科学数据共享中的元数据技术研究

王国复[1] 涂 勇[2] 王卷乐[3] 徐 枫[4]
(1. 国家气象信息中心,北京 100081;2. 中国科学技术信息研究所,北京 100038;
3. 中国科学院地理科学与资源研究所,北京 100101;4. 国家信息中心,北京 100045)

摘 要:本文结合元数据技术的介绍,对元数据在科学数据共享平台的作用、元数据的分类、元数据的管理、元数据系统和元数据的应用模型进行了阐述,试图为科学数据共享平台的设计和建设提出一套元数据技术应用的方案,以期对地球科学领域的信息系统建设提供参考。

关键词:元数据;数据资源;共享平台;数据发现
中图分类号:P208 **文献标识码**:A DOI: 10.3772/j.issn.1674－1544.2008.01.006

Application of Metadata Technology in Scientific Data Sharing Service

Wang Guofu[1], Tu Yong[2], Wang Juanle[3], Xu Feng[4]
(1. National Meteorological Centre, Beijing 100081;
2. Institute of Scientific & Technical Information of China, Beijing 100038;
3. Institute of Geographic Science and Natural Resources Research, CAS, Beijing 100101;
4. National Information Centre, Beijing 100045)

Abstract: We made a summary of the application of metadata technology in scientific data sharing, including function, classification and management of metadata, framework of metadata system, and metadata application model. We tried to give a schematic map of metadata technology's application in scientific data sharing service platform.

Keywords: metadata, data resources, sharing system, data discovery

随着信息技术的发展和数据资源共享步伐的加快,用户对各类数据资源的认识逐渐深入,而且对数据使用的要求也越来越高。元数据作为数据资源存放、管理和应用的重要手段已经成为当今信息技术领域的研究热点之一。

1 元数据的概念及作用

1.1 元数据的概念

元数据最经典的定义是:“关于数据的数

第一作者简介:王国复(1970－),男,河南扶沟人,高级工程师,研究方向是科学数据共享平台、元数据和数据管理等。
基金项目:国家科技基础条件平台建设项目“气象科学数据共享中心建设”(2005DKA31700)。
收稿日期:2007年9月15日。

据”。但这一定义往往不能清楚描述元数据是什么。许多专家学者[1-4]在不同时期给出了元数据的不同描述，总体可以归为两类：一类为元数据是对数据的描述信息，一类为元数据是数据应用系统的辅助信息。这可以看作是从两种角度对元数据进行定义：一是从提高数据的使用价值角度,一是从数据的应用和服务角度。其实,这两种关于元数据的表述在我国科学数据共享平台建设中都得到了体现。

我们认为，元数据是对数据资源的规范化描述，它是按照一定标准（即元数据标准），从数据资源中抽取出相应的特征属性，组成的一个特征元素集合(即元数据元素)。这种规范化描述必须准确和完备地说明数据资源的各项特征。

通过元数据，一方面能够对数据资源进行详细、深入的了解，包括数据资源的格式、质量、处理方法和获取方法等细节；另一方面能够实现网络共享，使得数据资源的用户可以迅速地发现与其需求相匹配的数据资源，进而通过网络或其他途径获得这些数据资源并加以利用，从而促进数据资源的共享。所以元数据在科学数据共享中能发挥极其重要的作用。

1.2 元数据的作用

元数据不仅起到数据描述的作用，而且起到管理数据的作用。随着网络技术的发展和数字化资源的猛增，元数据在数据共享、资源发现以及知识管理方面的作用越来越明显，越来越为人们所重视[5]。可以说，目前元数据已经从简单的描述或索引发展成为用于管理数据、发现数据、使用数据的一种重要的工具与手段。在科学数据共享平台中,元数据将为各种形态的科学数据(集)提供规范、灵活的描述方法和检索工具；元数据为分布的、由多种数字化资源有机构成的信息体系提供整合的工具与纽带。脱离元数据的各种数据信息将是一盘散沙，将无法提供有效的检索和处理。

当前，不同领域的数据种类不断增加，新的数据格式不断涌现，数据量呈现几何增长的趋势,从而极大地扩展了元数据的功能。例如,元数据是数据的抽象，这一抽象在一定程度上屏蔽了数据在格式以及其他实现方面的差异，为基于内涵的数据共享和互操作提供了基础；在绝大多数情况下，作为数据描述信息的元数据比数据资源本身小，并且是格式化的，这为组织获取海量数据资源的描述信息，开发发现并定位信息资源的发现技术提供了条件和基础。概要而言，元数据在科学数据共享平台中起到联结数据生产者（如各级信息或数据中心）、使用者（如部门业务单位、科研单位、科学家和公众）和管理者（如中国气象局、国家测绘局、水利部等政府部门）的纽带作用。具体来说有以下几个方面的作用：

