



Data Management Workgroup

Burt.Walsh@ast.myflorida.com

Michael.Avello@ast.myflorida.com

October 25, 2018

Open Data



“Open data enables governments, citizens, and civil society and private sector organizations to make better informed decisions. Effective and timely access to data helps individuals and organizations develop new insights and innovative ideas that can generate social and economic benefits, improving the lives of people around the world.”

(Open Data Charter, 2018)

Principles, Open Data Charter

<https://opendatacharter.net/principles/>

Project Open Data

- U.S. Federal Office of Management and Budget, Project Open Data, 2018
<https://project-open-data.cio.gov/>
- Framework designed around to help institutionalize the principles of effective information management
- Promotes interoperability and openness

Project Open Data

- This has value for all state data, not just open data
- U.S. Federal Office of Management and Budget, Project Open Data, 2018
<https://project-open-data.cio.gov/>
- Principals based upon many experts
- Technologies including open source



Open Data Guidance/Approach

Survey of Other States

- Arkansas, Hawaii, Illinois, Indiana, New Jersey, New York, Oregon, Texas, Virginia, Washington
- Other features and guidance
- Consistent with Project Open Data
- Licensing, legislation and handbooks



Open Data Guidance/Approach

Principals of Open Data

- Public
- Accessible
- Described
- Reusable
- Complete
- Timely
- Managed Post-Release



