

# 国家资源与环境数据库元数据管理研究<sup>①</sup>

曹彦荣<sup>②1</sup> 吴洪桥<sup>1</sup> 毕建涛<sup>1</sup> 黄裕霞<sup>2</sup> 何建邦<sup>1</sup>

(1 中国科学院地理科学与资源研究所信息室, 北京 100101)

(2 中国科学院遥感应用研究所, 北京 100101)

**摘要:** 本文从建立国家资源与环境数据库的元数据库入手, 用国际通用的面向对象的通用建模语言 UML 来设计元数据的结构, 并辅以数据字典加以说明。用适合于描述元数据的可扩展语言 XML 来描述元数据标准, 很大程度上实现了元数据的有效管理和可扩展性, 为资源与环境数据库的资源快速导航和高效查询提供了有力的支持。

**关键词:** 元数据标准; 元数据管理; UML; XML

**中图分类号:** P208

## 1 引言

国家资源与环境数据库拥有自然、社会、经济、人口等各方面的数据库群。不仅数据量庞大, 而且分属不同硬件平台, 由不同的软件支持, 具有不同类型的结构和内容。如何解决数据的生产、管理、组织和共享已经成为急需解决的问题。作为数据生产者, 迫切需要一套有效的数据管理和维护办法, 而数据用户同样也要求能够从生产者那里获取快捷、安全、有效、全面的服务, 以便从海量的资源与环境数据中快速、准确地发现、访问、获取和使用所需的数据。但是, 在网络上如何共享这些信息。还存在若干问题:

(1) 数据一旦在网络上共享, 就涉及到版权的问题;

(2) 数据的使用者在请求使用前, 首先要了解他所需的空间数据的内容、覆盖范围、质量、管理方式、数据的所有者和数据的提供方式等有关信息。

(3) 由于网络速度的瓶颈, 没有元数据而直接下载将浪费大量人力、财力。

空间元数据及其提供信息服务为整个问题的解决提供了一个可行的方法。

一般认为, 研究空间元数据在空间数据库共享的过程中的作用, 必须从 3 个方面入手: ①元数据的标准(Profile); ②元数据管理系统(Spatial Metadata Management System, SMMS) 的研制; ③以元数

据为基础提供的信息服务(Information Service)。这里着重讨论前两个问题, 即元数据管理方面的问题。

笔者所在的地理信息共享研究组(Research Group on Geographic Information Sharing, RGIS)在参照了 ISO/TC211 的最终提交的版本(ISO/TC211 Geographic information/Geomatics: N 1142), 以及本项目内部的实际需求后, 进行了研制项目的空间元数据标准的工作: 使用通用建模语言 UML 对元数据的结构进行了设计, 附以相应的元数据数据字典, 形成了与 ISO/TC211 接轨的标准; 在软件的编制方面, 采用了工业标准的 XML 和 JAVA 开发工具开发了运行于 Internet 上的元数据管理系统。

## 2 元数据标准的设计

地理信息元数据标准一直是国际地理信息社会的研究热点, 包括美国联邦地理数据委员会(FGDC)、欧洲地理信息标准化委员会(CEN/TC287)、国际标准化组织地理信息委员会(ISO/TC211) 等机构都一直致力于地理信息元数据标准的研究。目前, ISO/TC211 经过近 10 个版本的更新, 已经进入委员会草案阶段。我们国家在标准研究方面也做出了大量的努力, 现在已经由国家地理信息基础中心完成了国家地理信息标准。因此, 国家资

① 收稿日期: 2002-01-21.

