Wendy Thomas
Minnesota Population Center
NADDI 2014

MANAGING RESEARCH DATA WITH DDI-L: SUPPORTING INTEROPERABILITY BETWEEN MULTIPLE SYSTEMS

Coverage

- Problem statement
  - Why are there problems with interoperability with external search, storage and delivery systems
- Minnesota Population Center situation
  - Legacy model, increased requirements for interconnectedness, and internal needs
- Approach and Progress

Problem Statement

- System differences
  - Convergence of 3 primary systems for managing information
  - Content coverage, organization, and entry point
- Differences in content standards
  - Can have a different primary focus and purpose
  - Content coverage, organization, and entry point
  - Depth of searchable content
- Combining contents with systems
  - Ingest expectations
  - Delivery expectations

A little historical background

Systems Perspective

Library Perspective

- Libraries are collections of individual objects selected and organized by topical content
  - Descriptions (metadata) are traditionally held external to the object and are designed to support discovery via title, author, topical, temporal, and geographic coverage
  - Collections are fluid (libraries access and deaccess objects)
  - When objects became electronic with searchable content, descriptions were linked to OR bundled with the object to allow “keyword” searching of the object itself
  - Descriptions are “high level” and “generic” (i.e. they describe the object overall and support description of a wide range of object types)

Archives Perspective

- In general, archives consist of records that have been selected for permanent or long-term preservation on grounds of their enduring cultural, historical, or evidentiary value. Archival records are normally unpublished and almost always unique, unlike books or magazines for which many identical copies exist.

http://en.wikipedia.org/wiki/Archive

Archives cont.

- Archive metadata
  - Normally separate from the object/record itself
  - Focuses on relationships between records particularly in terms of organizational source, time, and the processes that created them (provenance)
  - Preservation is a key provision (archives ingest and preserve)
  - Queries often focus on relationships within the collection rather than on a “piece of information”; descriptive records support this via the use of fond, series, file, and item descriptions

Information Technology Perspective

- Focus on storing, retrieving, manipulating and communicating information
  - Storage is electronic (an object and/or description can be stored)
  - Retrieval is based on unique addresses discovered by searching:
    - Structured indexed content
    - All electronic content
    - Following chains of relationships (explicit or virtual)
  - Optimization occurs around speed of delivery and accuracy of the delivered content

Implications

- Each external system we interact with comes out the perspective of a different primary system, prioritizing some aspects over others
- Each has integrated other perspectives into their system approach to varying degrees

Content differences: There’s metadata and then there’s METADATA

- metadata
  - Bibliographic+ metadata is the high level discovery objects common to a broad range of objects. Think Dublin Core, OAI-ORE, MARC, etc.
- METADATA
  - Content metadata varies by discipline or group of disciplines. It carries the detailed information required to accurately determine the fit of data for a specific use and how to access datum within a data object

Bibliographic+ metadata

- Carries standard title, author, publisher, identifier, distributor information
- Provides structured coverage information (temporal, topical, spatial)
- May provide unstructured topical searching by leveraging access to content metadata through keyword searching of some or all text content
- Bibliographic metadata is associated with an object or aggregation of objects

Examples of bibliographic+ metadata

- Dublin Core – the basics
- MARC, DMARC, other bibliographic record standards
- METS – a means of wrapping a common structure of bibliographic metadata with the content metadata and objects (Digital Library Federation)
- OAI-ORE – a structure that adds the archival perspective of aggregations and flexible resource mapping (OAIS)

METADATA

- Content metadata is designed for specific purposes including but not limited to
  - Supporting deep topical discovery
  - Describing how to access a single datum within the object
  - Determining fitness of data to a specific use
  - Informing users of quality and facilitating use
  - Capturing process and provenance information
  - Driving production
  - Supporting comparison, analysis, and repurposing
  - …and more

Examples of content metadata

- EML – Ecological Metadata Language
  - Resource module containing information describing dataset, literature, protocol, and software resources
- FGDC – Federal Geographic Data Committee
  - Information on identification (bibliographic), data quality, organization of data, spatial reference, entity and attributes, distribution, and metadata reference
- DDI – Data Documentation Initiative
  - Study, conceptual framework, data collection/ capture, methodology, data processing, logical content of the data, physical storage, summary statistics, archival management, lifecycle events, comparison, groups, reusable metadata, source data, collections of data, etc.

