Metadata for Scientific Data in China: an Overview

Hou Yanfei

Computer Network Information Center, CAS

2006-10-25

科学数据库

Scientific Database

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

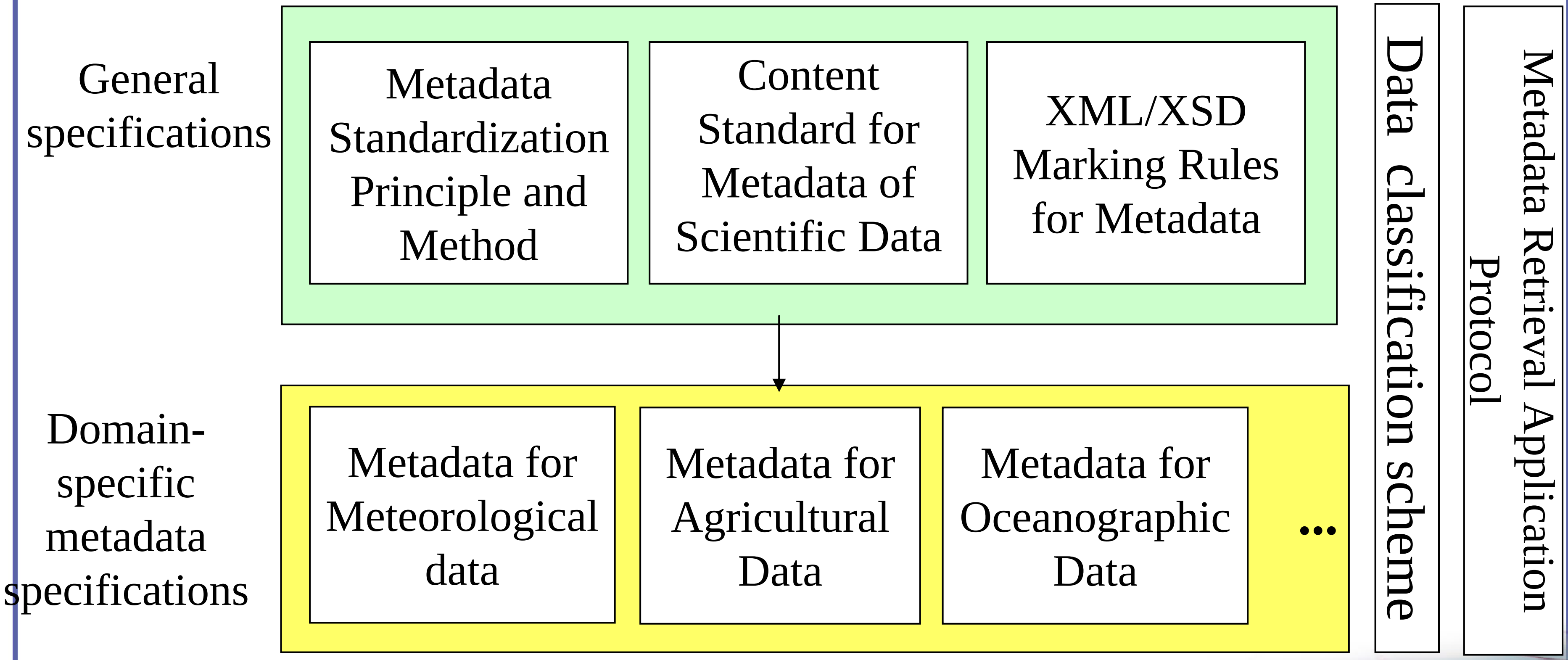
- Metadata for scientific data has been being studied since the mid-1990s in China.
- For scientific data, “metadata” was first studied and used by communities involved with the management and utilization of geospatial data.