基金项目: 十五国家攻关课题(2001BA608B-01); 国家自然科学基金(KJ951-B1-703); 中科院知识创新工程项目(KZCX2-308-02) 资助。

② 作者简介: 曹彦荣(1974-), 男, 在读博士研究生, 研究领域为地理信息共享, 元数据和信息服务。

源与环境数据库元数据标准是在国际和国家标准即将发布的前提下完成的。

考虑到标准的完整性、准确性、结构性和国际标准的一致性, 参照了许多最新国际标准、国家标准或行业标准, 如: ISO/TC211 在 2001 年 8 月提出的最终草案《inal Text of CD 19115 Geographic information – Metadata》, 国家基础地理信息中心在 2001 年提出的元数据标准《基础地理信息数字产品元数据》, 同时考虑到国家资源与环境数据库现存数据主要是以空间信息( 矢量、影像、栅格等) 为主的空间数据库和以属性数据为主且具有空间定位信息的空间数据库, 以及图书文献档案资料目录库、法律法规数据库等非空间数据集。制订了适用的标准。

2.1 元数据结构设计

标准包括 7 大类和 3 类公共数据:

7 大类包括: 标识信息( Identification) 是关于数据集的基本信息; 数据质量信息( Data Quality) 为对数据质量进行总体评价的信息; 空间数据表示信息( Spatial Data Organization) 是数据集中空间信息的组织方法; 空间参照系统信息( Spatial Reference) 为数据集中坐标的参考框架以及编码方式的描述; 内容信息( Entity and Attribute) 是关于数据集内容的细节信息; 分发信息( Distribution) 是关于数据发行

和获取的信息; 元数据参考信息( Metadata Reference), 即元数据当前状况及其负责部门的信息。

3 类公共数据包括: 引用信息( Citation), 引用和参考数据集时所需的简要信息; 时间信息( Date), 提供有关事件的日期和时间的信息; 负责单位联系信息( Contact), 在主要子集中被引用的有关个人或组织的联系信息。公共数据不单独使用, 作为其他元索引用的对象。

由于元数据各大类之间存在复杂的逻辑结构和关系, 如果用面向对象的方法来分析: 包括继承( 单一和多重)、组成、聚集、关联等关系。公共数据作为引用对象, 又频繁地被其他类所引用。如果仅仅用常用的二维表格来描述元数据, 将很难表达清楚。因此需要用图形化的手段来表示。

通用建模语言( Unified Modeling Language) 作为一种面向对象设计结构的图形表示法, 已经作为一个标准, 被对象管理组织( Object Management Group, OMG) 及其它组织所采用。ISO/ TC211 正在制订的地理信息系列标准中包括目前提出的最终草案, 也都采用 UML 作为其概念模式语言。因此, 这里我们采用了 UML 静态结构类图来表达元数据各个类的逻辑结构和关系, 类和类的属性用数据字典来描述。形成了完整的元数据标准。这一点也是和国际标准相接轨的。

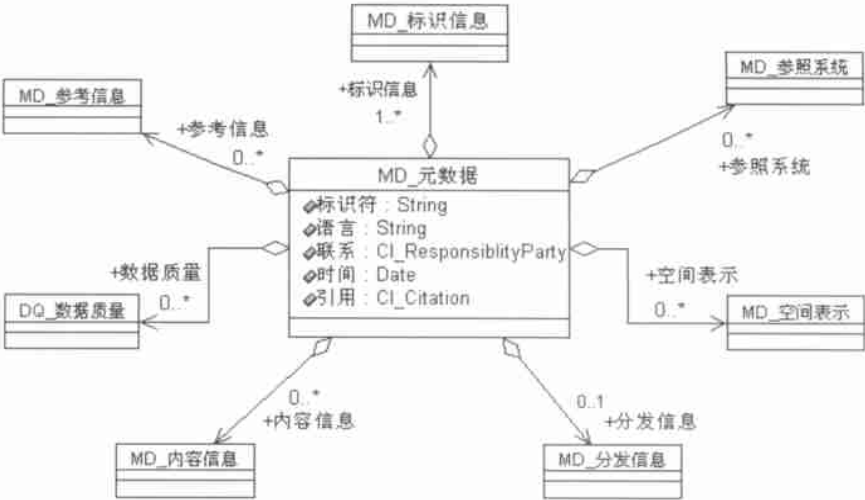


图 1 MD \_\_元数据类的定义以及与其他元数据类之间的约束关系

Fig. 1 The class MD \_\_metadata’s definition and containment relationships with the other metadata classes