Common features

- Provides high-level metadata with detailed, coverage relevant metadata
- Binds metadata and data within the metadata or through explicit external links
- Perspective is generally data file centric
- Common stated purpose is to support discovery and access

Combining the content with systems

- Ingest expectations:
  - There is an assumption that because we all cover the basic metadata that it is organized in similar ways
  - That metadata has related data
  - That the focus of the metadata is the data file/set
- Delivery expectations
  - All over the board

Comparison of purposes

DDI-L

- The **Data Documentation Initiative (DDI)** is an effort to create an international standard for describing data from the social, behavioral, and economic sciences. Expressed in XML, the DDI metadata specification now supports the entire research data life cycle. DDI metadata accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving.

http://www.ddialliance.org/what

FGDC

- The standard was developed from the perspective of defining the information required by a prospective user to determine the availability of a set of geospatial data; to determine the fitness and the set of geospatial data for intended use; to determine the means of accessing the set of geospatial data; and to successfully transfer the set of geospatial data.

Federal Geographic Data Committee. FGDC-STD-001-1998.
Content standard for digital geospatial metadata (revised June 1998).
Federal Geographic Data Committee. Washington, D.C. (pg. iv)

Comparison of purposes

DDI-L

- The **Data Documentation Initiative (DDI)** is an effort to create an international standard for describing data from the social, behavioral, and economic sciences. Expressed in XML, the DDI metadata specification now supports the entire research data life cycle. DDI metadata accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving.

http://www.ddialliance.org/what

FGDC

- The standard was developed from the perspective of defining the information required by a prospective user to determine the availability of a set of geospatial data; to determine the fitness and the set of geospatial data for intended use; to determine the means of accessing the set of geospatial data; and to successfully transfer the set of geospatial data.

Federal Geographic Data Committee. FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee. Washington, D.C. (pg. iv)

Comparison of purposes

DDI-L

- The **Data Documentation Initiative (DDI)** is an effort to create an international standard for describing data from the social, behavioral, and economic sciences. Expressed in XML, the DDI metadata specification now supports the entire research data life cycle. DDI metadata accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving.

http://www.ddialliance.org/what

FGDC

- The standard was developed from the perspective of defining the information required by a prospective user to determine the availability of a set of geospatial data; to determine the fitness and the set of geospatial data for intended use; to determine the means of accessing the set of geospatial data; and to successfully transfer the set of geospatial data.

Federal Geographic Data Committee. FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee. Washington, D.C. (pg. iv)

DDI-Lifecycle

- Pushed the focus from a data file firmly onto the Study defining the StudyUnit as a coordinated data capture process
  - A one time data capture through one or more instruments
  - A single wave or data capture cycle of a repeated study
- Allowed Grouping of Study Units into series or other relationships

- DDI-L does not come SOLELY from a discovery perspective
- Its no longer “data file” focused

So……

- When we interact with external systems that use a Library/IT discovery/access based approach its difficult to know what the primary access point is

Resulting issues with various systems

- METS
  - What is the primary entry point?
- Da|ra
  - If the “data file” is the primary object what about derivatives?
  - What about multiple forms of primary content metadata?
- DataONE
  - Where do we store the relational information for OAI-ORE (Resource Maps, Aggregation, etc.)
  - How can we support scrapping multi-relational descriptive metadata out of DDI content?

Minnesota Population Center (MPC)

