AST

AGENCY FOR STATE TECHNOLOGY

AST Data Management Efforts

burt.walsh@ast.myflorida.com

michael.avello@ast.myflorida.com

June 26, 2018

Chief Data Officer (CDO) Office

- Part of the Agency for State Technology (AST)
- Created in the 2017-2018 Session
- Has authority under Proviso
- Charge is Open Data, Interoperability, and Reduction of Duplicate Data
- Hosts and Coordinates a Data Management Workgroup

Open Data

“Open Data … is content can be freely used, modified, and shared by anyone for any purpose”

(https://opendefinition.org/)

Why 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.”

(https://opendatacharter.net/principles/).

Open Data Goals

- Standards for Open Data
- Technical Standards for Open Data
- State Open Data Portal

Open Data Goals

- Many states have open data portals
- The government put a major effort around open data and transparency

  (https://project-open-data.cio.gov/policy-memo/)
- We have many good resources for definitions and efforts

  (https://opendefinition.org/)

The Good News

- The quality of the data is proportional to the effectiveness of efforts leveraging said data.
- Ensuring that data has context and is truly open requires data management.
- Data management efforts have benefits for open data and the government’s use of data.

Data Management Workgroup Goal

Establish Data Management standards and promote their use to ensure the responsible use of the full power of the State of Florida's data so that it can be used to serve the citizens of Florida.

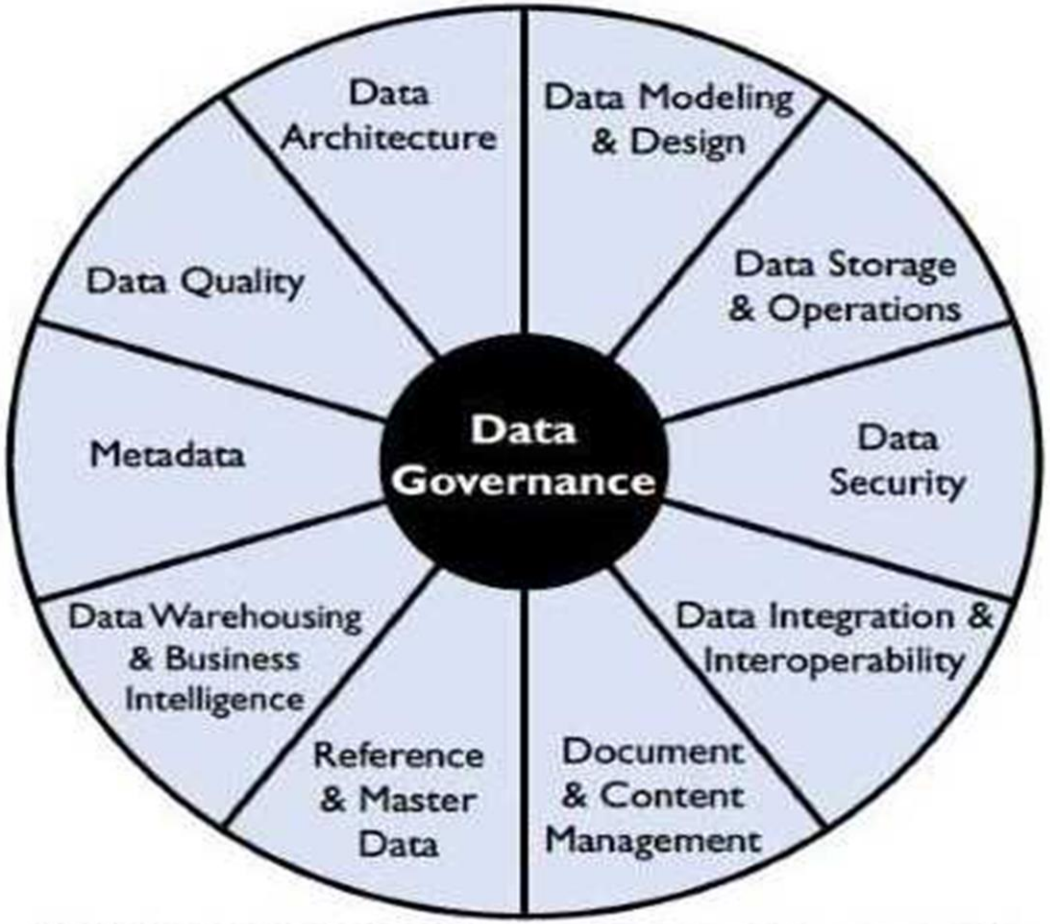
Data Management Workgroup Strategy

- Discuss topics and generic best practices
- Understand the needs of each agency
- Asset current level of maturity
- Develop plans for improvement (starting with KEY datasets)
- Serve agencies in the execution/governance of plans
- Memorialize best practices as standards