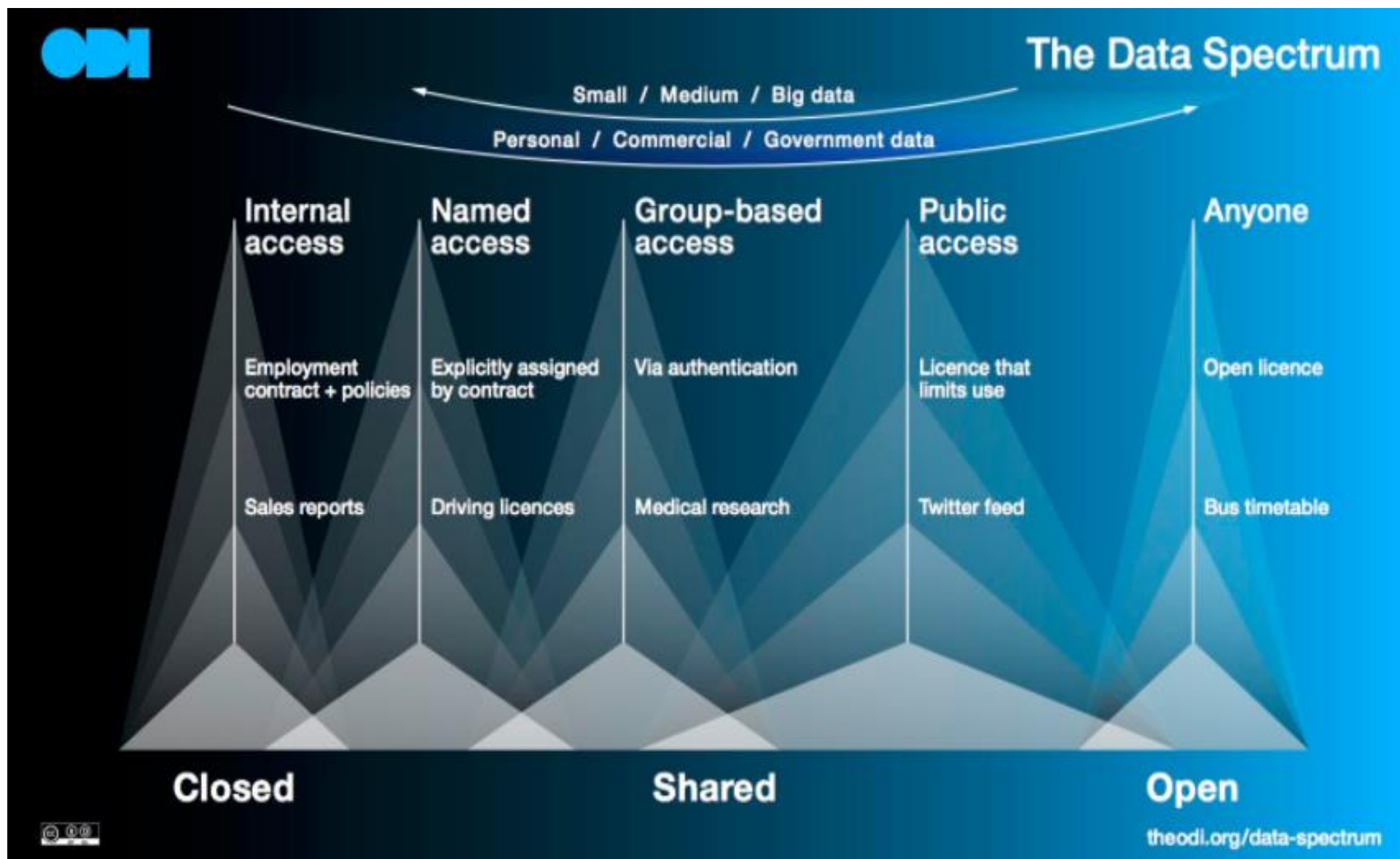
Open Data Guidance/Approach

Public

- “It is the policy of this state that all state, county, and municipal records are open for personal inspection and copying by any person. Providing access to public records is a duty of each agency.” (Section 119.01 (1), F.S.)
- Presume openness to extent permitted by law
- Privacy concerns (PII, PHI)
- Compliance concerns (HIPPA, FERPA)
- Exceptions like those under Section 119, F.S.

