Metadata for Research Data: An Overview

Jian Qin
School of Information Studies
Syracuse University
Syracuse, NY, USA
The context

• Data-driven X
  – Science
  – Policymaking
  – Healthcare
  – And more...

Demand for organized, curated, discoverable, reusable data
Metadata for research data

“all the information, additional to the raw data itself, which a potential user of the data would need to know to be able to make full and accurate use of the data in a subsequent scientific analysis...”

(Sufi & Mathews, 2004, p. 4).
The goal

“The real challenge is not to store countless petabytes (1 million gigabytes) of information, but to selectively retrieve and analyze that information in real time” (French, 2012, p. 165).

An important goal of metadata for research data is to provide a complete description of datasets and services for discovery, preservation, and use purposes.
Metadata requirements

- Automatic or semi-automatic metadata generation
- Interoperable
- Secure

- Verifiable
- Replicable
- Reproducible

- Interworkable
- Analysis-ready
- Visualizable

- Findable
- Identifiable
- Selectable
- Obtainable
The alphabetic soup

Astronomy data

Biodiversity and ecology data

Climate and forecast data

Clinical trials data

Access to Biological Collections Data

Darwin Core

Ecological Metadata Language (EML)

NetCDF Climate and Forecast (CF)
Metadata Convention

ClinicalTrials.gov
Protocol Registration System
The alphabetic soup (cont’d)

DNA and genome sequence and protein data
- Minimum Information about any (x) Sequence (MixS)
- Minimum Information about a (Meta) Genome Sequence (MIGS/MIMS)
- Minimum Information about a MARKeR gene Sequence (MIMARKS)

Geospatial data
- Content Standard for Digital Geospatial Metadata (CSDGM) 1998
- ISO 19115:2003 Geographic Information – Metadata
- North American Profile (NAP) of ISO 19115: Geographic information – Metadata

Social, behavioral, and economic sciences data
- Data Documentation Initiative – Lifecycle (DDI-L)
- Data Documentation Initiative – Codebook (DDI-C)

Provenance data
- Open Provenance Model (OPM v1.1)
Architectural view of standards

Identity metadata
Agent / Investigation / Publication / Data set or collection

Semantic metadata
Taxonomy / Thesauri / Classification / Ontology / Relations

Scientific context
Workflow / Provenance / Parameter / Processing

Geospatial metadata

Temporal metadata

Miscellany
Relationships between standards

Biological sciences
- Biological Data Profile
- Shoreline Metadata Profile
- Darwin Core (DwC)
- Ecological Metadata Language (EML)
- Georeferencing elements

CSDGM Profiles
- CSDGM
- ISO 19115: 2003 Geographic information - Metadata

Climate
- NetCDF Climate and Forecast (CF) Metadata Conventions

Astronomy
- Astronomy Visualization Metadata Standard
The landscape

Metadata activities
Research Data Consortia

Data repository services

RDA
Research Data Sharing without barriers

DRYAD

Dataverse Project

DataCite

EARTH CUBE

TreeBASE
A Database of Phylogenetic Knowledge

DataONE
Data Observation Network for Earth

ICPSR

Metadata Standards Directory
Things to watch

• Diverse standards, diverse applications
• Infrastructure, infrastructure, infrastructure
  – Vocabularies
  – Entities
  – Linking
• Service-oriented
  – Metadata as a service
  – Infrastructure as a service
• Challenges in filling the gaps between fast data growth and slow metadata creation
References