Office of the Chief Data Officer

*“Data Governance is a quality control discipline for managing, using, **improving and protecting organizational information**. Effective data governance enhances the **quality, availability, integrity, and protection** of a company's data by fostering **cross-organizational collaboration and structured policy-making**.”*

- IBM


Data Architecture
Data Modeling & Design
Data Storage & Operations
Data Security
Data Integration & Interoperability
Document & Content Management
Reference & Master Data
Data Warehousing & Business Intelligence
Metadata
Data Quality
Data Governance

DAMA-DMBOK2 Data Management Framework

Copyright © 2017 by DAMA International

First Steps

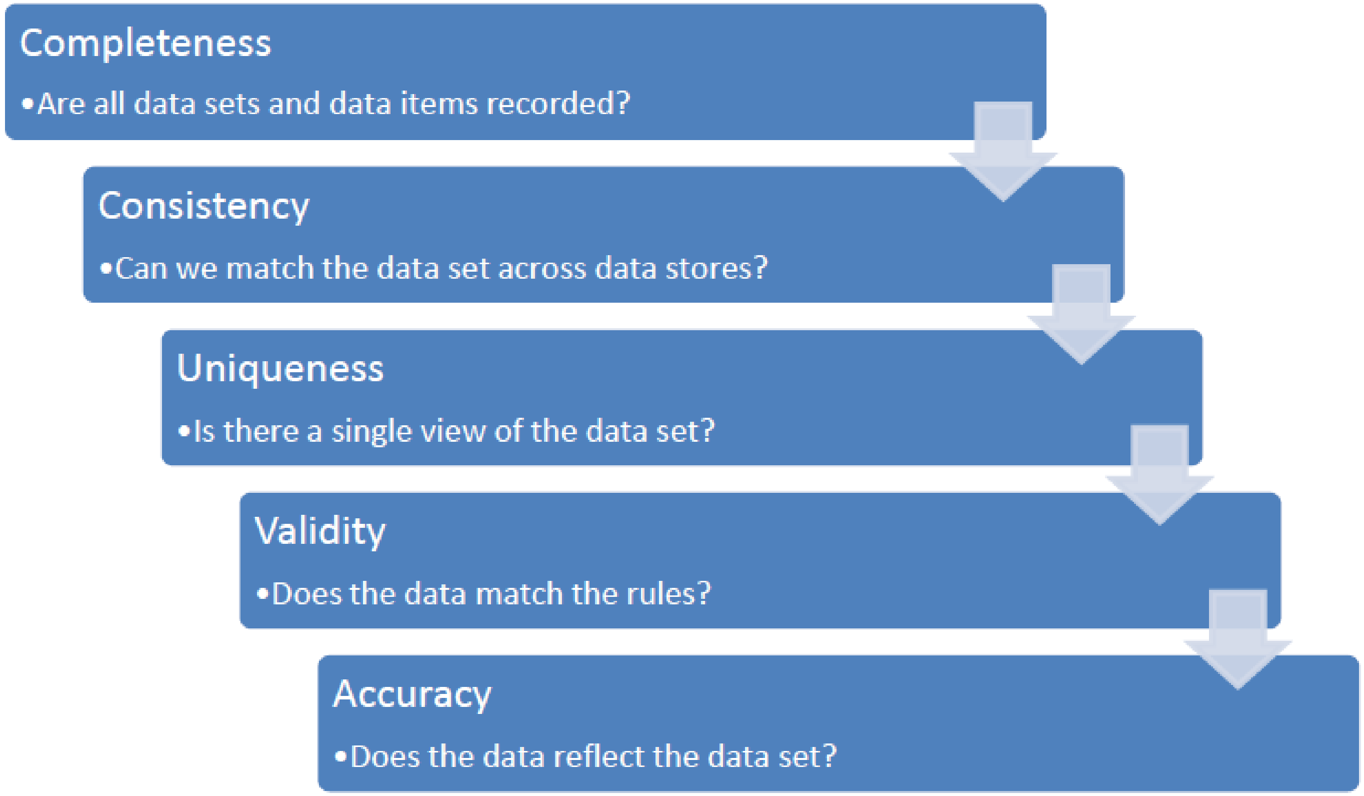
- Determine key data which is used by management to determine Key Performance Indicators (KPI) and Key Risk Indicators (KRI)
- Catalog data quality rules (metadata) around key data
- Catalog key entities and business context around entities in key datasets (metadata)
- Determine entity data attributes (Master Data Management) which can be leveraged for data deduplication and record linkage