Representative Projects

- 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

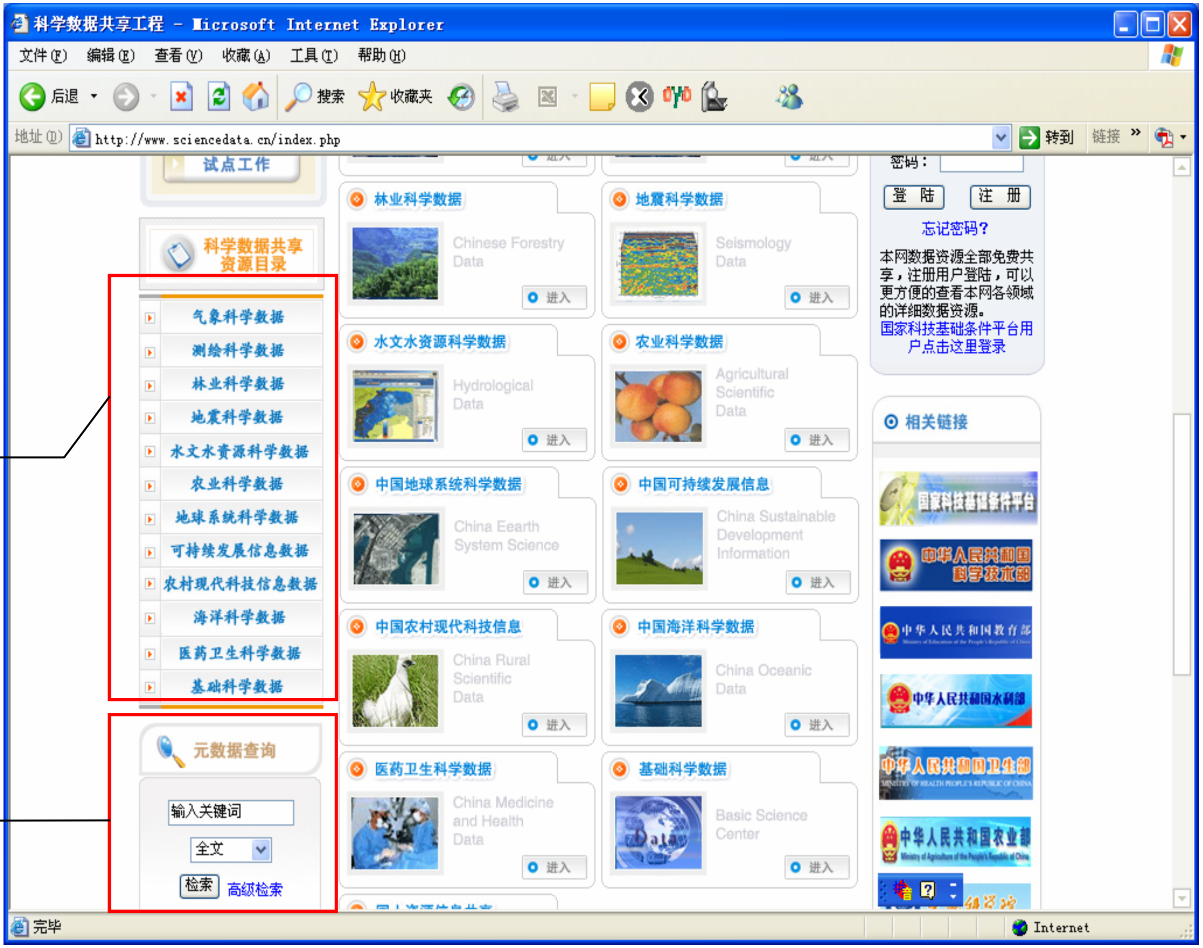
Metadata for Oceanographic Data

…

Data classification scheme

Metadata Retrieval Application Protocol

Representative Projects-CSDSP


科学数据共享工程 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹
地址(D) http://www.sciencedata.cn/index.php
转到 链接
试点工作
科学数据共享资源目录
元数据查询
输入关键词
全文
检索
高级检索
密码：
登 陆
注 册
忘记密码？
相关链接
国家科技基础条件平台
中华人民共和国科学技术部
中华人民共和国教育部
Ministry of Education of the People's Republic of China
中华人民共和国水利部
中华人民共和国卫生部
MINISTRY OF HEALTH PEOPLE'S REPUBLIC OF CHINA
中华人民共和国农业部
Ministry of Agriculture of the People's Republic of China
进入
完毕
Internet

- 气象科学数据
- 测绘科学数据
- 林业科学数据
- 地震科学数据
- 水文水资源科学数据
- 农业科学数据
- 地球系统科学数据
- 可持续发展信息数据
- 农村现代科技信息数据
- 海洋科学数据
- 医药卫生科学数据
- 基础科学数据