Open Data Guidance/Approach

Classifying Data Sets





Open Data Guidance/Approach

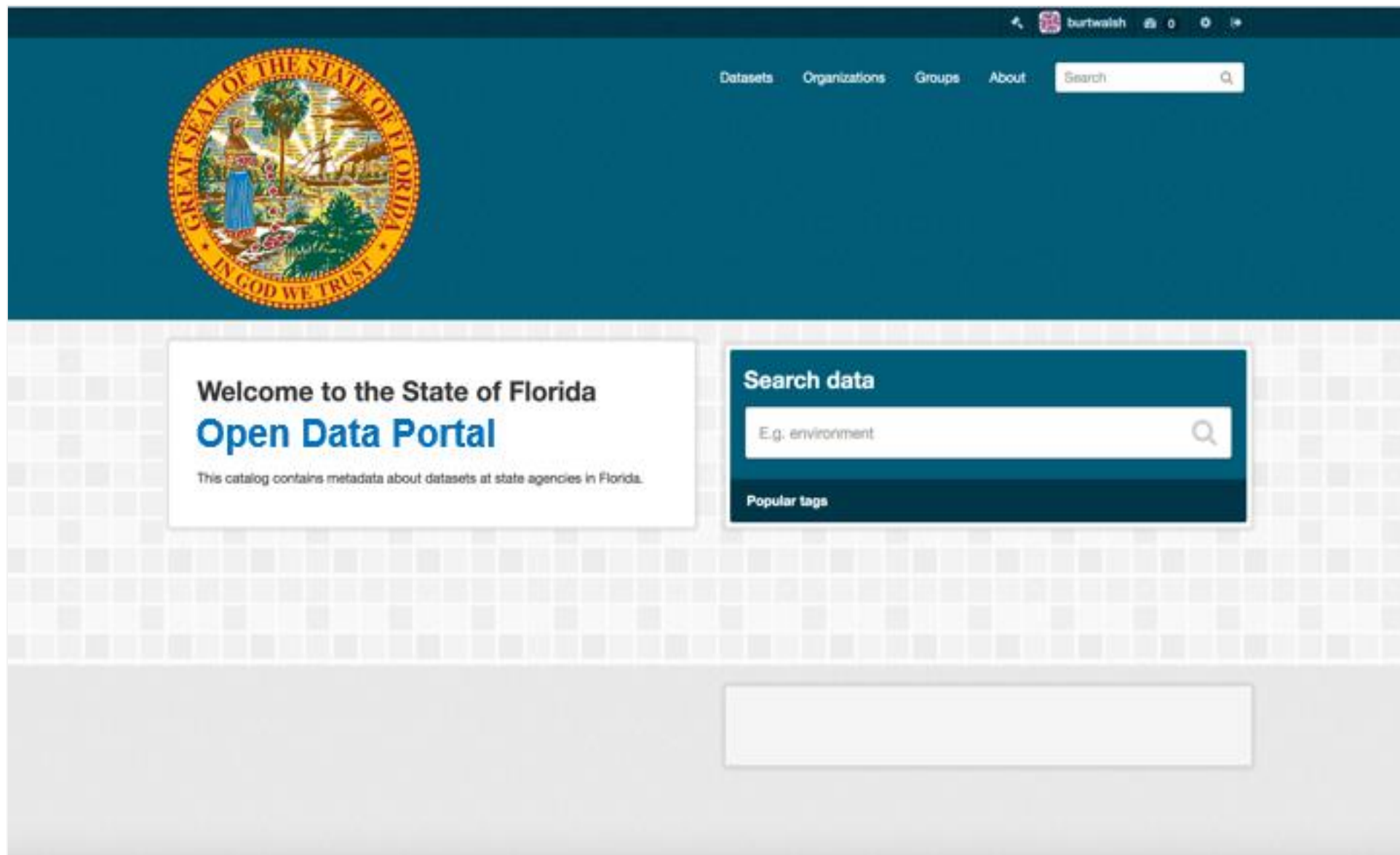
Open Data Portal

- Comprehensive Knowledge Area Network (CKAN)
- Web based/Searchable
- Secure/Users/Rights
- Supports upload of datasets and metadata about data sets
- Open source/Open standards
- Can be integrated with Identity as per Florida Rule Chapter 74-5 (SAML)



Open Data Guidance/Approach

Open Data Portal





Open Data Guidance/Approach

Accessible & Described

- Open formats (XML, JSON, GeoJSON)
- Open data format standards:
 - FHIR® – Fast Healthcare Interoperability Resources
 - National Information Exchange Model
 - Standards Information Base, The Open Group
- Metadata (context information and assumptions)
- Project Open Data metadata standard, natively supported by CKAN and other open data portal technologies

Reusable & Complete

- Data made available in an open, ubiquitous, machine-readable format
- Enables citizens to use their software tools and apps to reuse the data for their particular needs
- Complete means the availability of underlying metadata that references the source of aggregated data (without exposing the actual source data)
- Includes master (reference) data and the use of a Business Glossary



Open Data Guidance/Approach

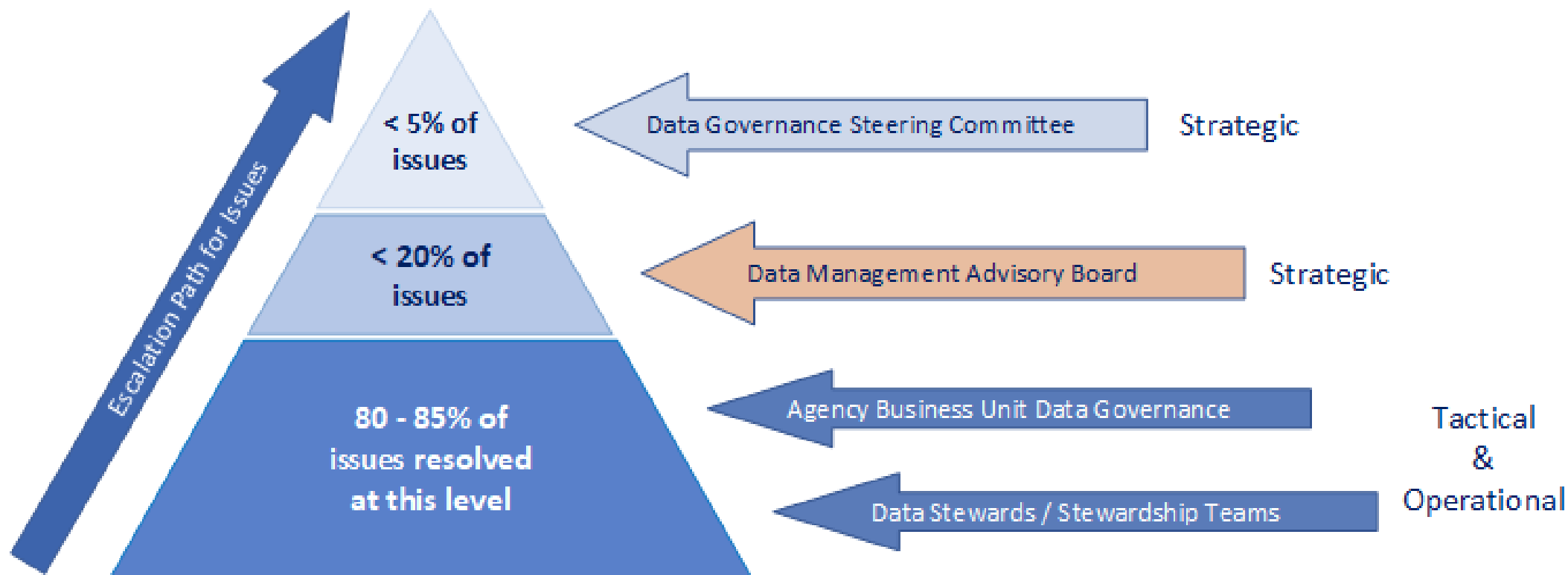
Timely & Managed Post-Release

- Timeliness ensures that the data is as relevant as needed for decision-making
- Open data where data can be used in many unknown contexts requires update as frequently as possible
- The metadata associated with the dataset should have the data steward's contact information
- Data Stewards should provide timely updates and visibility into corrections requested by the public (governance)