Data Quality

Dimensions & Measurements for Data Quality Assessment

Data Quality Dimensions

Describe a feature (characteristic, attribute or facet) of data that can be measured or assessed against defined standards in order to determine the quality of data.



Completeness

- Are all data sets and data items recorded?

Consistency

- Can we match the data set across data stores?

Uniqueness

- Is there a single view of the data set?

Validity

- Does the data match the rules?

Accuracy

- Does the data reflect the data set?

Source: 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.

First Steps (continued)

- Data deduplication/record linkage
- Just enough theory (cluster analysis/measure in this case)
- Tools (Python/R/commercial)
- Standards (HIPPA, FERPA, CJIS…)
- Privacy preserving record linkage (Bloom filters/measure)
- Data Lineage (metadata)

Data Quality Assessment Dimensions & Measurements

(data deduplication with Python)

Techniques & Tools

Soundex functions

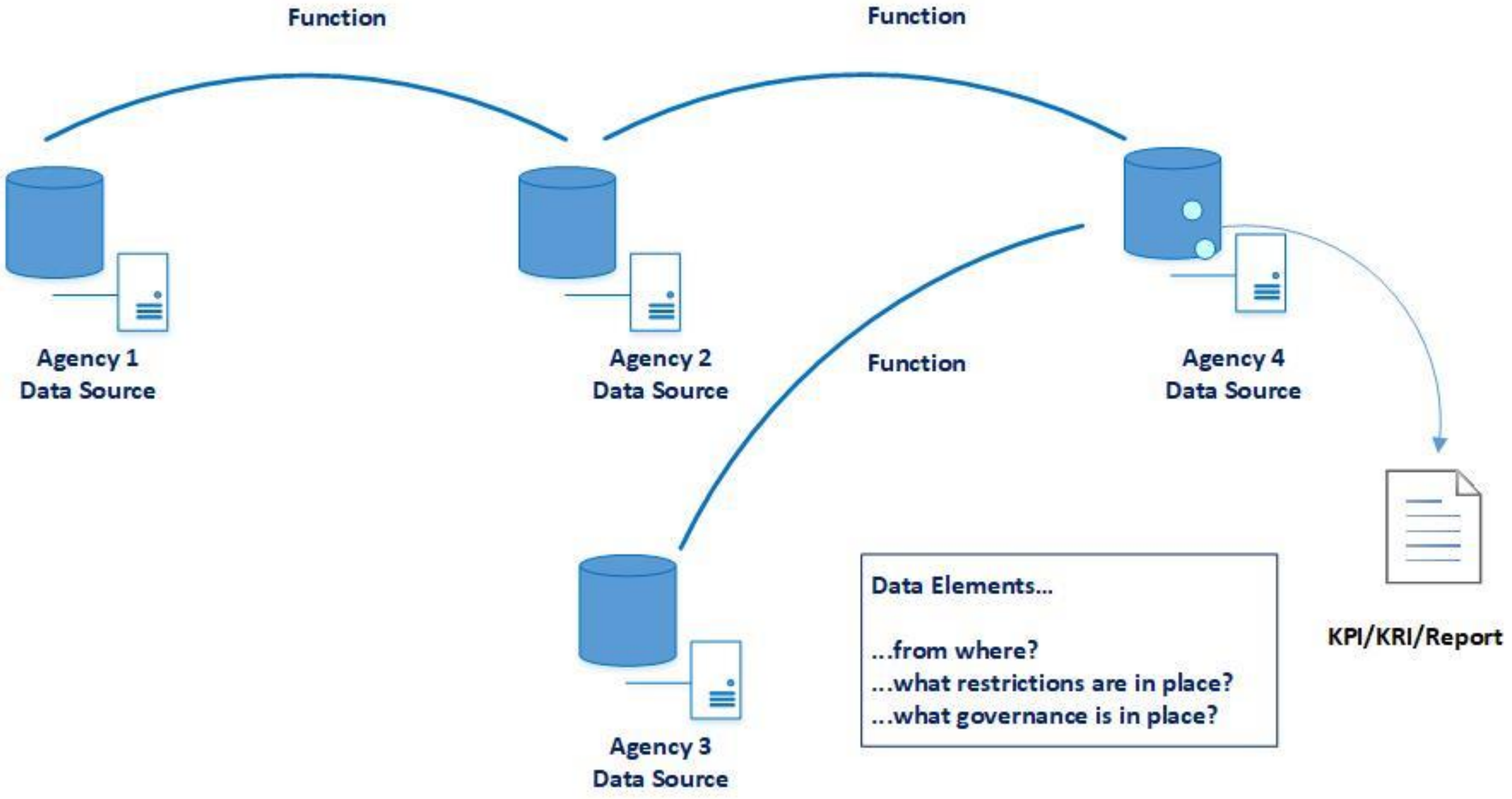
Temporal Data

Business Rules Emphasis