林业科学数据 Chinese Forestry Data

地震科学数据 Seismology Data

水文水资源科学数据 Hydrological Data

农业科学数据 Agricultural Scientific Data

中国地球系统科学数据 China Eearth System Science

中国可持续发展信息 China Sustainable Development Information

中国农村现代科技信息 China Rural Scientific Data

中国海洋科学数据 China Oceanic Data

医药卫生科学数据 China Medicine and Health Data

基础科学数据 Basic Science Center

本网数据资源全部免费共享，注册用户登陆，可以更方便的查看本网各领域的详细数据资源。

国家科技基础条件平台用户点击这里登录

12 data centers

Metadata retrieval area

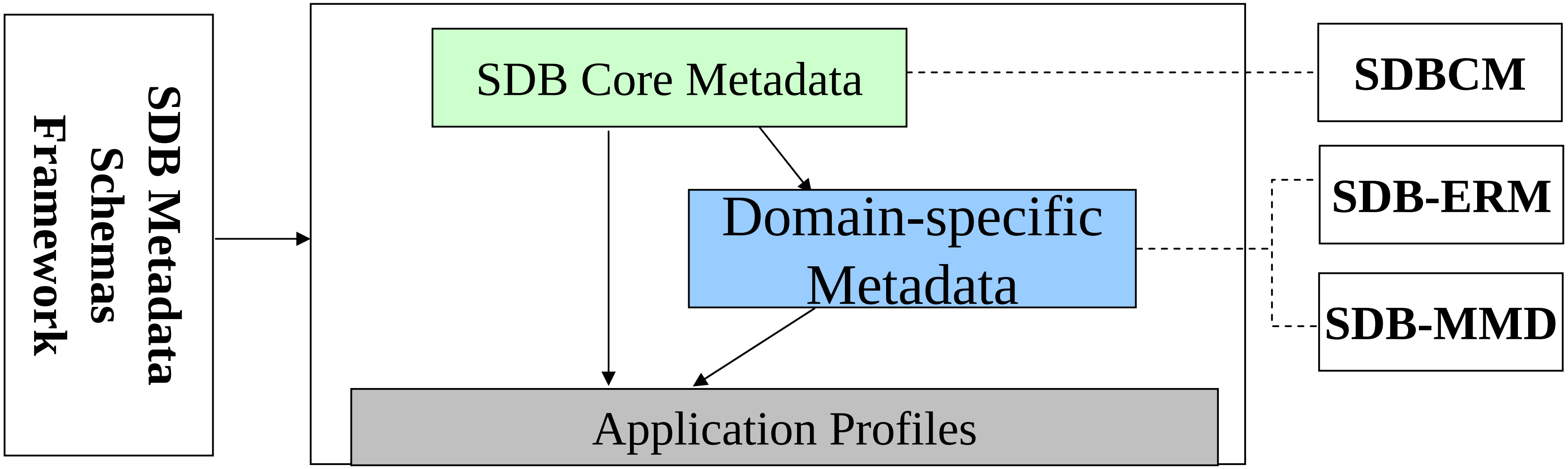
URL: http://www.sciencedata.cn/index.php

