of data governance in 2018. "It was a bit like the Wild West between convincing people that we need a data governance program and actually building the program," she recalled.

With a lack of visibility, availability, and accessibility, the State's data assets were a mess. "It immediately became apparent that it's really hard to govern data if you don't know what data you have," O'Connell said.

The team chose a data catalog to inventory their existing data; the resulting, curated catalog helps people discover what data is available fast. This has helped O'Connell achieve her goal of driving culture change across the State to be more data fluid.

The second reason the team chose to implement a data catalog came down to standardization. With a catalog, they could instill enterprise standards for data management. This is critical, since they manage dozens of agencies operating as individual businesses, which creates inconsistencies. "A few of our systems have data dictionaries," O'Connell said. "Some of them are accessible; others are not. And few divisions have their business terms defined."

As an example, the Department of Mental Health has several different divisions, each of which defines "subacute" differently. So the answer to "How many many subacute patients do we have?" varies depending on whom you ask.

"The data catalog would standardize definitions and agreed-upon terms to really help us improve our data quality and trust in our data," O'Connell said. Once they made the business case for a data catalog, they went through the traditional procurement process and selected Alation.

> ***"The data catalog standardizes definitions and agreed-upon terms to really help us improve our data quality and trust in our data."***
>
> Amber O'Connell
> Director of Data Governance, State of Tennessee

Early Wins and a Roadmap for Success

A critical element of data governance involves implementing standards and definitions. These are the guidelines informing how to design data solutions, interpret information, evaluate performance, and ensure quality for data — from source to consumer. The goals of this phase of data governance include having everything documented and required for all fields in the data set and that safe sources of data are identified.

Data definitions describe the meaning of a data element in a way that is readily understood by a data consumer. "In the case of the 'subacute' example, the data definition now means the patient has been in bed for five days," O'Connell said. "The data definition is readily understood." Standardizing this language eases communication and unlocks interoperability across the wider State.

Data standards are documented specifications for a specific data element or sub element. They're derived from data models, data schemas, naming rules, and business rules, and can include things like valid values and acceptable ranges, such as:

- An age range between 0 and 120
- Specific formatting rules such as never allowing free-text fields
- Minimum data quality levels
- Required fields, such as first and last name

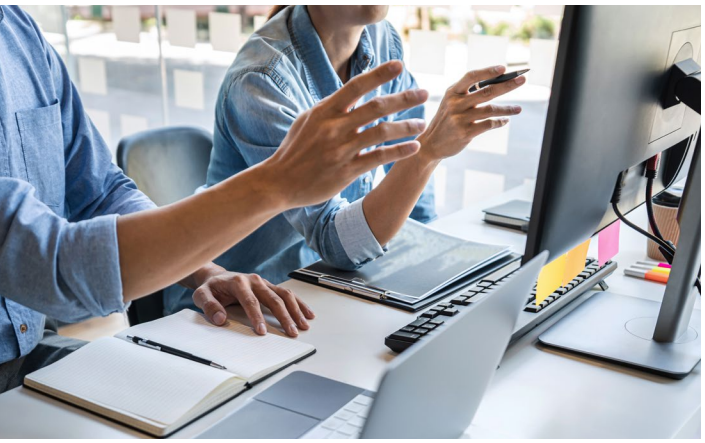
"You also want safe sources of data in case there are conflicting sources, as well as formulas for any derived data that's been created," O'Connell added. "If you do a report once a year and one year you wonder, 'How did we get there?,' that's a problem with data quality. Standards and definitions really help us."

Standards and definitions are housed in the data dictionary, which is stored in Alation with the business glossary. "You can import a data dictionary if you have one, and then can export it back out of it, if you need to," O'Connell elaborates. "Alation literally is a massive data repository, [and it] really helps us build and understand our data so we can know what our data models look like."

O'Connell and her team use Alation to classify and regulate all State of Tennessee data — which is vital to government agencies. "Alation really helps us designate which data should be public or which, like HIPAA and PII data, need to be restricted," she shares.