元数据标准的类之间的关系包括泛化( Generalization, 相当于面向对象中的继承)、聚集( Aggregation)、组成( Composition)、关联( Association) 等等。如图 1 中 MD \_\_元数据是由 MD \_\_标识信息、

MD \_\_质量信息、MD \_\_参考信息、MD \_\_内容信息、MD \_\_参照系统、MD \_\_内容表示、MD \_\_分发信息聚合而成的,而且这种关系是一种单向聚合关联。图中的数字表示多重性,如MD \_\_元数据和MD \_\_标识信息之间是 1..\*,表示元数据类有一个或者多个标识信息类。此外,重要的是还可以根据UML 的构造型(UML model stereotypes)在已经定义的模型元素之上构造新的模型元素,实现元数据的扩展。

2.2 元数据数据字典

数据字典描述了用UML 设计的元数据的特征,以子集、实体和元素为单位,在这个层次结构描述了实体和元素的结构关系和属性。有如下属性:名称、缩写名、定义、约束条件、最多出现次数、数据类型和值域。

表 1 国家资源环境数据库元数据数据字典示例  
Tab.1 Example of metadata for National Resource and Environment Database data dictionary

序号	中文名称	定 义	约束条件	最多出现次数	数据类型	值域
1	元数据-标识	关于数据集的数据和服务的标识信息	M	1	元数据子集	序号 1-55
2	数据集中文名称	数据集中文名称,全称	M	1	文本	自由文本
3	数据集英文名称	数据集英文名称,全称	O	1	文本	自由文本
4	日 期	数据集发布或更新日期	M	1	整形	YYYYMMDD

用UML 表示的元数据类图和数据字典组成了完整的元数据标准,并且有清晰的逻辑结构,易于理解,易于编程实现。

3 元数据的管理

3.1 元数据管理应考虑的重点

UML 描述的元数据结构以及相应的数据字典清楚地为用户描述了如何用元数据来说明数据库,所有数据集都可以用元数据来描述。但是这些如何管理元数据库,如何更有效的帮助用户获得数据,需要有良好的元数据管理系统。

元数据的管理,目标是为了对地理元数据进行

获取、检查、存储、处理和应用。地理信息共享的问题都是在网络的基础之上的,因此,所谓管理问题也是在网络上而言,涉及到WEB 浏览器、WEB 服务器、元数据服务器和数据库服务器等软件部件之间一系列的请求和响应过程。

目前国际上和国内都建立了许多元数据系统。如:由FGDC 推荐的I-Site 免费软件包,它是FGDC 推荐的用于建设空间信息交换中心(Clearinghouse)的软件包。想建立自己的空间信息交换中心的机构可以下载安装这个软件包,进行配置就可以使用。国家空间信息交换中心(NGIEC)已经有这种网站(www.nsi.gov.cn),用户可以通过浏览器来查询各个节点上相关的空间信息的元数据。其他比较著名的有已经商业化的由Blue Angel Technologies 开发的MetaStar 系列,ARC/INFO 的metadata DOCUMENT 等等。

分析这些元数据管理系统,可以得出它们都具有以下主要功能模块:

元数据浏览器:负责空间数据库的浏览和导航,提供查询界面,以及数据预览功能;元数据编辑器:实现元数据的各种编辑功能,如新建、插入、删除、更新等。

元数据服务器:管理元数据数据库,并在网络上进行发布。

联系到实际应用中的问题,除了实现上述功能,在实现国家资源与环境空间数据库的元数据管理系统的时候,还应该重点考虑以下几个问题:

(1) 元数据的组织应该清晰地映射出元数据UML 类图中各个类及它们的关系,具备良好的查询策略。

(2) 由于项目涉及到自然、社会、经济和人口各个领域,元数据标准不可能涵盖所有方面。因此,只是取了一些重要的、公共的元数据实体(否则的话,元数据标准中的实体将会变得庞杂不堪,实际应用中却产生大量冗余)。但是,对于有些领域来说,需要扩展自己的元数据(必须按照一定的规则,并通过一致性测试),否则将不能有效描述数据集。所以元数据的可扩展性非常重要。