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

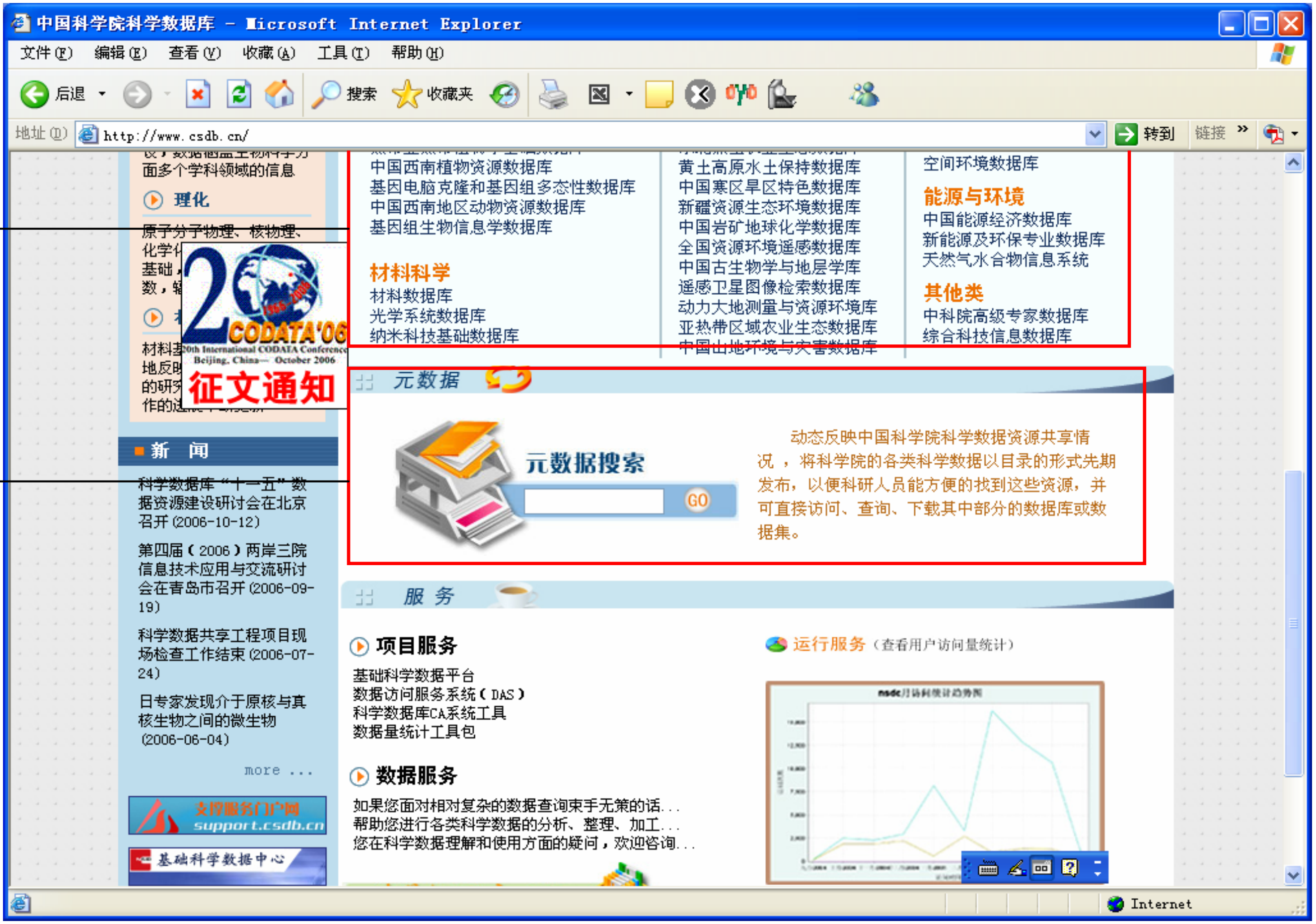


- Related systems in SDB
  - General Metadata Management System (in XML)
  - Metadata Management System for Ecological Research Data (in MySQL)
  - SDB Metadata Registry System

Representative Projects-SDB

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

Metadata retrieval area



中国科学院科学数据库 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退
搜索
收藏夹
地址(D) http://www.csdb.cn/
转到
链接

面多个学科领域的信息

理化

原子分子物理、核物理、

化学化

基础，

数，辐

材料基

地反映

的研究

作的进

20 CODATA'06

1966-2006

20th International CODATA Conference

Beijing, China— October 2006

征文通知

中国西南植物资源数据库
基因电脑克隆和基因组多态性数据库
中国西南地区动物资源数据库
基因组生物信息学数据库

材料科学

材料数据库
光学系统数据库
纳米科技基础数据库

黄土高原水土保持数据库
中国寒区旱区特色数据库
新疆资源生态环境数据库
中国岩矿地球化学数据库
全国资源环境遥感数据库
中国古生物学与地层学库
遥感卫星图像检索数据库
动力大地测量与资源环境库
亚热带区域农业生态数据库
中国山地环境与灾害数据库

空间环境数据库

能源与环境

中国能源经济数据库
新能源及环保专业数据库
天然气水合物信息系统

其他类

中科院高级专家数据库
综合科技信息数据库

新　闻

科学数据库“十一五”数据资源建设研讨会在北京召开 (2006-10-12)

第四届（2006）两岸三院信息技术应用与交流研讨会在青岛市召开 (2006-09-19)

科学数据共享工程项目现场检查工作结束 (2006-07-24)

日专家发现介于原核与真核生物之间的微生物 (2006-06-04)

more ...