1.2.1 数据描述作用

数据生产者可以利用元数据对他们的数据集进行详细的说明，帮助用户了解数据的基本特征，从而判断他们是否使用该数据并能否有效地应用数据。同时，为用户提供数据转换方面的信息。使用户在获取科学数据的同时便可以得到相关说明信息（元数据），并通过元数据，接受并理解数据资源，最终与自己的科学信息集成在一起,进行不同方面的科学分析和决策。

1.2.2 数据发现作用

帮助数据使用者查询所需资源信息。比如，它可以按照不同的地理区间、内容属性以及具体的时间段来查找气象数据集。这使得数据发现、检索和重复使用变得更加容易。用户能更好地通过网络准确地识别、定位和访问数据资源。

1.2.3 数据管理作用

通过特定的元数据接口，可以便于对元数据和数据集实体进行统一管理。例如，元数据与数据集的同步更新、异地访问以及元数据安全、用户权限等设置。

1.2.4 目录交换作用

科学数据共享平台中包含不同类型、存储在不同地域的多种数据资源，通过数据目录、元数据和元数据目录服务系统，用户便可以很容易地发现它们,并可以共享数据集、维护数据结果,并对它们进行优化等。

1.2.5 资源整合作用

数据资源整合就是通过元数据对分布式的数据资源进行标准化、结构化，形成一个通过目录交换体系发现并获取的虚拟数据资源整体。这个过程“把无序信息变成有序的信息”，使用户可以透明地了解数据。即，不需要深入到各个物理数据存储单元，只通过门户目录系统就可以发现和使用这些数据资源，这样把科学数据共享平台建成一个共建、共用的“大平台”。

1.2.6 知识产权保护作用

元数据可以确保数据资源得到有效的保护，进而保护国家的投资。在各类数据集开发完成后，往往随着单位人员的变换以及时间的流逝，后期接替该工作的人员对先前的数据了解甚少或一无所知，对先前数据的可靠性产生置疑。而通过元数据内容，则可以充分描述数据集的详细情况。同样，当用户使用数据引起矛盾时，数据提供单位也可以利用元数据维护其利益。

2　元数据的分类与设计原则

科学数据共享平台的数据库系统存储着海量的数据资源信息。而面对这些数据，要对用户提供方便、快捷的服务。建设的各类科学数据共享平台正是通过元数据信息描述和管理这些数据资源，并且依靠元数据实现各类数据的检索应用。所以对这元数据进行科学的分类和设计具有非常重要的意义。

2.1　元数据的功能分类

从存储形式来说，元数据可以分为物理型元数据（physical file metadata）和逻辑型元数据（logical file metadata）。物理型元数据主要是准确地描述数据存储的结构，以及数据资源存储系统的特征，特别是分布式共享系统的各节点信息和元数据库信息等。逻辑型元数据除包括经典的描述型元数据，还包括对数据属性、数据格式、数据更新，以及数据管理和应用服务等元数据信息。

在科学数据共享平台设计与建设中，通常按功能把元数据划分为三大类：描述型元数据、管理型元数据和应用型元数据。

（1）描述型元数据：对数据库中所存储的数据资源的内容、属性的描述。这是元数据最基本的功能特点。这类元数据可以完整地反映数据库中所有信息对象的全貌，帮助用户了解具体的信息对象。

（2）管理型元数据：记录数据库中信息的存储地方、存储方式、操作信息（如更新、备份等管理操作）。

（3）应用型元数据：为用户提供分层检索方式，帮助用户方便、快捷地获取其所需要的资料。

描述型元数据信息是在数据生产或处理过程中形成的，且按照相应的元数据标准对数据集进行描述形成元数据文件（如 XML 元数据文档）。管理型元数据信息往往是在数据库建设过程中形成的，如存储管理信息和操作信息等。应用型元数据信息是根据科学数据共享平台开发和管理的需要形成的，如用户信息、数据检索应用信息等。

2.2　元数据设计原则

元数据是在数据资源管理者和用户之间实现良好互操作性的基础之一，没有统一、全面的规划和考虑，则对今后工作的开展造成很大的障碍。在进行元数据设计时，建议遵循以下几个原则：

