Metadata for Scientific Data in China: an Overview

Hou Yanfei
Computer Network Information Center, CAS
2006-10-25
Outline

- Introduction: Scientific Data and Metadata
- Representative Projects
- National/Industry Standards on Metadata for Scientific Data
- Journal Articles about Metadata for Scientific Data during 1997-2005
- Conclusion and Further Works
Introduction: Scientific Data

Scientific Data

- refer to any and all complex data entities from observations, experiments, simulations, models, and higher order assemblies, along with the associated documentation needed to describe and interpret the data.

--NSF’s Cyberinfrastructure vision for 21st Century Discovery

- differ from the information objects that libraries and librarians have concern about.

- access to scientific data is increasingly crucial for scientific research and discovery.
Introduction: Metadata for Scientific Data

- Metadata: data about data
- Metadata for scientific data
  - summarize the content, context and structure of scientific data
  - usually regarded as a subset of data
  - be an infrastructure for the access to and the utilization of scientific data
Introduction: Some Experiential Observations

 Metadata schemas for scientific data VS. DC and application profiles based on DC

 complexity, structuralism VS. simpleness, minimalism/cautious structuralism

 much richer context information, especially provenance information (information on history and origins) on the objects VS. less context information

 usually give emphasis on the description of structure information of the object VS. usually ignore that
Introduction: Metadata for Scientific data in China

- Metadata for scientific data has been being studied since the mid-1990s in China.
- For scientific data, “metadata” was first studied and used by communities involved with the management and utilization of geospatial data.
Representative Projects

aN China - Scientific Data Sharing Project (CSDSP)

• initiated in 2001, supported by Ministry of Science and Technology of the People’s Republic of China

• goal: to form a distributed scientific data sharing system on the national level that collects and re-organizes all possible scientific data funded by the government, and makes them available to the public

• standardization is an important task in the project.
Representative Projects-CSDSP

 Metadata specifications for scientific data in CSDSP

- General specifications
  - Metadata Standardization Principle and Method
  - Content Standard for Metadata of Scientific Data
  - XML/XSD Marking Rules for Metadata

- Domain-specific metadata specifications
  - Metadata for Meteorological data
  - Metadata for Agricultural Data
  - Metadata for Oceanographic Data
  - ...
Representative Projects-CSDSP

12 data centers

Metadata retrieval area

URL: http://www.sciencedata.cn/index.php
Representative Projects

Scientific Databases Project in CAS

- initiated in 1986, mainly funded by Chinese Academy of Sciences.
- a large-scale system of multi-discipline, distributed scientific databases which collects scientific data produced by the institutes (and the scientists) in Chinese Academy of Sciences and provides them to the scientists and others to share.
- standardization began from 2001.
Representative Projects-SDB

- Metadata specifications for scientific data in SDB

**Related systems in SDB**

- General Metadata Management System (in XML)
- Metadata Management System for Ecological Research Data (in MySQL)
- SDB Metadata Registry System
Representative Projects-SDB

Databases in SDB system. More than 500 databases, 13TB

Metadata retrieval area

URL: http://www.csdb.cn
Representative Projects

Other projects

- Network of Chinese Sustainable Development Information (URL: [http://www.sdinfo.net.cn/](http://www.sdinfo.net.cn/))
- Chinese Ecosystem Research Network (URL: [http://www.cern.ac.cn/](http://www.cern.ac.cn/))

- ....
National Standards

** GB/T 19710-2005 Geographic Information – Metadata （地理信息 元数据）
- Local adoption of ISO 19115:2003
- There are several small modifications to ISO 19115:2003.

** GB/T xxxx-xxxx Metadata for Ecological Data （生态科学数据元数据）
- One national standard which will be announced soon
- Data may be spatial data or non-spatial data.
Industry Standards

- QX/T 39-2005 Core Metadata Content of Meteorological Dataset (气象数据集核心元数据)
- TD/T 1016-2003 Core Metadata Standard for Land and Resources Information (国土资源信息核心元数据标准)
- Metadata for Spatial Information in Agriculture (农业资源空间信息元数据)

All of them are tightly related with GB/T 19710. Metadata instances conforming to anyone can be easily converted into metadata instances conforming to GB/T 19710.
Journal Articles about Metadata for Scientific Data during 1997-2005

The authors collect the journal articles related to metadata for scientific data through searching in China Journal Fulltext Database (CJFD ) in CNKI (http://www.cnki.net.cn/).

- Retrieve by the keyword “metadata”
- Browse the title, abstract or full text of every article in the collection of search results, and identify whether it’s relevant to metadata for scientific data or not.
- Assign a relevance value (s, m, or w) to every article relevant to metadata for scientific data according to the following criterion.
- Supplement controlled keywords to the articles, through which the articles can be well grouped by subject.
### Journal Articles about Metadata for Scientific Data during 1997-2005

<table>
<thead>
<tr>
<th>Relevance value</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>S (Strongly relevant)</td>
<td>All or mostly all content of the article is about metadata. The topic discussed in the article may be metadata schema, metadata implementation, metadata application, or others.</td>
</tr>
<tr>
<td>M (Middlingly relevant)</td>
<td>At least one important part in the article is about metadata.</td>
</tr>
<tr>
<td>W (Weakly relevant)</td>
<td>The article mentions metadata, but metadata is unimportant to the topic which the article discusses.</td>
</tr>
</tbody>
</table>
Results and Analyses

- The first relevant article was issued in 1997. It discusses metadata for geographic information.
- 360 relevant articles were published from 1997 to 2005, and of them 312 articles are strongly or middlingly relevant to metadata for scientific data.
- The topics discussed in the articles include metadata schema, systems or tools for metadata implementation, application of metadata, metadata creation, description language and metadata, metadata and ontology, data integration and metadata, metadata quality, metadata interoperability, case study, and so on.
Results and Analyses (Cont.)

number of journal articles relevant to metadata for scientific data issued every year
number of journal articles strongly and middlingly relevant to metadata for scientific data every year
Results and Analyses (Cont.)

The pie chart shows the percentage of the amount of articles issued in the journals in different domains. Specifically:

- 46% of the articles are in the domain of geoscience.
- 21% of the articles are in the domain of computer science.
- 33% of the articles are in other journals.

The chart illustrates the distribution of articles across these domains.
About 33% of articles were issued in the journals in the domains of biology, environmental science, medical science, agriculture, chemistry, engineering, library and information science, and in some comprehensive journals.

12 articles (about 3% of total articles) about metadata for scientific data were published on the journals in the domain of library and information science.
About 70% of the articles are relevant to metadata for geospatial data.

- Geospatial data dominate in point of data amount in the kingdom of scientific data
- Geospatial data are produced not only in the domain of geoscience (geography, geology, meteorology, geodesy, hydrology, oceanography, geophysics and geochemistry included), but also in many other domains, such as ecology, environmental science and agriculture.

Other scientific data discussed in the articles include biological data, medical data, chemical data, and engineering data.
Conclusion

- In China, the importance of metadata has been recognized in many domains including geoscience, biology, agriculture, environmental science, medical science and others.
- In large-scale projects ‘CSDSP’ and ‘SDB’, many efforts to develop and implement metadata schemas and apply them in the discovery and use of scientific data have been done.
- Two national standards of metadata for scientific data, at least three industry standards, and many metadata schemas in specific projects or data systems.
- Further works on the topic of the presentation are needed.
Further Works

صب Further works will be done:

صب More and deeper analyses by bibliometric methods on the journal articles, for examples, statistics and analysis of the topics discussed in the articles, quotation analysis, …

صب A considerate and deep survey to the application of metadata for scientific data in China.
Thanks for your attention!