科学数据库元数据注册系统研究与实现*

曾　炜[1,2]，黎建辉[1,2]

(1. 中国科学院 计算机网络信息中心，北京 100080；2. 中国科学院 研究生院，北京 100049)

摘　要: 首先介绍了科学数据库中元数据管理和应用存在的问题,根据元数据特点给出了元数据分层模型,提出了一个满足科学数据库元数据标准规范建设需求的元数据登记系统设计方案,分析并解决了该登记系统中的关键性理论和技术难题,并在应用实践中进一步完善。

关键词: 元数据; XML Schema; 元数据注册; ISO11179

中图法分类号: TP311.13　　文献标识码: A　　文章编号: 1001-3695(2006)03-0073-03

Research and Implementation of Science Database Metadata Registry System

ZENG Wei[1,2], LI Jian-hui[1,2]

(1. *Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100080, *China*; 2. *Graduate School, Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: This paper first gives a full description of metadata management and application in Scientific Database. Then presents a layer-based model, finally proposes the design of Scientific Database metadata registry to meet the needs of metadata users. The design will undergo the examination in the future practice.

Key words: Metadata; XML Schema; Metadata Registry; ISO11179

中国科学院科学数据库经过近 20 年的发展,到目前共有专业数据库 300 多个,总数据量达 8.2TB。元数据理论和技术是实现数据标准化以及数据共享、交换和整合的主要手段。目前,中国科学院计算机网络信息中心已经制定完成了科学数据库元数据体系当中的《中国科学院科学数据库核心元数据标准》[7] 以及以它为基础的多个面向具体应用的扩展元数据标准。而各种类型的元数据标准常常缺少兼容性的要求,科学数据库中目前多种元数据标准的共存以及未来更多标准的出现,使得实现不同元数据标准相互兼容,进而按照不同元数据标准著录的数据之间能够相互访问和检索成为了目前亟待解决的问题。一般有两种方法来达到统一访问的目的:①定义一个统一的元数据标准;②建立各元数据到基本标准的映射关系。

采用一个公认的元数据标准(如 DC 或 MARC)会有好处,但这是不现实的,这样无法准确地描述各种类型的数据。一个所谓万能的标准,或者不能提供描述对象的所有特征或者就是极难构造,从而导致描述的不准确性。即使像 DC 这样的标准取得了部分成功,经验显示其数量有限的元素(实现互操作的关键)限制了很多元数据使用者。为了解决这个问题,更多的元数据元素(Element)已经加入到 DC 中以增强描述的能力,这种策略允许元数据更复杂地应用,但是同时威胁到了互操作性这个初始的目标。只建立并只采用一个单一的元数据标准是不现实的,其他不兼容的元数据是不可避免的而且会一直存在下去。因此我们相信任何一种元数据互操作问题的解决方案一定要处理互不兼容的元数据。

收稿日期: 2005-01-18; 修返日期: 2005-04-14

基金项目: 中国科学院知识创新工程重要方向项目(171711)

1　元数据分层模型

元数据就是关于数据的数据,但是这个定义过于宽泛。为了避免产生元数据术语上的混淆,下面将定义三类元数据术语[4]:

(1) 元数据元素(Metadata Element),代表一个抽象同时详细而精确的概念来描述数据,如 DC 中的 Creator 元素就定义了一个对于创造一个资源负主要责任的实体的概念。

(2) 元数据标准(Metadata Schema),代表一组任意但是特定的元素。一个元数据标准就是 DC,它是 15 个元数据元素的总称。

(3) 元数据标准的实例(Metadata Schema Instance),描述了依据某个元数据标准中元素的概念建立的一组特定的数据,如本文的元数据都是使用了 DC 元数据元素的实例。

从下到上具体分层模型如图 1 所示。

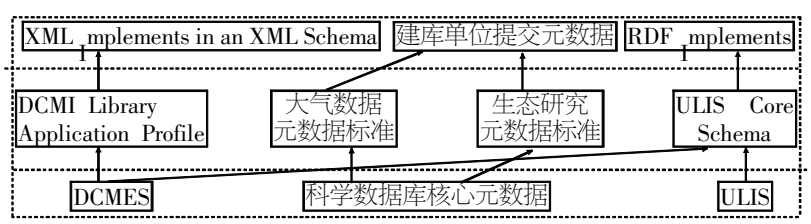

图 1　元数据分层模型

2　元数据描述

为了达到更广泛的元数据互操作的目标,最主要的要求在于描述和标志各种元数据标准及其相应的元素的能力。没有描述,一个标准就是一个随意的一组术语的集合,除了它们的建立者谁也不会知道它们被创造的目的;没有元数据标准