Data Management & Governance

- Disciplines and processes supported by full agency
- Requires Business and technical data stewards to ensure data is being governed according to policy and effectively controlled and used
- Data governance is structured to function at strategic, tactical, and operational levels to address data management issues

Data Management & Governance



Data Quality

Good quality data is fit-for-purpose and is measured against these quality dimensions:

- **Completeness:** Are all datasets and data items recorded?
- **Uniqueness:** Is there a single view of the dataset?
- **Timeliness:** Does the data represent reality from the required point in time?
- **Validity:** Does the data match the business rules?
- **Accuracy:** Does the data reflect the real-world entities that it represents?
- **Consistency:** Can we consistently match the dataset across data stores?

Data Quality



The Plan-Do-Act-Check Cycle

- **Plan:** Assess the scope, root causes, impact, and priority of known issues.
- **Do:** Address the root causes of issues and plan for ongoing monitoring of data.
- **Check:** Actively monitor data quality as measured against quality dimensions (see above).
- **Act:** Address and resolve emerging data quality issues.

Each cycle promotes continuous improvement and typically begins when data quality measurements fall below certain thresholds or when business rules or requirements change.

Metadata Management

- For any organization to be data-driven, it must be metadata-driven.
- Metadata is “data about data” ... it describes the data itself and the concepts and meaning that the data represents
- Metadata helps agencies to understand their data, systems, and business processes.
- It is the common thread that enables the ability to process, maintain, integrate, secure, audit, and govern data.

Business Glossary

- Used to document and store business concepts and terminology, definitions, and relationships between those terms
- Business Glossary content is not static—its implementation should follow a basic set of procedures centered around metadata elements:
 - **Components** (how metadata is captured, categorized, and used)
 - **Types** (Descriptive, Structural, Administrative)
 - **Categories** (business, technical, operational)
 - **Subject Areas** (operations, business intelligence, performance management, etc.)

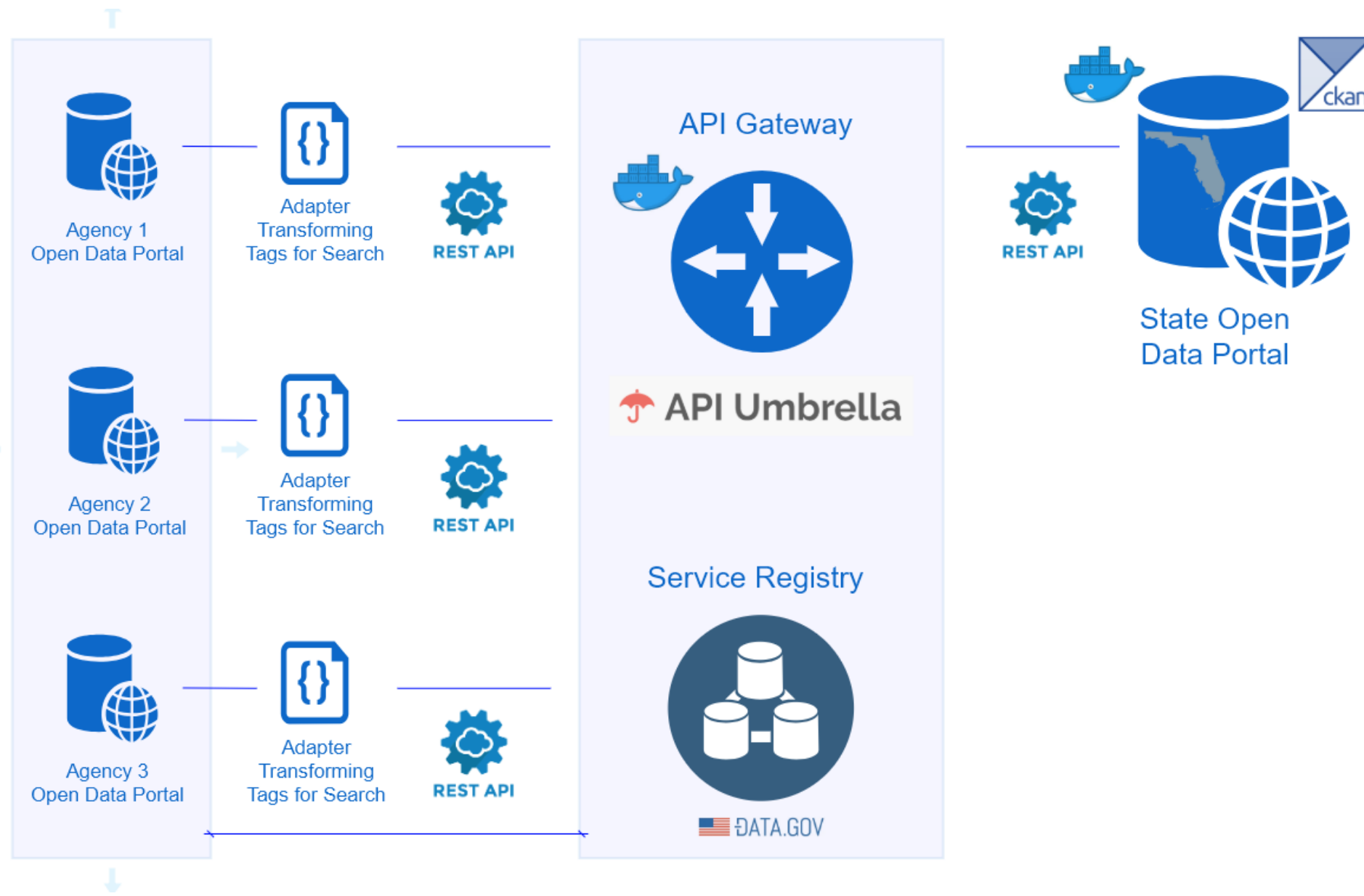
Data Extraction, Transformation and Loading (ETL)

