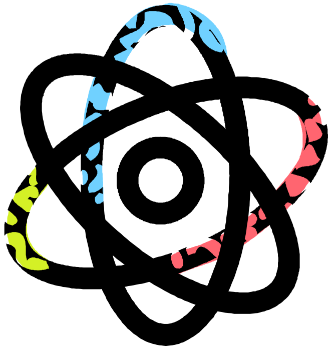


Metadata as the Underpinning of Sustainable and Effective Access to Scientific Data

THE SCHOOL OF
INFORMATION
STUDIES
at Syracuse University
"Expanding human capabilities through information"

Jian Qin

School of Information Studies, Syracuse University, Syracuse, NY, USA

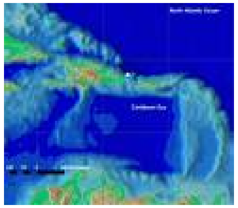
jqin@syr.edu

Topics in this presentation

- Scientific data objects
- Problems in metadata for scientific data
- Who is addressing the problems
- Solutions to the problems
- Implications for CODATA

Scientific data objects

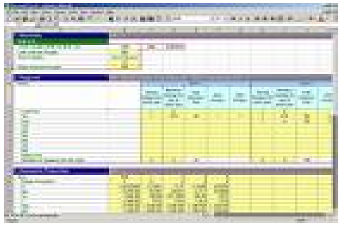
Oceanographic data



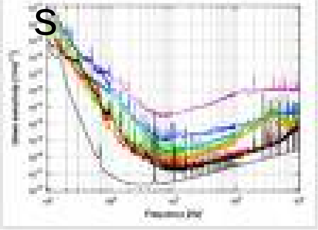
DNA sequence

Plant specimens

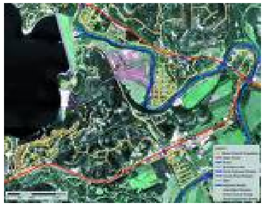
Spreadsheets



Field observations

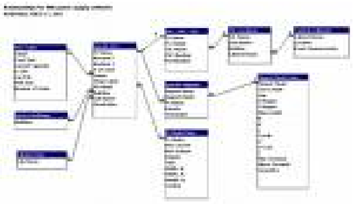


Geospatial data



XML

Databases



Scientific data are referred to the raw data that have been collected or generated in many ways, such as by measurement or observation of the environment, by carrying out an analytical experiment or by running a computer simulation.

Metadata for scientific 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, S., & Mathews, B. (2004). CCLRC scientific metadata model: version 2. CCLRC Technical Report: DL-TR-2004-001. Retrieved July 27, 2006, from http://epubs.cclrc.ac.uk/bitstream/485/csmdm.version-2.pdf

Data discovery problems

- Data set archives are becoming increasingly distributed
  - – Institutions may be unable to serve their own data and need to send it to a publicly accessible repository
  - – Contributing groups need to maintain their own data remotely
  - – Security
  - – Data discovery cross data servers

Data description problems (1)

- Insufficient metadata description of models, experiments, etc. that generate data
  - – e.g., weather and climate modeling
- Dissimilarity among metadata from diverse sources

  e.g., multiple metadata conventions exist for weather and climate models and experiments:
  - NetCDF (Network Common Data Format)
  - Climate and Forecast (CF)
  - COARDS (Cooperative Ocean/Atmosphere Research Data Service)

Kinter III, J. L. & Taylor, K. E. (2005). Data Issues for WCRP Weather and Climate Modeling, a white paper based on presentations at the First Session of the WCRP Modeling Panel, Exeter, UK, October 6, 2005. http://copes.ipsl.jussieu.fr/Organization/COPESStructure/Reports/WMP1/WMP1_KinterTaylor.doc

Data description problems (2)

- Metadata descriptions for scientific data
  - are operated under different paradigms
  - have different terminology for similar concepts
  - need to be able to compare data collected in different contexts with different technologies
- Different metadata needs for
  - current and future users of primary data
  - expert/familiar users and other users
- Cost/benefit imbalance for metadata creators (who bear the burden) and users (who enjoy the benefits)

Gritton, B. (1994). Metadata comments. http://www.llnl.gov/liv_comp/metadata/papers/comments-gritton.html

Who’s addressing the problems?

Metadata standards in scientific domains:

- ISO 19115:2003 Geographic Information – Metadata: http://www.isotc211.org/
- FGDC Metadata Standards: http://www.fgdc.gov/metadata/
  - Extensions:
    - Biological Data Profile: http://www.nbii.gov/datainfo/metadata/standards/
    - Shoreline Metadata Profile of the Content Standards for Digital Geospatial Metadata: http://ioc.unesco.org/Oceanteacher/OceanTeacher2/02_InfTchSciCmm/02_Meta/06_MetaStds&Form/FGDC/sprofile.pdf
- NetCDF Climate and Forecast (CF) Metadata Convention:

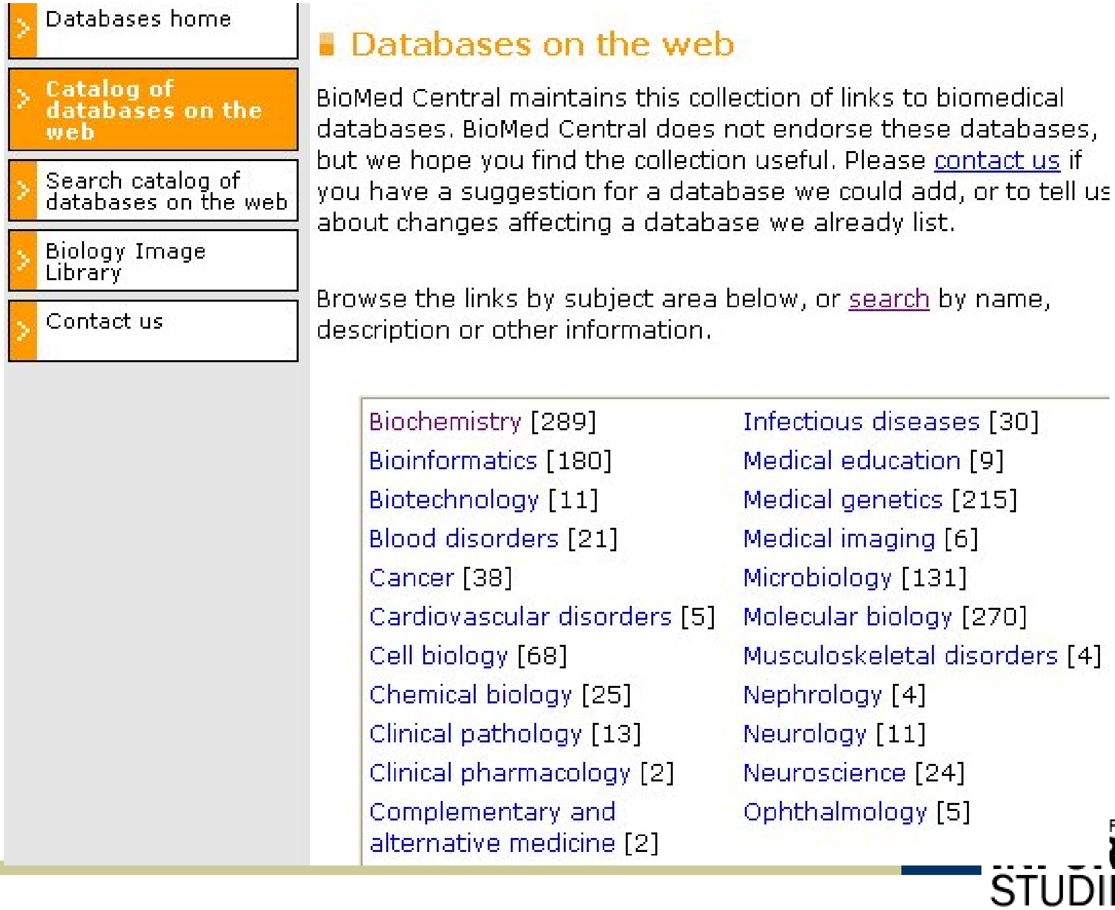
  http://home.badc.rl.ac.uk/lawrence/blog/2005/11/09/the_future_of_cf

General metadata standards:

- Dublin Core, Data Documentation Initiative, METS, etc.

What are the solutions? (1)

- Service-oriented metadata:
  - Directory services of datasets, databases, repositories


Databases home
Catalog of databases on the web
Search catalog of databases on the web
Biology Image Library
Contact us

Databases on the web

BioMed Central maintains this collection of links to biomedical databases. BioMed Central does not endorse these databases, but we hope you find the collection useful. Please contact us if you have a suggestion for a database we could add, or to tell us about changes affecting a database we already list.

Browse the links by subject area below, or search by name, description or other information.

Biochemistry [289]
Bioinformatics [180]
Biotechnology [11]
Blood disorders [21]
Cancer [38]
Cardiovascular disorders [5]
Cell biology [68]
Chemical biology [25]
Clinical pathology [13]
Clinical pharmacology [2]
Complementary and alternative medicine [2]

Infectious diseases [30]
Medical education [9]
Medical genetics [215]
Medical imaging [6]
Microbiology [131]
Molecular biology [270]
Musculoskeletal disorders [4]
Nephrology [4]
Neurology [11]
Neuroscience [24]
Ophthalmology [5]

What are the solutions? (2)

- Object-oriented metadata:
  - Description of scientific data objects
    - Documentation of data origins, processing history, collection methodologies, measurement precision, etc.
    - Content description of data objects
    - Administrative and maintenance of data objects

How metadata is created?

- Manual creation:
  - Collection level
  - Mainly for discovery purpose
- Automatic or semi-automatic generation
  - Object level
  - Mainly for identifying, locating, and selecting data resources
  - Many approaches exist

Approaches in automatic generation of object-level metadata

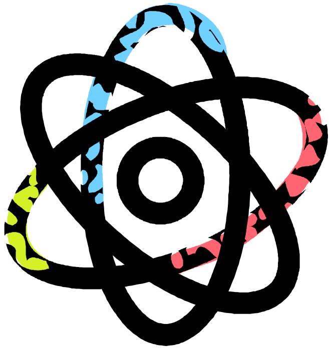
- *Extraction*: use natural language processing, machine learning, or other methods to extract metadata from data objects
- *Assignment*: assign metadata values to data objects against a controlled vocabulary or metadata scheme based on automatic analysis result
- *Combining automatic extraction/assignment* with human intervention

Implications to CODATA (1)

- Service-oriented metadata:
  - Vital for data discovery in distributed, remote access to data archives
  - Vital for data repository administration and maintenance in a distributed environment
  - CODATA metadata directory service: a distributed and coordinated effort is needed to
    - enable data discovery by geographic coverage
    - enable data discovery by temporal coverage
    - enable data discovery by disciplinary or cross-disciplinary domains
    - enable data discovery by cross-sections of all the above

Implications to CODATA (2)

- Object-oriented metadata:
  - – Vital for identifying and using data objects
  - – CODATA inventory of domain metadata standards for scientific disciplinary and interdisciplinary fields is needed to:
    - enable data interoperability
    - enable data quality assessment
    - enable long-term preservation and use



Question?

THE SCHOOL OF
INFORMATION
STUDIES
at Syracuse University
"Expanding human capabilities through information"