支撑服务门户网 support.csdb.cn

基础科学数据中心

元数据

元数据搜索

GO

动态反映中国科学院科学数据资源共享情况，将科学院的各类科学数据以目录的形式先期发布，以便科研人员能方便的找到这些资源，并可直接访问、查询、下载其中部分的数据库或数据集。

服　务

项目服务

基础科学数据平台
数据访问服务系统（DAS）
科学数据库CA系统工具
数据量统计工具包

数据服务

如果您面对相对复杂的数据查询束手无策的话...
帮助您进行各类科学数据的分析、整理、加工...
您在科学数据理解和使用方面的疑问，欢迎咨询...

运行服务（查看用户访问量统计）

URL: http://www.csdb.cn

Representative Projects

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

National/Industry Standards on Metadata for Scientific Data

- 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.

National/Industry Standards on Metadata for Scientific Data (Cont.)

- 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.

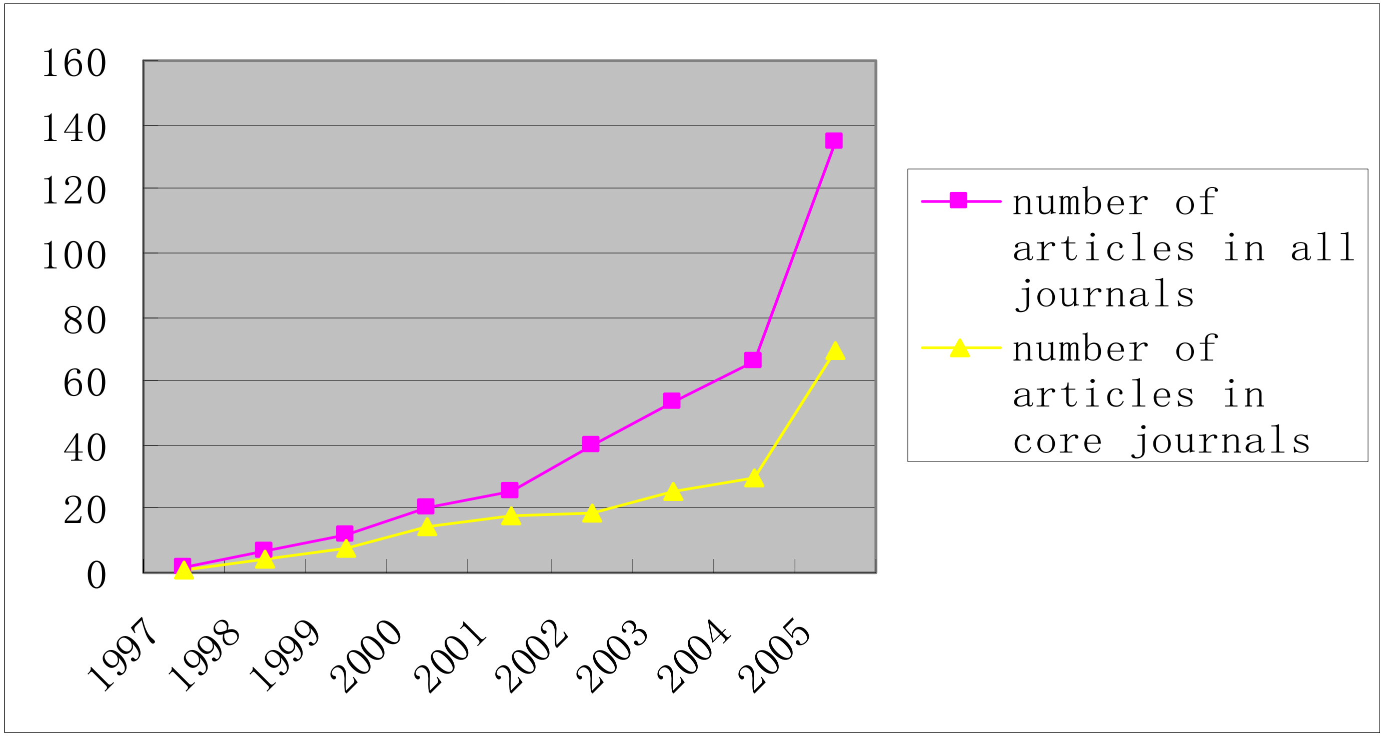
Journal Articles about Metadata for Scientific Data during 1997-2005

Relevance value	Characteristics
S (Strongly relevant)	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.
M (Middlingly relevant)	At least one important part in the article is about metadata.
W (Weakly relevant)	The article mentions metadata, but metadata is unimportant to the topic which the article discusses.