- Data can be uploaded to CKAN or referenced if presented through an interface such as a REST interface.
- Significantly impacted by data quality, but ETL is also an opportunity to fix and enhance data quality by:
 - Using rules documented in the Business Glossary to place value constraints on the data
 - Mapping raw data values to reference values to establish validity and assist record linkage
 - Enriching or complementing the data to provide more context and value, such as adding longitude and latitude coordinates to complement street address data

System Modernization and New System Development

- Consistent with Section 119, F.S. and Project Open Data guidance
 - “When designing or acquiring an electronic recordkeeping system, an agency must consider whether such system is capable of providing data in some common format such as, but not limited to, the American Standard Code for Information Interchange.” (Section 119.01(2)(b), F.S.)
 - “Providing access to public records by remote electronic means is an additional method of access that agencies should strive to provide to the extent feasible.” (Section 119.01(2)(e), F.S.)
- Metadata provides a cost-effective means to manage record exemptions (see Section 119.07(1)(e), F.S.)

Open Data Guidance/Approach





Open Data Portal

Open Data Guidance/Approach

A screenshot of a web browser displaying the Open Virginia data portal. The browser's address bar shows the URL 'data.openva.com/dataset'. The page has a dark blue header with the 'Open Virginia' logo and the tagline 'Creating an API for the commonwealth.' Below the header, there is a navigation bar with links for 'Datasets', 'Organizations', 'Groups', and 'About'. A search bar is also present. The main content area shows a search results page for '141 datasets found'. On the left, there are filters for 'Organizations', 'Groups', and 'Tags'. The 'Groups' filter is expanded, showing categories like 'Virginia General As...', 'Transportation (3)', 'Courts (3)', 'Code of Virginia (2)', 'Regulations (1)', and 'Elections (1)'. The 'Tags' filter shows 'GIS (28)', 'IRS (24)', 'Water (23)', and 'Income Taxes (23)'. The search results list includes 'Virginia Building Footprints' and 'County-Level Pesticide Use Estimates - 1993' and '1992'. Each result shows a brief description and available file formats like 'geojson', 'CSV', 'tsv', and 'TXT'.

Open Data Portal – Adding a Dataset

The screenshot displays the Open Data Portal interface. At the top, a dark blue header features the Great Seal of the State of Florida on the left and navigation links for Datasets, Organizations, Groups, and About on the right. A search bar is also present. Below the header, the breadcrumb trail reads: Home / Organizations / AgencyForStateTechnology. The main content area is divided into two columns. The left column contains the AgencyForStateTechnology logo, which includes a stylized map of Florida and the text 'ASTA AGENCY FOR STATE TECHNOLOGY'. Below the logo, it states 'AgencyForStateTechnology' and 'There is no description for this organization'. It also shows 'Followers 0' and 'Datasets 0' with a green 'Follow' button. The right column has tabs for Datasets, Activity Stream, and About, along with a 'Manage' button. A dark blue 'Add Dataset' button is prominently displayed. Below this is a search bar labeled 'Search datasets...'. The main content area of the right column displays 'No datasets found' and an 'Order by: Relevance' dropdown menu.