(3) 为方便使用,可以导出和导入各种符合规范的元数据文件。元数据的存储可以采取文件系统和数据库系统相结合的模式。

(4) 用户界面友好,易于操作。

### 3.2 XML 技术对库元数据标准的描述

#### (1) 用 XML 映射 UML 类图

考虑到以上需求, 采用目前流行的 XML 技术应该是理所当然的。可扩展标记语言 XML (Extensible Markup Language) 是继 HTML 之后的又一种 web 标记语言, 它为用户提供了灵活的标记扩展机制, 使得不同内容的资源能以格式良好的 (well-formed) 自定义的标记元素来表现。本质上, XML 是一种元语言, 是一种用于描述其它语言的语言。它具有以下特点: 自描述性, 可以自己定义标签 (tag), 开发者可以根据自己的需求, 定义自己的文档类型说明 (Document Type Definition, DTD); 半结构性, 适合描述层次型数据; 具备良好的扩展性; 而且与平台无关, 适于在网络上传输等等。因此, 对于开发元数据管理系统而言, XML 可以完全映射 UML 定义的元数据的各种类和类之间的关系, 并且可以扩展元数据, 同时满足网络运行的要求。

#### (2) 用 XML 描述元数据

利用 XML, 我们可以描述国家资源环境数据库元数据标准。具体工作是定义标准的 XML DTD。定义的 DTD 部分如下 (主要以标识信息中一部分为例, 其他以 “.....” 省略):

```
<!-- 国家资源环境数据库元数据的 DTD -->
<!ELEMENT metadata ( idinfo, dataqual?,
continfo?, distrib?, spatrep?, refsystem ?,
metainfo) >
<!-- 标识, 数据质量, 内容, 分发, 空间数据
表示, 参照系统, 元数据参考 -->
<!-- 标识信息部分 -->
<!ELEMENT idinfo(cn __name, en __name,
date, version, purpose, status, geobox.....)>
<!ELEMENT chinesename (# PCDATA)>
<!ELEMENT englishname (# PCDATA)>
<!-- 时间信息在 timeinfo 中有描述 -->
<!ELEMENT version (# PCDATA)>
<!ELEMENT purpose (# PCDATA)>
<!ELEMENT status EMPTY ) >
<!-- ATTLIST status
progress ( Complete | In work | Planned
) "Planned"
update ( Continually | Daily | Weekly |
```

```
Monthly| Annually| Unknown| As needed | Irregu-
lar | None planned ) "Unknown" >
<!-- ELEMENT geobox EMPTY >
<!-- ATTLIST geobox
westbc CDATA # REQUIRED
eastbc CDATA # REQUIRED
northbc CDATA # REQUIRED
southbc CDATA # REQUIRED> .....
```

可以看出, 定义了 DTD 之后, 可以用 XML 准确无误地表达 UML 所描述的元数据。如果用户需要扩展元数据, 遵守元数据标准中的扩展性规则, 从系统提供的界面即可以自定义 DTD, 从而达到元数据扩展的需求。

#### (3) XML 元数据的存储和查询

由于 XML 表达的元数据是文本文件, 它的存储如果只靠文件系统, 那么查询效率将会低下。但是目前还没有 XML 数据库。在这里可以采用这样的策略: 将 XML 文件存入数据库的二进制大对象 (BLOB) 中。这样做存储非常简单, 但是对于查询来说, 就损失了一部分效率。只能采取基于关键词的全文检索。现在的主流数据库系统 SQL Server 2000 和 Oracle 9i 都能较好地支持文件全文检索 (full-text index), 而且都可以在 BLOB 字段上建索引。