rowID	firstName	lastName	age	street	apt	city	state	zip	ssn
R01	John	Willams	38	643 Gulf Ln		Half Moon Bay	CA	94013	A2B-4C-678D
R02	Richard	Alpert	151	15 Black Rock St		Cannon Beach	OR	97110	A3B-5C-78D1
R03	Ana Lucia	Cortez	39	48 Ocean Park Ave		Santa Monica	CA	90405	A4B-6C-891D
R04	Joh	Williams	42	642 Gulf Lane		Halfmoon Bay	CA	94013	A2B-4C678D
R05	Daniel	Faraday	48	23 Martin St		Essex	MA	01929	A5B-7C-9123
R06	Jonathan	Williams	42*	643 Gulf Lane		Half Moon Bay	CA	94013	A2B4C-678D
R07	Penny	Widmore	43	1623 Hawthorne Road		Palos Verdes	CA	90275	A6B-8C-123D
R08	Austen	Kate	38	1516 Ontario Street		Ames	IA	50014	A78-9B-234D
R09	John	William	47	403 Stadium Dr	B-005	Tallahassee	FL	32304	A2B4C78D
R10	Benjamin	Linus	63	815 Oceanic Ave		Portland	OR	97205	A8B-C2-345D

Putting it all together—with an understanding of the primary business rules—increases the confidence level for finding duplicate records in the database.

Impact Analysis/Business Lineage



Data Elements…

…from where?
…what restrictions are in place?
…what governance is in place?