Results and Analyses

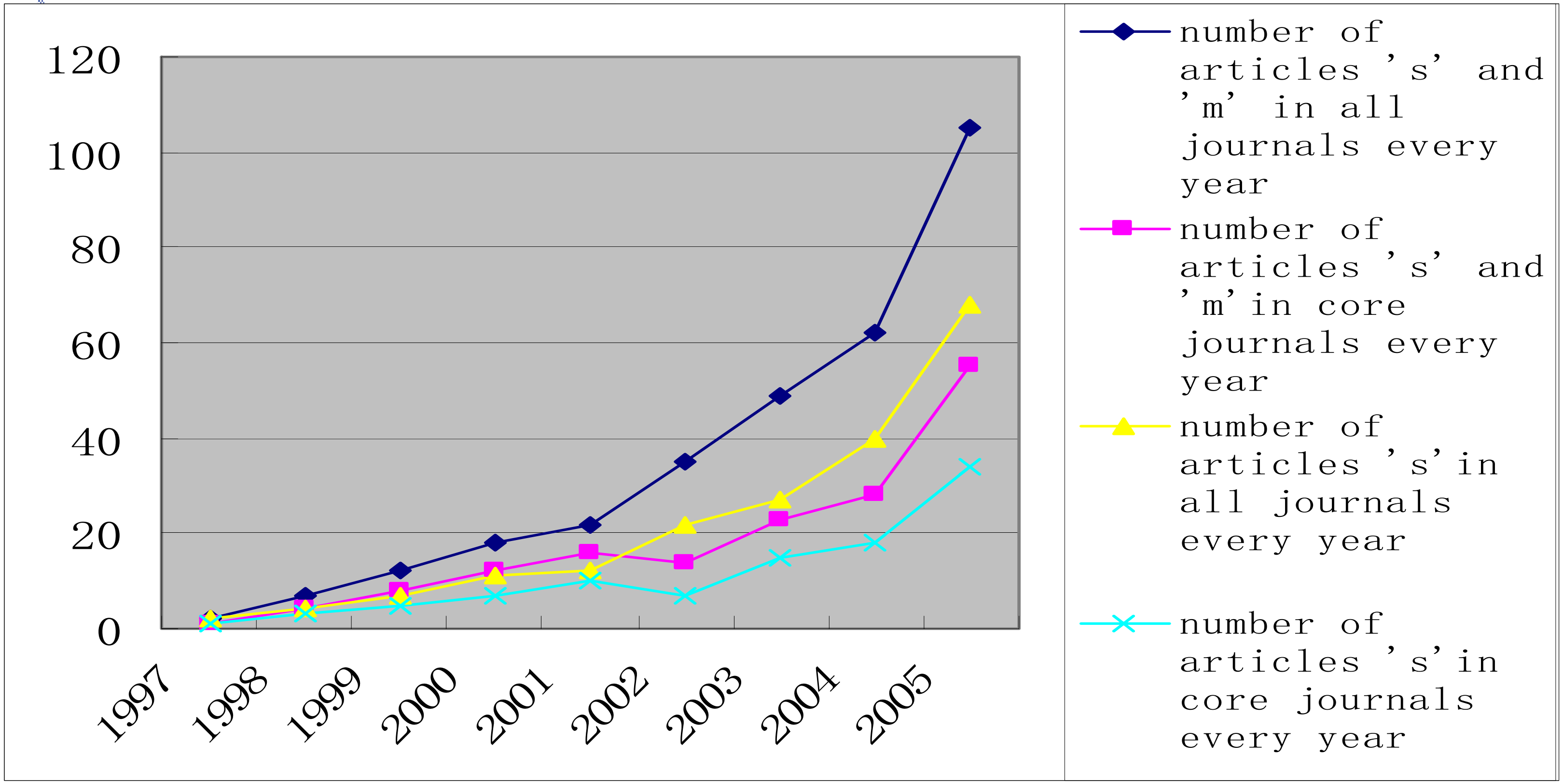
- The first relevant article was issued in 1997. It discusses metadata for geographic information.
- 360 relevant articles were published from 1997 to 2005, and of them 312 articles are strongly or middlingly relevant to metadata for scientific data.
- The topics discussed in the articles include metadata 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.)



