Enabling Data Discovery & Analytics Through Curation in a Federated environment

Presenter: Kathy Rondon

Why a Federated system?

- Authority is "divided between a central...authority and smaller, locally autonomous units...usually formed through the political union of two or more formerly independent states under one sovereign government that does not, in any case, arrogate the individual powers of those states."

- A federated system allows both autonomy is some respects and centralization in others.

- Large organizations generally benefit from such a federated governance system.

What to Centralize?

- The data assets themselves almost never need to be centralized, stored or accessed in a single repository
- Centralized tasks or content: reference data, master data, and/or metadata
- The Open Archival Information System is one example of a federated architecture

Leveraging the OAIS

- Borrowing from OAIS tenets to create a federated governance environment:
  - Negotiate with and accept data from data providers
  - Obtain sufficient control of data assets to ensure long-term preservation
  - Ensure that data is understandable to the user community without having to consult with data creators or providers
  - ***Follow documented policies and procedures which enable information to be disseminated as authenticated copies and/or traceable to the original***

Standardizing Contextual Metadata

- The Submission Information Package (SIP): Information that providers send to the central repository
- The Dissemination information Package (DIP): Information the central repository sends to data requesters
- The Archival Information package (AIP): Information stored by the central repository for preservation and archival purposes

What Is Data Curation?

- A discipline that enables data discovery and retrieval, maintaining its quality, adding value, and providing for reuse over time.*

- The ongoing documentation of **CONTEXTUAL METADATA** about key data assets according to standard operating procedures and consistent terms of reference

*Source: The University of Illinois School of Library and Information Sciences

CURATION – IT'S ABOUT THE METADATA

CURATION [illegible] and/or articulation of caveats and margins of error

How to Curate Your Data

- Develop and maintain **terms of reference** for describing your organization's data.
- Document and communicate to users **data handling policies** (such as restrictions on use, copyrights and licenses).
- Determine the **metadata** (structured and unstructured) that your organization's data users need—and think ahead to future reusability of the data.
- Train your workforce on consistent **documentation standards**.
- Data curation is essentially a data management maturity issue.

Constructing a Data Dictionary

- Best Practices
  - **Try not to boil the ocean**: Choose the most important and possibly most misunderstood terms and functions and define them in as straightforward and simple language possible
  - **Coordinate**: ensure that stakeholders agree on the terms of reference, but don't let a single hold out take your data dictionary hostage
  - **Communicate**: Once you have a data dictionary, don't stick it in a drawer or keep it a secret; publish it, share it and train it in as many forums as possible

Determining Metadata Standards

- To consistently curate data, the organization must have a metadata standard, the contextual metadata fields that apply to every curated dataset

- Contextual metadata can be structured, unstructured, or both, depending on the needs of the organization

- If a relevant metadata standard already exists, don't reinvent the wheel, but don't be afraid to create something new if you need to

Determining Metadata Standards

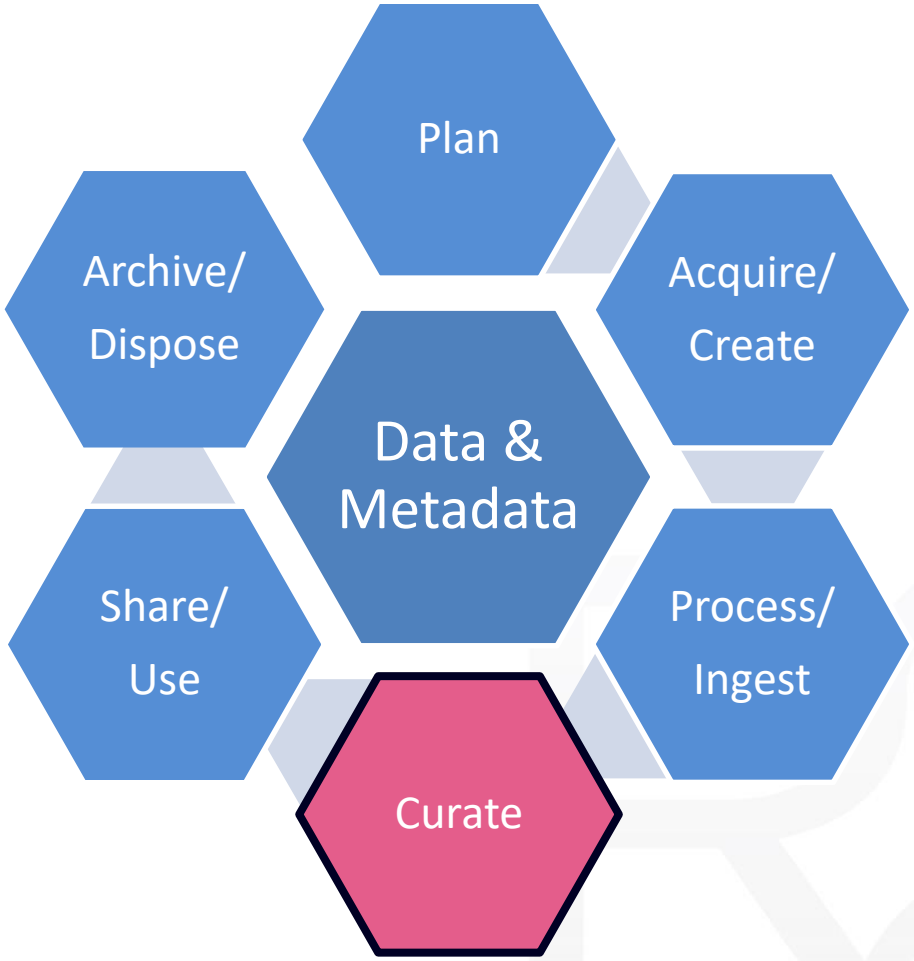
"Standards are like toothbrushes. Everyone agrees that they're a good idea, but nobody wants to use anyone else's." *

* From *Metadata* by Jeffrey Pomerantz; attributed to Murtha Baca, Getty Research Institute

All will be for naught if…

- Data Curation: everyone thinks it's a great idea…for someone else to do.
- Training and professionalization of a curation workforce is key to the success of a curation effort
- Skill set of a data curator is one that bridges the gap between IT and business or mission.
  - A curator should understand data formats and data exploitation systems at a basic level.
  - The curator must have excellent writing and communication skills
  - The curator must in depth understanding of the intended Business or mission uses of the data.

Curation in the Data Lifecycle



QUESTIONS?