（1）元数据描述的单元是各类数据集：一个数据集在数据库中对应一个数据表或一个存储目录。由于科学数据共享平台数据库系统中数据资源的存储、管理和服务的单元是数据集，元数据在设计时也是以数据集为单位进行的。

（2）检索应用有关的元数据单独存储：与用户检索应用有关的元数据与其他管理和应用（如建库、数据结构等）所涉及的元数据分表存储，以提高检索应用的效率。

（3）最全面地反映用户需求：元数据用以向用户更好地揭示各类数据资源，因此用户需求应作为衡量元数据设计是否合理的标准之一。

（4）通用性、标准性、完备性和可扩展性原

则：因为科学数据的数据种类非常多，各类资料的应用方式多有差异，所以通用性、标准性、完备性和可扩展性应该成为所有设计原则中最重要的内容。通用性是针对各类数据资源的特点进行分析，提出其共性的东西。标准性是指在元数据设计过程中，要体现或总结出一套标准，这些标准在信息管理中发挥很好的指导作用。完备性是指对所有需要的管理信息都有全面的描述或表示。可扩展性是指为了适应数据库管理资源种类的不断增加，为了更好地描述某些新的特殊种类信息，原来的设计方案需要有较好的扩充能力。

3　元数据的管理

3.1　元数据管理要解决的基本问题

作为对信息资源的描述，元数据只有完整地包含其描述对象的各种特征信息，并且其内容和组织方式需要遵循一定的规范，人们才能借助元数据正确地了解其所描述的对象，进而促进信息资源或产品的共享或交换，或者通过元数据实现流程有效控制[5]。元数据的完整性和规范性需要通过对元数据的有效管理加以保证，这就要求元数据管理系统必须能够适应元数据的应用目的和特点，在具备一般信息管理系统的共同功能之外，还应着重解决以下几个方面的问题：

（1）充分支持元数据内容的标准。元数据的目的是在不同人或系统之间共享有关信息资源、产品的说明信息，从而间接地实现对信息资源或产品的管理和交换。达到这一目标的前提是不同的人或系统能够一致地理解元数据中的各项信息，因而要求元数据在其使用范围内必须在内容组织和语意上遵循一定的规范。相应地，元数据管理系统必须对用户所采用的元数据内容标准给予有效地支持。根据需要，在元数据内容标准中，有时要定义描述元素之间的约束关系，例如描述元素之间的互斥、互为前提，乃至元素值之间相互限制。元数据管理系统必须能够正确地处理这些逻辑关系，严格按照标准规范对元数据的各种处理，以便正确地规范元数据的采集和维护工作。此外，不同领域的元数据内容标准必然有所不同，同一领域的标准也会随着应用需求的改变而发生变化。元数据管理系统必须具备足够的标准适应能力，以便用户能够及时根据需求的变化进行必要的调整。

（2）高效的元数据网络检索。元数据的主要作用是借助计算机网络交换信息，使人们能够通过元数据及时、准确地了解他们所需的产品（数字化或非数字化）。因此，元数据管理系统必须提供元数据的网络查询检索功能。元数据的网络查询既不同于关系型数据检索，也不同于一般网络搜索引擎常用的全文检索。元数据是非结构化的，而关系型数据的索引机制不能很好地适应元数据的不稳定结构。此外，元数据在信息组织上又存在数据域（描述元素）的划分，采用全文检索机制则不利于通过数据域的区分来减小查询命中范围。因此，元数据管理系统需要采用与元数据特点相适应的新的检索机制，以提高元数据查询的整体效率。

（3）标准的网络搜索协议。不同部门之间在元数据共享方面的合作要求各部门的元数据管理系统之间必须能够互联，并实现元数据的网络交换。为实现这一目标，元数据管理系统的网络查询服务必须遵循一种通用的协议实现对元数据网络搜索的提取。目前，在网络信息搜索和提取方面最重要的协议是 Z39.50 协议。该协议由 ISO 建立，用于规范网络信息搜索和提取过程中的各种请求与响应，并对服务器和客户机的处理进行规范。

（4）元数据的维护和互操作。元数据可以用于许多方面，包括数据文档建立、数据发布、数据浏览、数据转换等。元数据对于促进数据的管理、使用和共享均有重要的作用，应当加强并加快元数据的研究。随着信息种类和数据量的迅速增长，元数据建立和维护的难度越来越大，需要数据生产者更多的努力，同时需要那些随后可能应用数据的用户，或可能修改数据以便符合其需求的用户做出相应的努力。