CDO, GIO, and Proviso

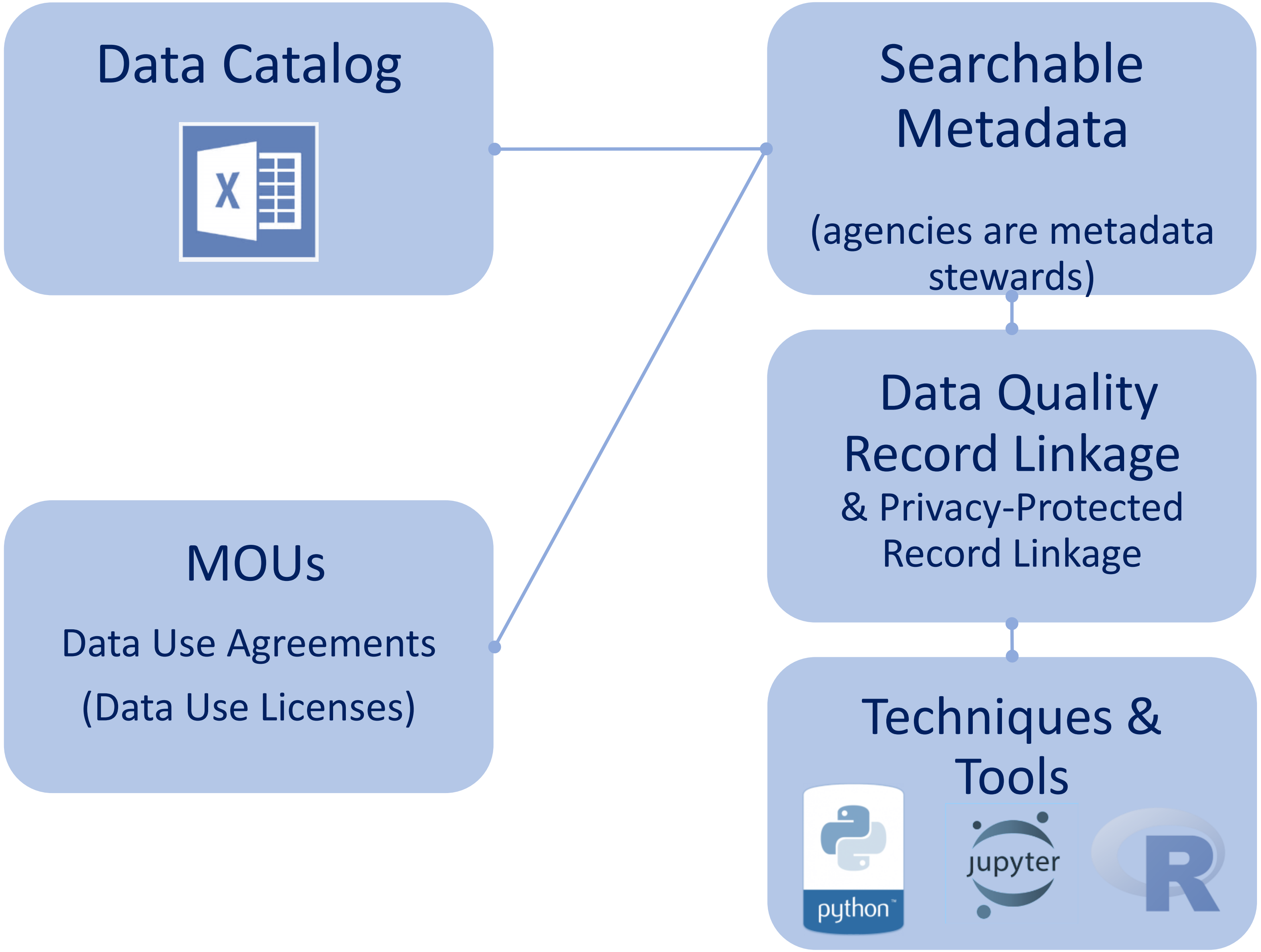
- Chief Data Officer and Geographic Information Officer established in 2017-2018

- For 2018-2019 responsibilities in Pages 381-382 of http://www.flsenate.gov/Session/Bill/2018/5001/BillText/er/PDF

2018-2019 Proviso

- Continuation of some work from 2017-2018
- Enterprise Data Inventory
- Methods for standardizing data to promote interoperability and reduce collection of duplicate data
- Identify data classified as open data
- Recommend open data technical standards
- Recommend options and associated costs for a state open data catalog

Agencies Leveraging State Data



Data Catalog

MOUs

Data Use Agreements

(Data Use Licenses)

Searchable Metadata

(agencies are metadata stewards)

Data Quality Record Linkage

& Privacy-Protected Record Linkage

Techniques & Tools

Enterprise Data Inventory

- Legislature has asked for the following (but not limited to):
  - The title and description of the dataset
  - Description of how the data is maintained including standards
  - Planned Application Programming Interfaces (API) to publish data and the data it exposes

Enterprise Data Inventory (continued)