Great Seal of the State of Florida

Datasets Organizations Groups About Search

Home / Organizations / AgencyForStateTechnology

ASTA
AGENCY FOR STATE TECHNOLOGY

AgencyForStateTechnology
There is no description for this organization

Followers 0 Datasets 0

Follow

Datasets Activity Stream About Manage


+ Add Dataset

Search datasets...

No datasets found

Order by: Relevance

Open Data Portal – Adding a Dataset



[Datasets](#) [Organizations](#) [Groups](#) [About](#)

[Home](#) / [Datasets](#) / [Create Dataset](#)

What are datasets?

A CKAN Dataset is a collection of data resources (such as files), together with a description and other information, at a fixed URL. Datasets are what users see when searching for data.

1 Create dataset

2 Add data

Title:

* URL: localhost:5000/dataset/my-test-dataset [Edit](#)

Description:

eg. Some useful notes about the data

You can use [Markdown](#) formatting here

Open Data Portal – Mapping Key-Value Pairs

Version:

1.0

Author:

Burt Walsh

Author Email:

burt.walsh@ast.myflorida.com

Maintainer:

Burt Walsh

Maintainer Email:

burt.walsh@ast.myflorida.com

Custom Field:

Key: Major Entities

Value: citizen full name, citizen email

Custom Field:

Key:

Value:

Custom Field:

Key:

Value:

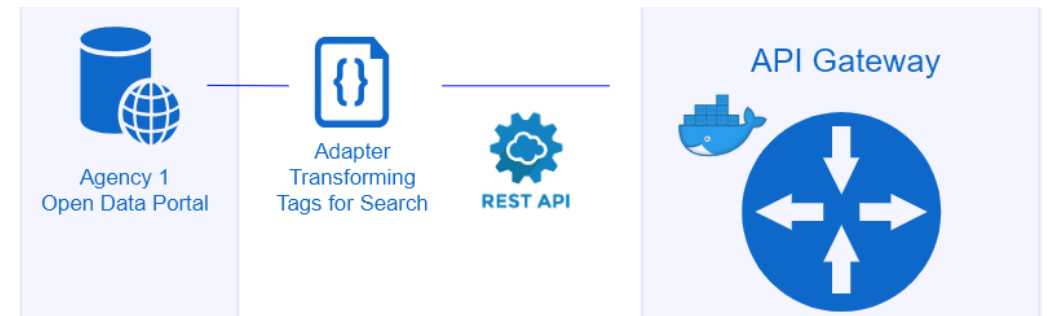
The data license you select above only applies to the contents of any resource files that you add to this dataset. By submitting this form, you agree to release the metadata values that you enter into the form under the [Open Database License](#).

* Required field

Next: Add Data

Metadata is essential for creating key-value pairs for each data set.

Agency-specific open data portals will require an adapter to translate and transform agency key-value pairs for use on the statewide open data portal.



Open Data Portal – Mapping Key-Value Pairs

[Datasets](#)[Organizations](#)[Groups](#)[About](#)

[Home](#) / [Organizations](#) / [AgencyForStateTechnology](#) / **my test dataset**

my test dataset

Followers

0

[Follow](#)

[Organization](#)



[Dataset](#)

[Groups](#)

[Activity Stream](#)

[Manage](#)

my test dataset

[PRIVATE](#)