MPC Metadata Systems

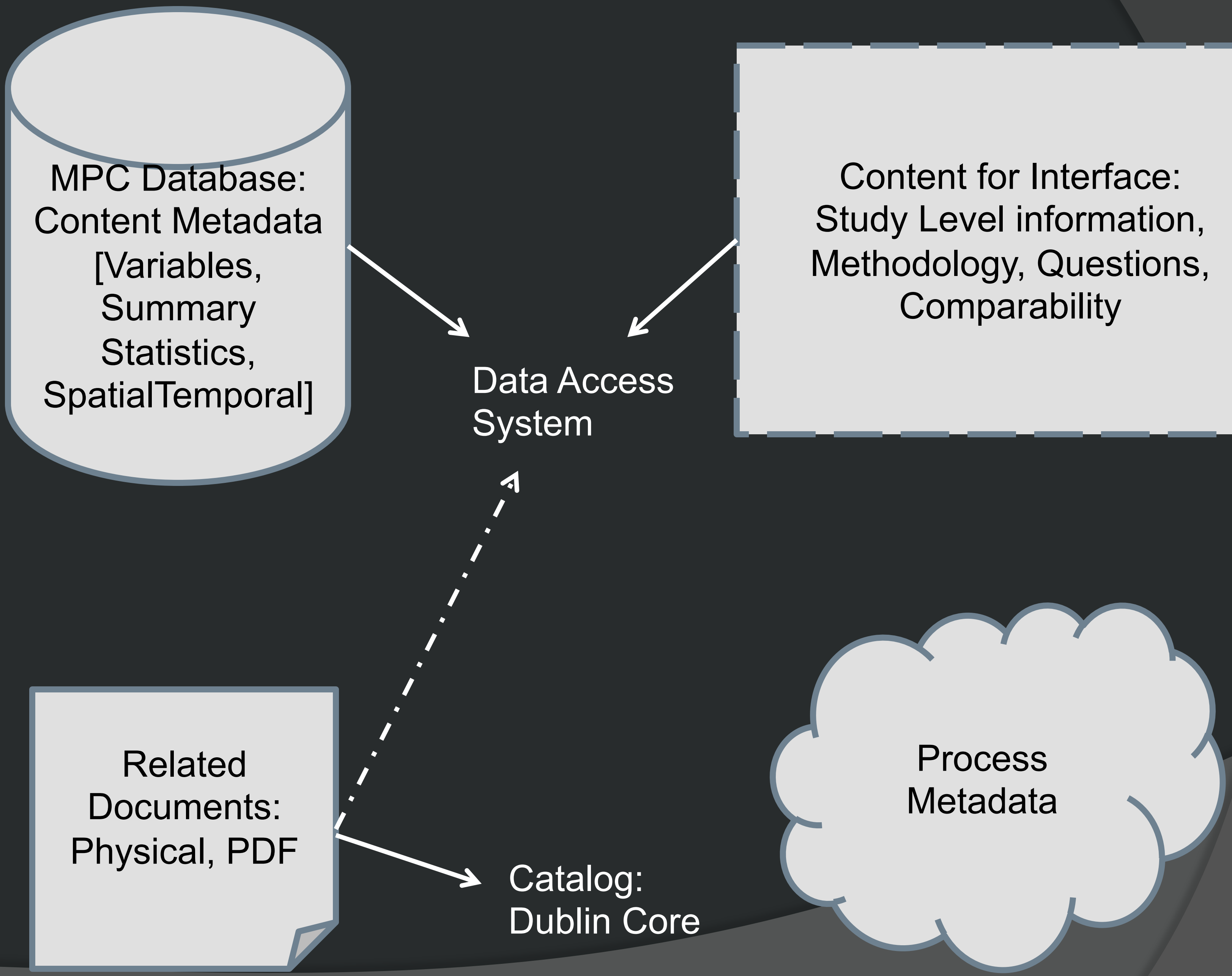
- Microdata storage and access system (IPUMS and related systems)
- Aggregate data storage and access system (NHGIS)
- Integration of access systems (TerraPop)
- Specialized access systems for some microdata projects (IHS, ATUS, ...)

The MPC as a hybrid institution

- Are we a research center?
  - Modify (integrate and harmonize) rather than collect data
  - Provide the data infrastructure for other people’s research
- Are we an archive?
  - Archival responsibility for our products
  - Archival responsibility for selected source data
- Are we a service center?
  - Provide support for proposal development and implementation
  - Forum for discussion

Current Data Metadata Structure

- Data is held in ASCII fixed format files
- Metadata is held in multiple formats
  - Standardized MPC Data Base (microdata and aggregate data)
    - Runs the dissemination and access system
  - Structured text documents
    - Study level information used in user interface
  - Physical and digital images of related materials and original metadata
  - Provenance and Process notes…varied



MPC Database: Content Metadata [Variables, Summary Statistics, SpatialTemporal]

Content for Interface: Study Level information, Methodology, Questions, Comparability

Data Access System

Related Documents: Physical, PDF

Catalog: Dublin Core

Process Metadata

Current level of standards compliance

- Dublin Core
  - Use an extended version of Dublin Core Terms to describe related documents and data files
- DDI-Codebook
  - Original input structure for aggregate data systems
  - Output structure for microdata products

