Defining 'Core' Metadata: what is needed to make data discoverable?

San Cannon

Federal Reserve Board

FedCASIC 2013

The challenge

- Massive influx of data from a variety of sources to a variety of business units spurred by assorted business needs
- Coordination of acquisition is improving: users need to find what we have before they try to acquire more
- Data sets are mostly catalogued – in many different places
- Work group of staff at the Federal Reserve Board and New York Fed began to ponder

The solution – with questions

- How to help with data discovery? Single catalogue? No! Build a federated search tool!
- Harvest metadata from a variety of catalogue instances of various formats.
- Each catalogue is built to serve a particular user base.
- Wide range of metadata concepts and names
- What is really required to make unified search work?

Defining (required) metadata

- Think of your iPod – how do you find music there?
- Title, artist, album, genre, year, etc. are all metadata fields.
- Some are more useful for finding music, others are used for other purposes.
- What do you consider “required?” Why?

The cost/benefit analysis

- Buying music online gets you lots of information with it but digitize your own music and information needs to be entered
- What is the minimum information you find useful for your purpose?
- It depends on the use – discovery, purpose, operational, usage
- iPod shuffle requires/displays no metadata yet is a useful tool
- But how do you create a playlist?

Same is true for datasets

- Providing metadata is costly but necessary for discovery and determining fitness for purpose as well as knowing how to access and use the data
- Require too many items and people won't comply, require too few and the information won't be useful
- Proposal: make minimal metadata required and encourage inclusion of other items
- Process: identify "core" information for tool to search as mapped from source catalogues

Our solution

- Examine existing catalogue entries and uses
- Look to “standards” or other tools
- Determine minimum requirements
- Categorize other entries by degree of “usefulness”
- Result: 27 metadata items of 4 types divided into 5 requirement categories

Six (6) Mandatory fields

- Dataset name - Official/formal name of the dataset.
- Dataset description - Relatively detailed description of the contents of the dataset as free from marketing jargon as is practicable.
- Dataset creator - Name of the individual, group, or entity who can claim intellectual property over the creation of the dataset if not its individual items.
- FRS Security classification - Classification under FRS information security policies.

More mandatory information

- System contact - Person or group who serves as the main contact to be able to answer or appropriately route security, access, format and content questions.
- Organizational unit responsible - Name of the group or organizational unit within the System to which the creator (or acquirer) belongs
- Three (3) “conditional mandatory” fields
  - Other FRS classification
  - Dataset Access Policies
  - Dataset Usage Policies

Next level

- Six (6) strongly suggested fields
  - Dataset Also Known As
  - Key Search Words
  - Subject Area
  - Dataset provider
  - Unit of analysis
  - Date Range Accessible

Other useful information

- Eight (8) optional fields
  - Data Frequency
  - Dataset Update Frequency
  - Date Range Existing
  - Dataset Storage Format
  - Dataset Storage Location
  - Notes
  - Documentation
  - Geographic Region (under review for more general use)

System information

- Four (4) “harvested” fields
  - Dataset Type (created, contracted, collected)
  - Dataset Creation Date
  - Inventory Load Date
  - Inventory Update Date

Challenges

- Three dataset types have very different requirements
- Not all catalogues contain even the basic 6 mandatory fields
- Not all data are catalogued
- Need for searching on variable level data
- Security concerns and access restrictions even for metadata

Current status

- Proof of concept implementation being currently tested with available technology and a subset of catalogues
- Outlining limitations of current application due to limited development time with proposal for further development
- Creating guidance for catalogue creators/maintainers to for better findability

Thanks for listening!

Questions?

Sandra.A.Cannon@frb.gov

(202) 452-3710