Data and Resources



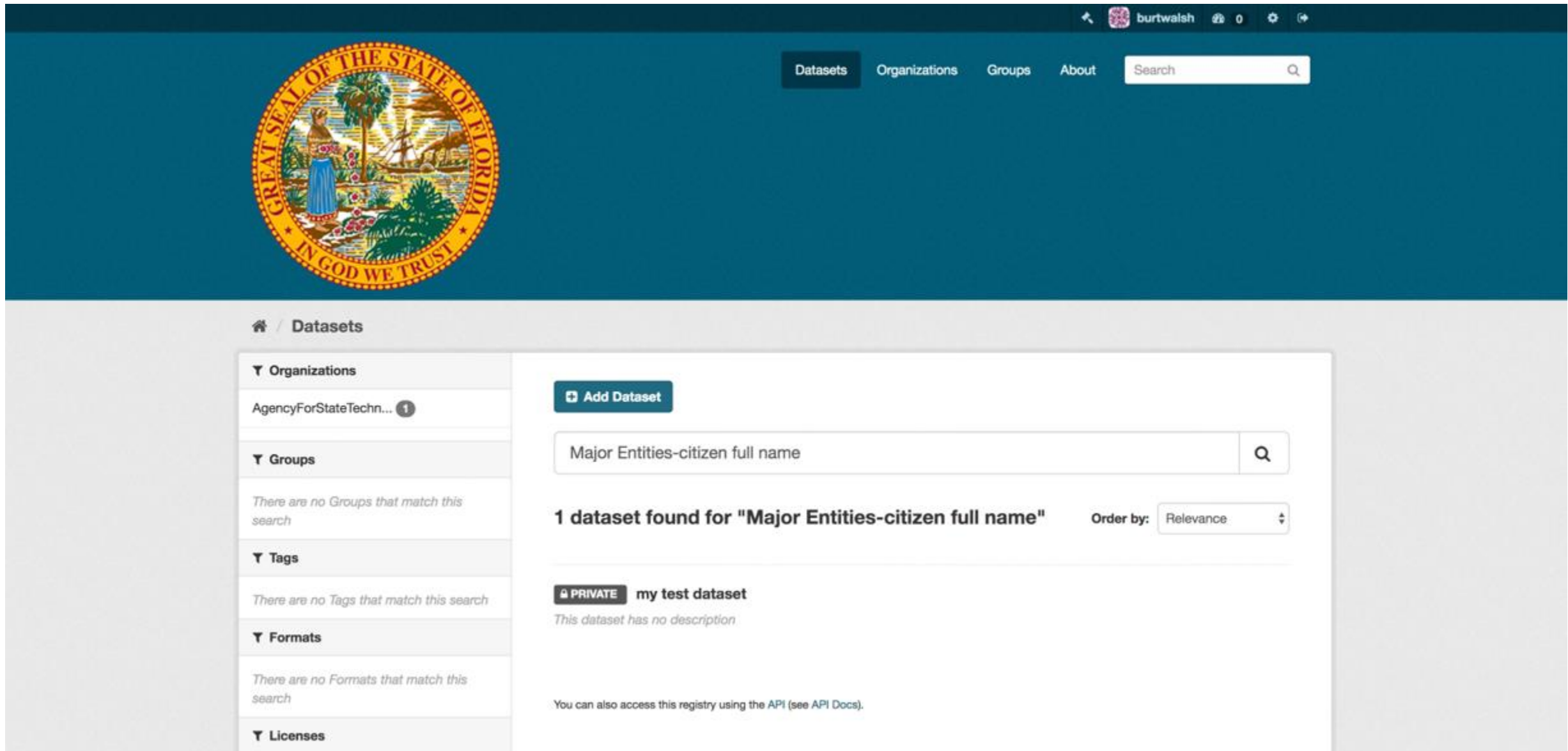
My data

[Explore](#)

Additional Info

| Field | Value |
|------------|------------|
| Author | Burt Walsh |
| Maintainer | Burt Walsh |
| State | active |

Open Data Portal – Search Based on Key-Value Pairs



The screenshot displays the Open Data Portal interface. At the top, there is a dark blue header with the Florida State Seal on the left and navigation links for 'Datasets', 'Organizations', 'Groups', and 'About' on the right. A search bar is also present in the header. Below the header, the main content area is divided into a left sidebar and a main panel. The sidebar contains filters for 'Organizations', 'Groups', 'Tags', 'Formats', and 'Licenses'. The main panel shows a search for 'Major Entities-citizen full name' with one result found: 'my test dataset'. The result is marked as 'PRIVATE' and has no description. A link to the API is provided at the bottom.

Open Data Portal – Search Based on Key-Value Pairs

Header:

- Navigation: Datasets, Organizations, Groups, About
- Search: Search

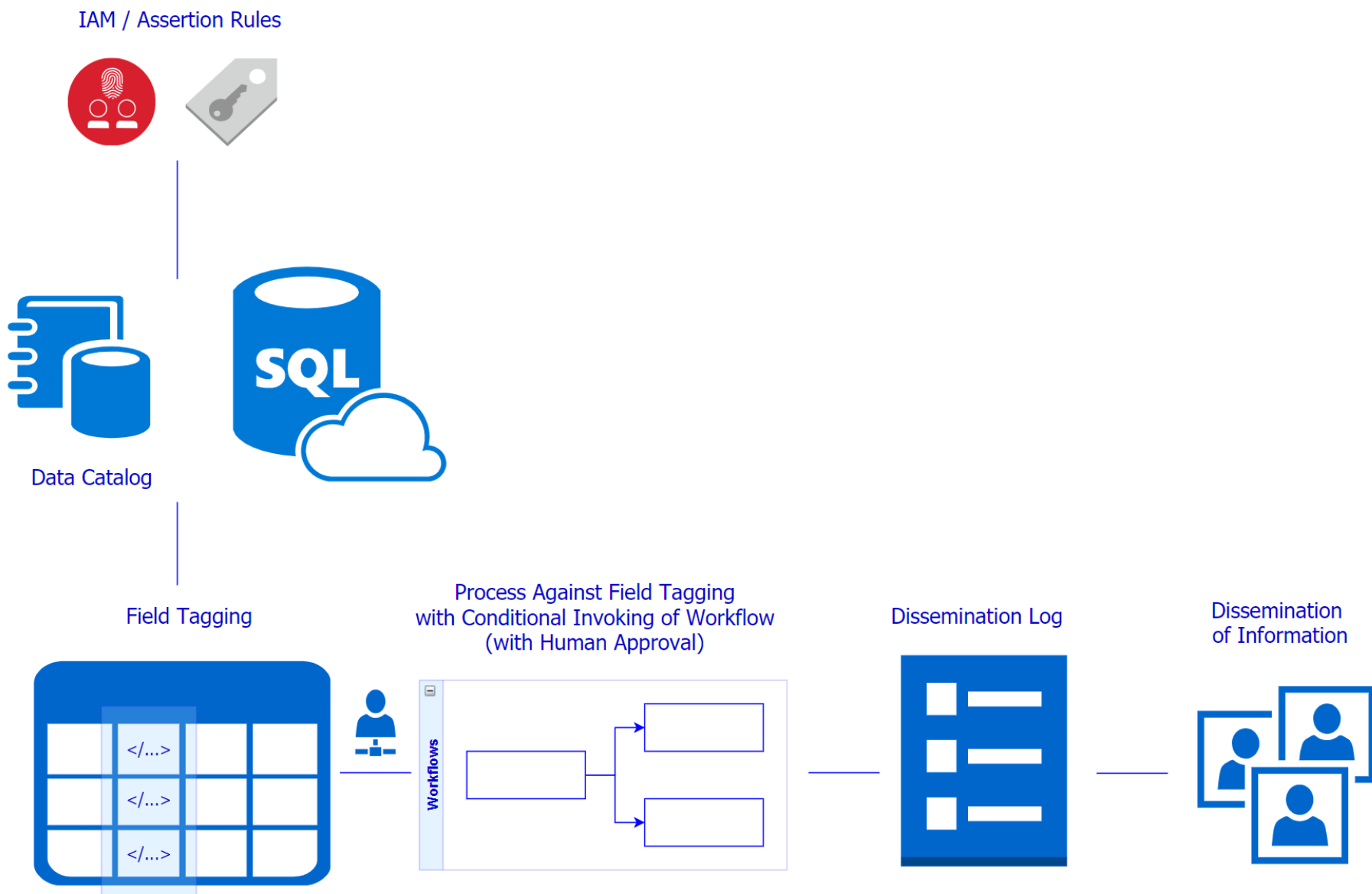
Left Sidebar:

- Organizations**
AgencyForStateTechn... 1
- Groups**
There are no Groups that match this search
- Tags**
There are no Tags that match this search
- Formats**
There are no Formats that match this search
- Licenses**

Main Panel:

- Add Dataset**
- Search: Major Entities-citizen full name
- 1 dataset found for "Major Entities-citizen full name"**
- Order by: Relevance
- PRIVATE my test dataset**
This dataset has no description
- You can also access this registry using the [API](#) (see [API Docs](#)).

Automated Information Dissemination



References

- Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on knowledge and data engineering*, 19(1), 1-16. Retrieved April 27, 2018 from the Purdue University, Department of Computer Science website, <https://www.cs.purdue.edu/homes/ake/pub/TKDE-0240-0605-1.pdf>
- Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., ... & Schwarzenbach, J. (2013). The six primary dimensions for data quality assessment. *DAMA UK Working Group*, 432-435. Retrieved April 23, 2018 from EM360Tech website, https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf
- De Jonge, E., & van der Loo, M. (2013). An introduction to data cleaning with R. Heerlen: Statistics Netherlands. Retrieved April 16, 2018 from The Comprehensive R Archive Network (CRAN) website, https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf
- Florida Department of State. (2012). Address Confidentiality Exemption Request Form Revised 08-2012 (2), Public Records Exemption Request to the Florida Department of State. Retrieved May 2, 2018 from the Florida Department of State website, <http://dos.myflorida.com/media/696331/dos119-public-records-exemption-form.pdf>
- American Society for Quality. (2018). Plan-Do-Check-Act (PDCA) Cycle [Website]. Retrieved September 28, 2018 from <http://asq.org/learn-about-quality/project-planning-tools/overview/pdca-cycle.html>



Questions?

Burt Walsh

burt.walsh@ast.myflorida.com

Michael Avello

michael.avello@ast.myflorida.com

Agency For State Technology
4050 Esplanade Way, Suite 115
Tallahassee, Florida 32399