由于元数据的种类复杂且用途殊异，将来多种元数据共存共荣的局面已成为共识，而元数据的互操作性要求在由不同的组织制定与管理且技术规范不尽相同的元数据环境下，能够作到对

用户保持一致性的服务，也就是说对一个应用或用户来说，能够保证一个统一的数据界面，保证一致性与对用户的透明。目前，元数据的重复使用和各种元数据的相互交换已成为元数据发展的趋势。

3.2　元数据管理系统的结构

元数据管理系统结构主要由元数据网关、元数据服务器和元数据库组成(图 1)。其中，各分节点安装元数据服务器，用于提供该节点信息中心元数据信息的发布，并按照统一的元数据标准建设元数据库；主节点部署安装元数据服务系统网关软件，用于连接各分节点元数据服务器，实现元数据和数据的全网发布[6]。

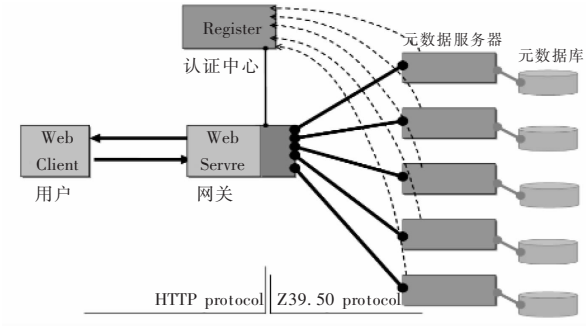

图 1　元数据管理系统结构

元数据网关是支持元数据服务的中心枢纽。一般地，其功能包括服务器代理功能：可以有效地避免远程客户对元数据库直接存取，屏蔽非法入侵，保证用户数据安全。服务器注册管理功能：对于加入到元数据共享系统的服务器，需要对其服务器名称、地址等进行注册登记，使其连接到元数据共享系统中。网络客户管理功能：提供用户注册、数据库访问权限管理等网络客户管理功能，便于网络客户权限的控制。

元数据服务器用于发布元数据。各元数据服务器一方面通过申请注册，把本节点元数据信息纳入到整个系统中；另一方面接受 Web 服务器对本节点的元数据和数据搜索指令。这样用户通过该系统就可以透明访问任一节点上的元数据和数据信息。

元数据库是元数据发布系统的核心内容，元数据的采集可以利用元数据编辑器通过手工方式进行采集，也可以进行自动采集，但都要按照统一的元数据标准进行处理。

3.3　元数据的发布流程

元数据发布过程分 4 个步骤(图 2)：

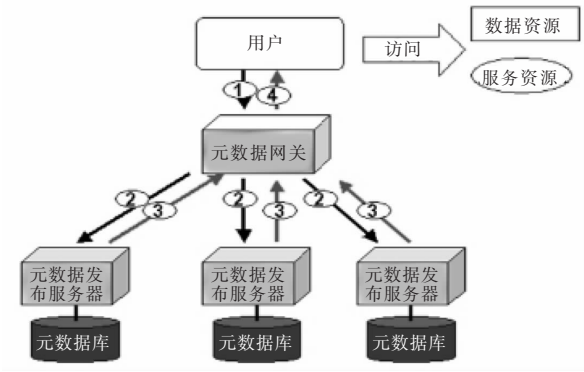

图 2　元数据发布流程示意图

(1) 用户向元数据网关发送元数据查询指令。

(2) 元数据网关将用户的指令发布到各节点的元数据发布服务器。

(3) 各节点元数据发布服务器搜索本地的元数据库，并将结果返回到元数据网关。

(4) 元数据网关将查询到的元数据记录进行综合后，返回给用户。用户对检索到的元数据进行评估，以决定是否访问该信息资源，经认证后获取数据服务。

3.4　元数据的更新与维护

在科学数据共享平台中，各元数据可以采用数据库表的形式以及 XML、LDAP 方式等进行存储和管理。元数据的更新及一致性维护是元数据管理的核心。元数据的更新和一致性维护采用人工更新与维护、应用程序自动更新与维护、数据库触发机制 3 种方式。