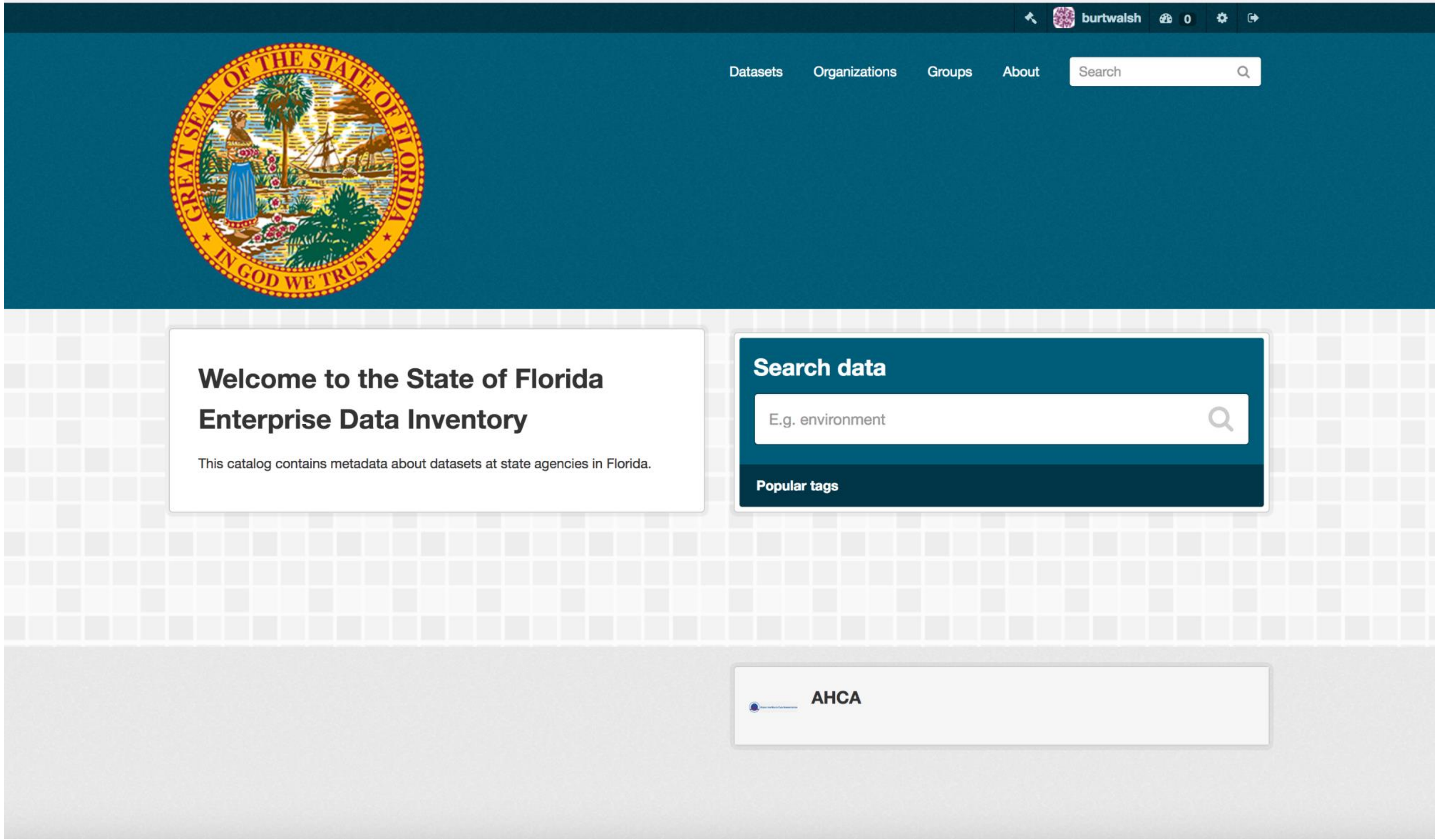
- Other information from 2017 Proviso
  - Is data sharing governed by an MOU
  - Ingress and Egress (interface) technologies and partners
  - Key entities in the dataset (Master Data Management/Metadata)
  - Business function (metadata) context around the data
  - **Support Data Sharing between agencies**

CKAN Software

- Searchable data and metadata repository (http://docs.ckan.org/en/2.8)
- Built from Open Source Software
- Built as part of the Project Open Data (https://project-open-data.cio.gov/)
- Help make data and metadata timely and current

AGENCY FOR STATE TECHNOLOGY

Office of the Chief Data Officer



Welcome to the State of Florida Enterprise Data Inventory

This catalog contains metadata about datasets at state agencies in Florida.

Other Drivers

- Education on technologies, techniques and processes
- Developing Data User **License** Agreements which promote sharing while protecting both the sharing data steward and the citizen’s data
- Metadata to help with understanding of data (requirement of the sharing party)
- Serve the Citizen’s of Florida while protecting their data!