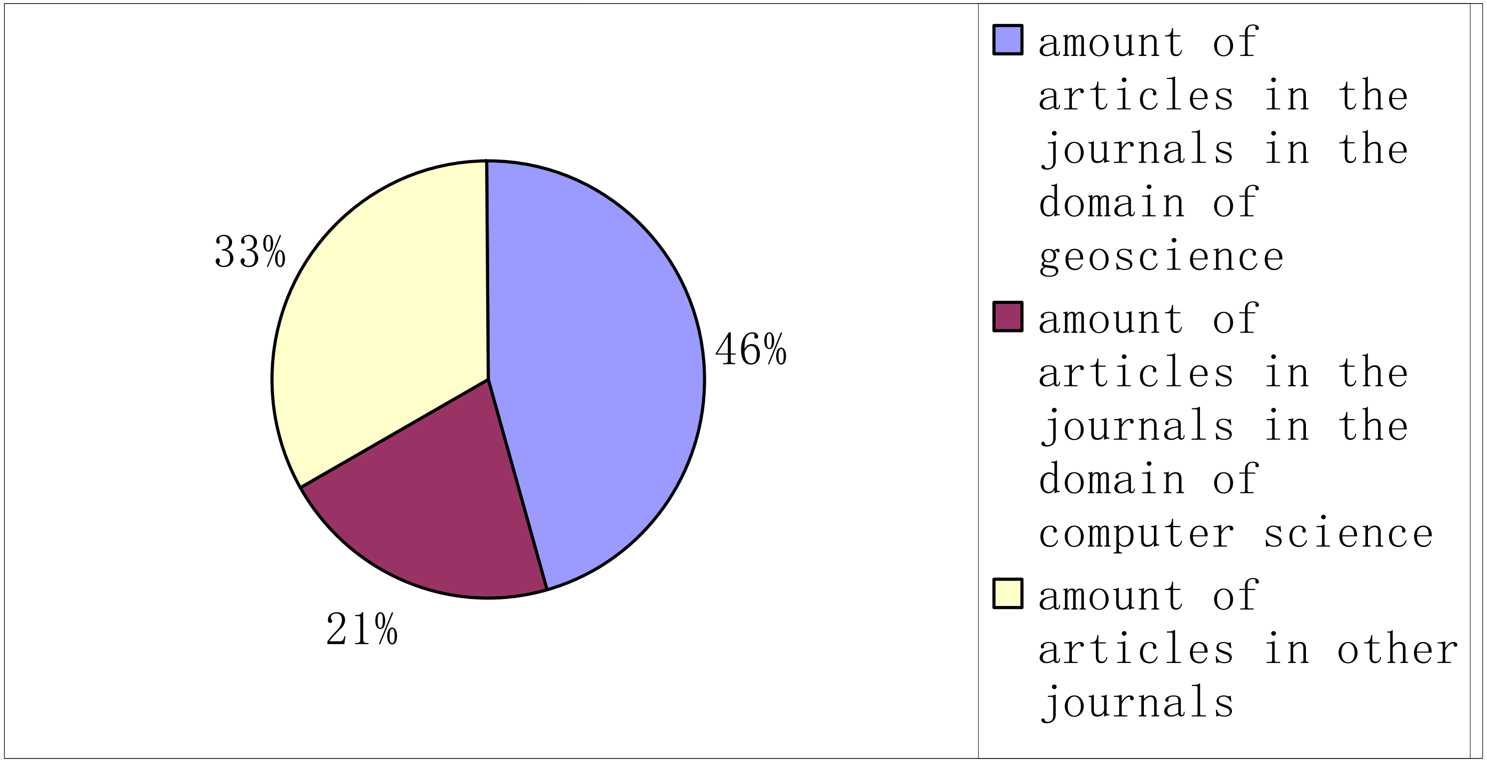
number of journal articles relevant to metadata for scientific data issued every year

Results and Analyses (Cont.)



number of journal articles strongly and middlingly relevant to metadata for scientific data every year

Results and Analyses (Cont.)



percentage of the amount of articles issued in the journals in different domains

Results and Analyses (Cont.)

- 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.

Results and Analyses (Cont.)

- 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

- Further works will be done:
  - 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, …
  - A considerate and deep survey to the application of metadata for scientific data in China.

Thanks for your attention!