对于大多数描述型和应用型元数据信息较适合于采用人工更新与维护的方法，人工更新与维护必须遵守相应的元数据标准。描述型和应用型元数据信息最主要、最常见的修改操作是对某些不太准确的描述内容或应用方式进行一定的修正，需要人为的判断和思考，而且这种更新一般是不定期的，因此人工的方式可以对其进行很好的维护和更新，人工对元数据的管理和维护可以通过研制界面友好的人机交互元数据管理系

统来实现。

大多数管理型元数据信息是在应用程序的运行中自动生成和更新的。管理型元数据一般是关于数据库的存储管理方面的，在应用程序运行时，如每天对新的数据进行追加的程序运行时，就可以随着程序的运行自动地对相关的管理型元数据信息进行一致性的更新。

一些元数据信息（如数据表描述信息）还可以通过预制的数据库系统存储过程等实现元数据记录的更新与维护，也是一种自动维护方式。

4 元数据的应用

4.1 元数据技术的应用

科学数据共享平台可以借助于元数据强大的管理功能，实现数据的有效管理和各种形式的服务。以元数据为核心的管理模式是科学数据共享平台的技术特征[7]，并使科学数据共享平台成为一个可查询的信息目录，为查询和管理用户实施信息查询、数据发布和管理等提供一个高效的工具，这个可查询的信息目录包含的不是科学数据本身，而是关于数据的信息，即元数据。

显然，以元数据技术开发建立的科学数据共享平台，是在现有网络、数据库、存储、目录信息发布、身份认证等信息技术的支持下，对数据资源进行统一管理，实现目录服务、数据服务、功能服务等功能的综合性业务系统。科学数据共享平台的功能体系如图 3 所示。

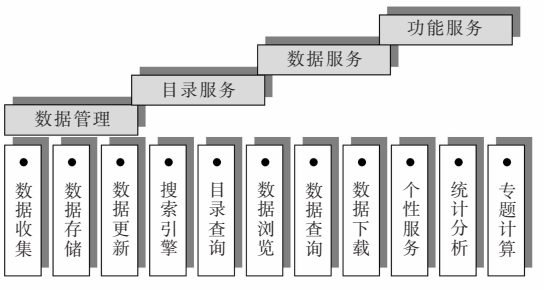

图 3 科学数据共享平台的功能体系

这些功能，从技术层面上更多地体现在以元数据为核心的分布式数据管理和服务技术，尤以元数据目录服务（MCS，Metadata Catalog Service）为代表。

以元数据为核心的目录查询，是元数据系统利用元数据技术提供信息服务的一种标准模式。它通过元数据标准的核心元素将信息以动态分类的形式展现给用户。用户通过浏览以数据库系统为核心建立的科学数据共享平台所提供的元数据摘要（核心元数据）可以快速确定自己所需的数据范围，然后通过共享平台在该范围内进一步搜索，进而定位数据资源，并通过查询、下载、个性化服务等系统功能获得数据。

4.2 元数据在数据发现中的应用

在科学数据共享平台中，用户从登录系统到下载数据，经历了数据发现（data discovery）和数据访问（data access）两个阶段。其中，数据发现的实现主要是由元数据系统来完成的，而数据访问是由各数据中心建立的网站系统来实现的。