Unlocking Access to Data

Data Use License

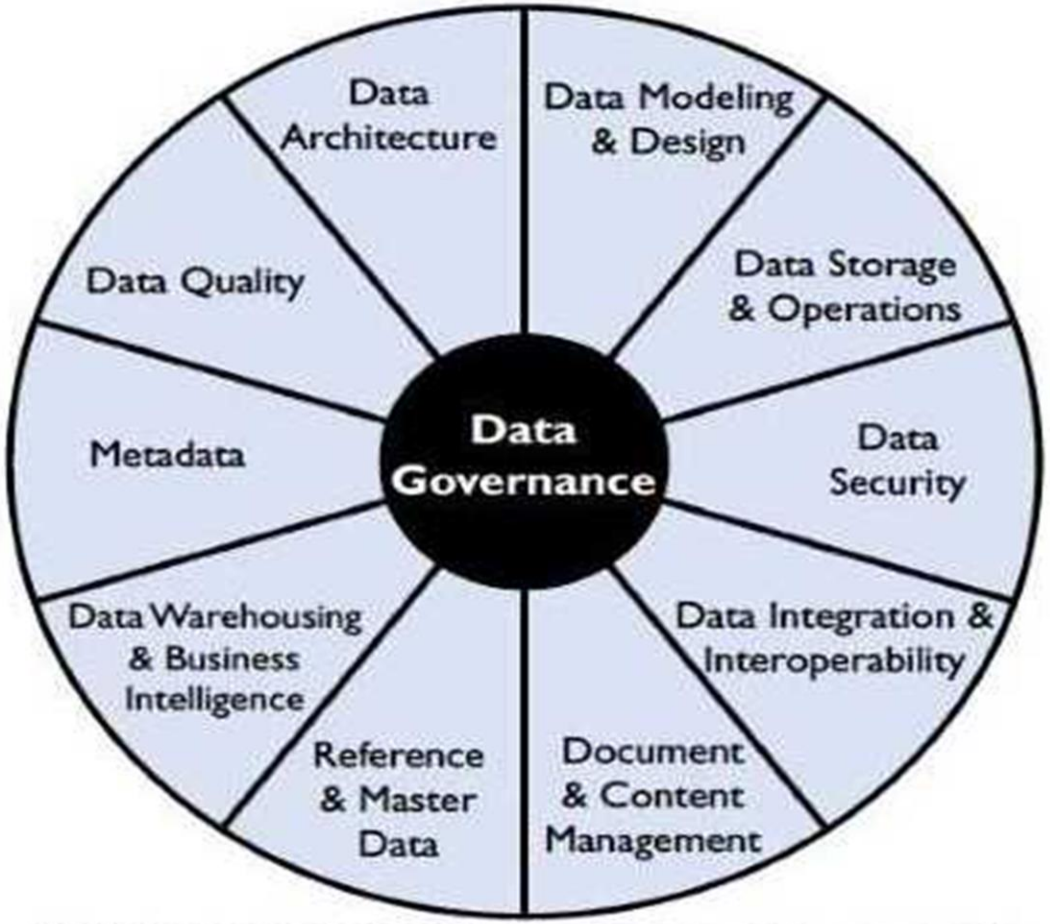
- What is required to have access to the data?
  - Credentials (training)
  - Purpose
  - Agreement to terms
  - Acceptance of responsibility of use
  - Indemnification for the provider

Data Use License

- What can and should be expected from the provider?
  - Metadata
    - Data quality measures
    - Business context for the data
    - Timeliness of data
    - Standards including compliance
    - Lineage (source and process for creation)

Office of the Chief Data Officer

"How do we get a holistic approach to data management and data sharing?"


Data Governance
Data Architecture
Data Modeling & Design
Data Storage & Operations
Data Security
Data Integration & Interoperability
Document & Content Management
Reference & Master Data
Data Warehousing & Business Intelligence
Metadata
Data Quality

DAMA-DMBOK2 Data Management Framework

Holistic Data Management

- Coordination between application owners, security professionals, data professions, legal professionals and business professionals
- AST supporting agencies in their service to the citizens of Florida
- Help agencies with solutions and/or technologies

Supporting the Agencies

- We are here to support the agencies in their data needs so that they can support the citizens of Florida.
- We can help with data management, data science and give guidance on technologies.
- Please feel free to contact us at any time!

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

AGENCY FOR STATE TECHNOLOGY

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