(Metadata databases could be mapped to DDI Lifecycle presumably without loss)

Model Selection

- Currently going with an integrated model using Premis, DDI, ISO 19115, and Dublin Core
- Working on developing a profile of objects from each that will be supported within the MPC (required/optional) and how they relate to each other
- Determine mapping to external metadata structures we need to interact with

The Issues

- Identification of gaps in metadata and determining how to fill them
- Involve individuals in resolving metadata capture issues on a process-by-process basis
- Minimize time requirements on research staff for analysis activities and process changes
- Relaying a sense of the larger picture – why metadata is captured and how it is used – without overwhelming individuals
- Develop a means of instituting these practices early in the project proposal stage for future projects

Specific Requirements

- Producing specific “flavors” of DDI to meet needs of DDI based systems (World Bank, other NADA systems)
- Generating and storing different required subject headings
- Organize profiles of DDI 3.2 to serve different functions
  - Publication
  - Internal management of specified content

MPC Approach and Progress

Initial decisions

- Continue to maintain internal systems
  - Move more content to database
- Define current system as the delivery system and explore what is needed for a processing/archival layer(s)
- Publish DDI 3.2 for archiving and dissemination purposes
- Publish other dissemination formats from DDI 3.2 (leveraging DDI 3.2 to X mapping activities)
- Use DDI 3.2 (4) to inform the content and structure of processing/archival layer(s)

Additional recommendations

- Clearly differentiate harmonized content from sample specific content
- Add a collection management layer to:
  - Capture cross collection relationships
  - Facilitate interface with external system
  - Integrate non-DDI related objects (50,000 documents related to census activities from around the world)
- Generate publication profiles and processes to meet external needs

Sharing perspective

- Our original approach was based on how we wanted to manage metadata internally
- Viewed DDI-L as a base output from which high level records or DDI-C could be created for external distribution
- We currently are working with 5 different organizations who want to provide access to our collections
- Everyone wants something different

External catalog

- IHSN has a specific format of NESSTAR's DDI-C for individual samples
- Da|ra wants a fuller DDI bibliographic record based on the study
- DataONE wants an OAI-ORE resource map based on the data file
- All have their locally supported search subjects

What I want

- To make sure all the metadata regarding our data files can be expressed in a DDI 3.2 instance
- Leverage the more detailed bibliographic information structure of DDI
- Maintain an set of bibliographic information (extended Dublin Core) to serve a source for generating records based on external profile requirements that covers all of our holdings (DDI and non-DDI)

Collection management

- Create extended Dublin Core records for non-DDI material
- Create collection level records that can serve as OAI-ORE Aggregations
- Automatically generate the subject headings for external systems based on our internal subject headings
- Capture all relationships between records in a way that supports a variety of objects being considered “top level” objects

Dublin Core Extensions

- Add MPC type codes that allow for selection of specific elements when creating a profile of metadata for a specific external system
- Addition of more specified OWL and OAI-ORE predicates for linking
- Addition of specialized links between a data file and it's primary metadata
- Content to support the consistent generation of RDF URN identifiers

DDI content

- Study level metadata
  - Bibliographic, spatial, concepts, coverage
  - Related data files (Physical Instance)
  - Instruments (Questionnaires)
  - Other Materials (bibliographic information)
    - Codes
    - Spatial metadata
- Group level metadata
  - Bibliographic, spatial, concepts, coverage
- Resource Packages
  - Bibliographic, coverage

I need to be able to "scrape" the following information from the DDI:

- Record for each object within a DDI Study Unit and Resource Package
- Record for each "collection"
- Links between records to support flexible aggregations
- Generate specialized subject headings from local subject content

Return metadata to DDI

- When objects are deposited in da|ra a DOI is generated and needs to be noted in the DDI
- When objects are deposited in DataONE an identifier is generated and needs to be noted in the DDI
- When a DDI instance (DDI-L or DDI-C) is generated the object is stored and the specific DDI identification (Agency, ID, Version) needs to be noted in the DDI store as a product

DDI ISSUES:
Supporting Library/Archive Management

Possible areas of enhancement

- Making the internal use of Dublin Core extensible in terms of adding DDI and/or Local type attributes
- Capturing more specific relational information (OAIS Resource Maps, DataONE link to specific metadata for a data file)
- Improved access control
- Provenance management

Questions

wlt@umn.edu