4.2.1 数据发现

数据发现是用户通过元数据评估并定位共享数据资源的过程。

具体来说，数据发现是通过共享平台主节点所提供的元数据搜索功能来实现的，主要包括如下过程（图 4）：

（1）用户访问主节点网站元数据搜索页面，将请求发送到 Web 服务器。

（2）Web 服务器对请求中所指定的元数据检索条件进行处理，构造 Z39.50 协议消息，并将其发送至元数据网关。

（3）网关将根据协议消息中所指定的检索条件确定节点检索范围，并发送到特定分节点元数据节点服务器。

（4）节点服务器将访问本地的元数据库（以 XML 格式存储的），执行 XML 查询命令，产生检索结果。

（5）元数据网关汇总、合并来自所有被检索节点服务器的检索结果。

（6）应用服务器将检索结果以网页形式返回用户浏览器。

（7）用户对检索结果进行浏览，并对结果所描述的共享数据资源的内容、质量、格式、大小等进

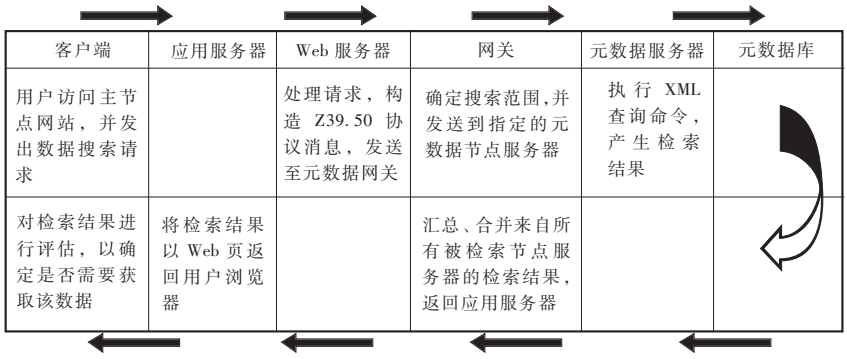

图 4　数据发现流程图

行评估，以确定是否需要获取该资源。

4.2.2 数据访问

数据访问是指在获得共享资源定位信息（URL）后，对共享资源的访问、下载、在线操作等。对共享数据资源的访问是通过分节点网站系统所提供的数据检索应用系统实现的。

分节点网站系统接收用户数据访问请求。若该用户是全网合法用户，分节点 Web 服务器将对请求进行处理，如执行标准的 SQL 语句检索共享数据库，或者通过 HTTP 或 FTP 直接下载共享数据资源。

从纵向结构来说数据访问分 4 层：浏览器客户端→应用服务器→Web 服务器→数据资源（数据库或文件目录结构），如图 5 所示。

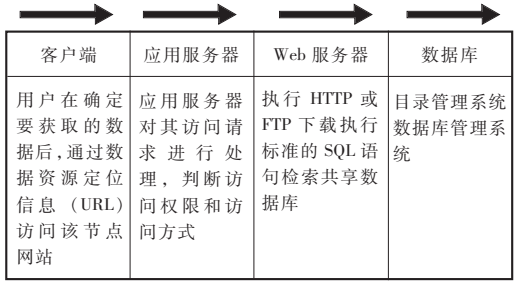

图 5　数据访问流程图

5　小　结

元数据对我国正在实施的科学数据共享中心（网）建设具有非常重要的意义，对信息技术的发展具有不可估量的作用。近年来，在科技部的大力推动下，一些行业首先通过采标、引标或自主研究等形式制定了不同领域的元数据标准，例如测绘、资源环境、气象、海洋、林业、水利等部门，这既是一项标准规范建设，又是开展共享和数据交换的基础。其后，通过联合研究，开发了元数据系统软件，并在气象科学数据共享服务网、地球系统科学数据共享网等共享平台上率先进行了部署，实现了分布式环境下的数据导航服务和元数据的统一发布。目前，气象科学数据共享服务网还开发和部署了气象元数据目录服务系统（Met. MCS），不但实现了元数据的统一发布和数据发现，而且实现了数据的统一目录服务（catalog service），即把元数据服务和数据服务统一起来。

可以看出，元数据技术在科学数据共享应用中经历了元数据标准、元数据系统和目录服务 3 个发展阶段，每个阶段有不同的研究内容，分别满足共享工作不断深化的需求，如数据交换、数据搜索和数据访问。随着科学数据共享工作的深入，用户也将对共享平台提出新的需求，例如共享服务的科学评价，数据的可视化服务等，需要研究利用包括元数据技术在内的信息技术来不断丰富科学数据共享的服务内容，为科技创新提供坚实的信息支撑。

参考文献

[1] Francis. P. Bretherton, Paul T. Singley. Metadata: a User's View[J]. Scientific and statistical Database Management, 1994, 30(28): 166 – 174.

[2] Consortium for International Earth Science Information Network. CIESIN Metadata Guidelines. http://www.ciesin.org/metadate/documentation/guidelines/, 1998

[3] Noushin Ashrafi. The Information Repository: A Tool for Metadata Management[J]. Journal of Database Management, 1995, 2(2): 183 – 190.

[4] 李军，周成虎．地球空间数据元数据标准初探[J]．地理科学进展，1998, 17(4): 55 – 63.

[5] 徐枫．元数据技术及其在科学数据共享中的应用．科学数据共享管理研究[J]．北京：中国科学技术出版社，2002: 178 – 196.

[6] 王国复，徐枫，吴增祥．气象元数据标准与信息发布技术研究[J]．应用气象学报，2005, 16(1): 114 – 121.

[7] 王卷乐，游松财，谢传节．元数据技术在地学数据共享网络中的应用探讨[J]．地理信息世界，2005(2): 36 – 40.