According to O'Connell, Alation was also critical for populating State business glossaries. After the privacy officer created and defined the office privacy regulations, any Alation user can see those details – and use them to inform their own work.

Plus, O'Connell noted, people can create new enterprise definitions, adding to the body of knowledge. "Let's say Mental Health defined a term and then the Department of Health saw the term and thought, 'Hey, that's what we do, too. Let's just use that definition,'" she elaborates. "Now you've got an enterprise standard that I didn't even have to talk to them about. They copied it themselves. That's what's really great about Alation and the glossaries."



For the State, the final element of data governance is audits. O'Connell stressed that she'd never seen audit preparation happen in any other data governance framework. She sought to make the process itself self-auditable and fully documented in one place. She also wanted the internal audit to be ready for external auditors, as well.

"The self-audit process documents any gaps in our data governance so you can address them immediately," O'Connell said. "It gives ownership of the process to the data owners. They feel more knowledgeable, have more trust, and are better enabled to share data within or outside of their organization. Alation absolutely helps us with our audits."

Alation Data Catalog contains agency policies and procedures, standards and definitions, and data flow diagrams through the lineage. It's also where all the data is classified and owners and stewards are identified, so one can see who has access to what data.

The Data Governance App includes features like Policy Center, which automates policy control and synchronization for Snowflake data. Policies can be applied and enforced on data (in the Data Cloud) at the time of use, helping meet security and compliance objectives. The ability to create and enforce policies is critical, as policies create the foundation of a data governance program, which provides real-time guidance for new users, and creates visibility for managers into how data assets are managed over their lifecycle.

These policies cover 5 key focus areas:

1. **Authoring.** The when, how and by whom data may be created, changed, or deleted.
2. **Access.** Which people or systems are authorized to see and get the data?
3. **Usage.** What are the authorized uses for the data and how are they mapped back to the authorized users?
4. **Maintenance.** How is the data maintained in the source systems and backed up for recovery?
5. **Retention and storage.** How long must the data be kept in what format and any defined lean times for retrieval?

An Optimistic Future

"Although we're fairly new at this, we feel like we've got a pretty solid plan," Wienzierl said. We have a number of agencies lining up, wanting to participate. They've seen the value and they're ready to go."

Wienzierl and O'Connell created a "data initiative status" chart that maps the status of each agency through the data governance framework, adoption of Alation Data Catalog, and use of the Snowflake data lake.

"We were very delighted to learn that you can plug Alation right into Snowflake and catalog it as well," Wienzierl said. "I've built data lakes in the past and it doesn't take very long for them to turn into a data swamp – if you don't know who the owners are or what the purpose of the data is or exactly what you have in there. We're going to have a lightweight data governance

process running against everything that is in Snowflake, and it will all be documented in Alation Data Catalog."

With the help of Alation and this data governance framework, Wienzierl, O'Connell, and their team have empowered more employees in the State with much-needed data fluency, which in turn serves the needs of their citizens.

A Data-driven Public Sector

Organizations everywhere are collecting, ingesting, and storing more data than ever before. But the value of this data is greatly diminished without a way to derive insights from it — and for Public Sector agencies, the deluge of data is only growing.

Whether your goal is to improve existing processes, promote greater transparency and dialogue, or create new initiatives to positively impact citizens, improving how you manage data can transform your ability to effectively serve the public. A data catalog is an integral platform by which to curate, catalog, and more effectively govern that data to this end.

Alation Data Catalog provides a single source of reference that increases the efficiency of data efforts, helping analysts find, understand, and trust the data they need in order to power the kinds of initiatives that improve the lives of constituents.

About Alation

Alation is the leader in enterprise data intelligence solutions, enabling data search and discovery, data governance, data stewardship, analytics, and digital transformation. More than 450 enterprises in myriad industries build data culture and improve data-driven decision-making with Alation, including local, state, and federal government agencies such as the State of Tennessee, the Ohio Department of Job and Family Services, and the Centers for Medicare and Medicaid Services (CMS). Learn more about how Alation helps Public Sector agencies transform data systems to facilitate mission delivery, fortify data integrity, and elevate service standards at alation.com/solutions/public-sector.