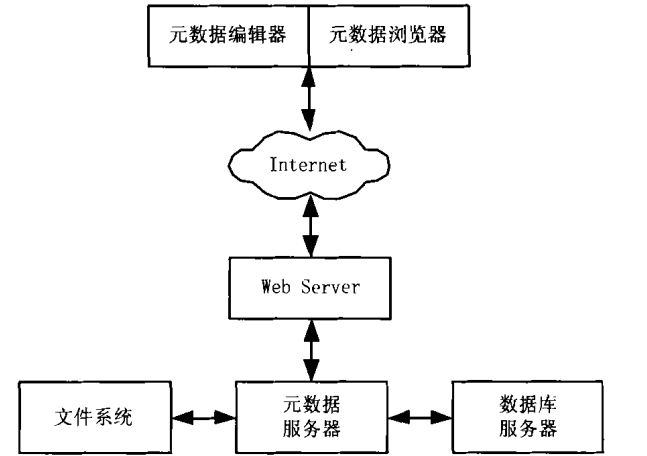


图 2 国家资源与环境空间数据库元数据管理系统  
Fig.2 Metadata management system of national resource and environment spatial database

### 3.3 元数据管理系统

由于国家资源与环境空间数据库项目的需求, 元数据服务器将集中在中国科学院地理科学与资源研究所, 同时数据也集中于此。因此, 在整个元数据

管理体系上,没有采取国际上通用的空间数据交换中心(Clearinhouse)方式,而是采取集中的管理模式。节点用户通过 web 服务器统一在元数据服务器注册各自的元数据信息,如图 2 所示。

## 4 结束语

本元数据管理系统是为实现国家资源与环境空间数据库共享而建立的。在元数据标准的设计上采用了与 ISO/TC211 保持同步的 UML,在存储和传输上使用了跨平台、可扩展、可自描述的 XML,为信息共享提供了有力支持。但是,还有一些工作有待进一步完成:

(1) 元数据 XML 文件目前是存储在数据库的 BLOB 字段中,检索效率不高,方式不灵活。如何实现高效率的存储和检索,需要进一步研究有关文献,做出一些软件编制方面的尝试;

(2) 元数据管理系统目前是集中式的,以后应该向分布式发展。需要参考空间数据交换中心的成功经验,以及研究 Peer-to-Peer、Agent 等先进软件的特点,使元数据管理更加灵活,方便快捷。

## 参考文献

- [1] ISO/TC211. Final text of CD 19115 geographic information - metadata. 2001.
- [2] 何建邦, 蒋景瞳, 刘若梅. 地理信息标准化研究与思考. 地理信息世界, 1998(2): 8 ~ 12.
- [3] 姚艳敏, 姜作勤, 严泰来. 国土资源信息核心元数据的研究. 测绘学报, 2001, 30(4), 349 ~ 354.
- [4] 戴超凡, 刘青宝, 邓苏. OIM XML 编码研究. 计算机工程与应用, 2001(3), 7 ~ 9.
- [5] Extensible markup language (XML) 1.0 (Second edition), <http://www.w3c.org>.
- [6] Metadata tools for geospatial data, <http://badger.state.wi.us>.
- [7] 汪小林, 罗英伟, 丛升日. 空间元数据研究及应用. 计算机研究与发展, 2001, 38(3): 321 ~ 327.
- [8] 蒋景瞳. 中国地理信息元数据标准研究. 北京: 科学出版社, 1999, 1 ~ 116.
- [9] 苏理宏, 黄裕霞. 资源和环境信息系统元数据的组织和应用. 中国图象图形学报, 2001.
- [10] How to set up a clearinghouse node, <http://www.fgdc.gov>.

# The Research of Metadata Management of National Resource and Environment Spatial Database

CAO Yanrong<sup>1</sup> WU Hongqiao<sup>1</sup> BI Jiantao<sup>1</sup> HUANG Yuxia<sup>2</sup> HE Jianbang<sup>1</sup>

(1 Institute of Geography Science and Natural Resources Research, Beijing 100101)

(2 Institute of remote sensing application, C.A.S. Beijing 100101)

**Abstract:** The paper discusses how to establish metadata management system of National Resource and Environment Spatial Database. It gives the designing method of Unified Modeling Language (UML) static diagram for metadata framework designing and data dictionary for detail element definition. Using Extensible Markup Language (XML) to describe metadata standard, it carries out metadata management and metadata expanding. It can help user to query effectively and find resource quickly in the database.

**Keywords:** Metadata standard; Metadata management; UML; XML