的标志,就不会存在一种机制去推演它的本质或者如何去使用它。

2.1 元数据标准描述

为了描述各种元数据标准,我们采用 ISO11179 第三部分。ISO/IEC11179 是数据元素的规范说明和标准化,是一个国际标准,是对用于数据库和文件中的数据元素进行描述的标准。它说明了数据元素组成的各个基本方面(包括元数据),当数据元素用于人机共享时,该标准适合确切描述数据元素的含义(Meaning)和标志(Representations)。现在的第三部分不仅仅是一个数据元素标准,更是一个元数据注册系统的标准。许多组织正在将这些标准作为元数据储存体的部分标准,如美国的环境数据注册系统(the Environment Data Registry)、美国的健康信息库(the United States Health Information Knowledgebase)、美国的人口普查团体的元数据库(the Census Bureau Corporate Metadata Repository)等。

ISO/IEC11179 定义了三类信息来描述被管理项(Administered Item,即登记对象):管理与标志信息(Administration and Identification)、命名与定义信息(Naming and Definition)和分类信息(Classification)[6],如图 2 所示。

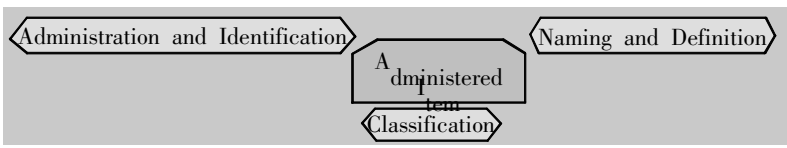

图 2 被管理项描述

为了便于构造元数据标准的描述,我们构造了 Schema 去规范元数据元素各种各样的属性,并且遵循 ISO11179 的规范。采用 XML Schema 建模元数据使得我们可以充分发挥 XML Schema 的优点:①XML Schema 基于 XML,没有专门的语法,可以像其他 XML 文件一样解析和处理;②XML Schema 还支持一系列的数据类型(Int, Float, Boolean, Date 等),并且提供可扩充的数据模型,简化元数据标准描述的编码过程,并且提供附加的完整性检查,同时还使独立产生准确的元数据标准定义的编码成为可能。

2.2 元数据标准的标志

在我们的方法中使用一个句柄(URI 命名空间)去唯一标志每个元数据标准以及它们的元素,用句柄作为元数据标志的优势在于它提供了一个简单的机制把每个元数据标准的标志和一组特定的资源指针(Pointer)联系起来,这些指针能用来定位一个特定元数据标准的描述和服务。因为元数据标准的句柄能够用作一个标志、描述或服务的指针,我们称句柄代表了元数据标准的类型。

2.3 元数据标准对象

因为要去寻找一个单一的元数据标准是不现实的,只能通过动态转换元数据的方法来达到互操作的目的。我们的方法集中在建立一个框架将元数据实例、标准以及服务映射成一级类的数据对象,这涉及到分类、标志和定义元数据,要求有一个能联系元数据和分布式服务的框架。

一个数字元数据对象是元数据模型在元数据标准层的对应对象,它可以描述自己、元数据标准和所拥有的属性。它提供一组服务可以将遵照它标准的元数据转换成一种或更多的符合异类元数据标准的元数据格式,而且能产生不同的元数据表示和编码方式。我们的框架允许新元数据和标准的动态引入以及作为分布式一级类的对象被访问。

3 可扩展的、可互操作的元数据注册系统

我们的互操作的元数据注册系统设计是建立在一个观点上的,即把元数据标准包装成一级类的网络对象,为操作元数据提供更强大的支持。下面将描述怎样利用数字对象体系结构来实现动态的、可扩展的、可互操作的元数据注册系统的设计。

3.1 基本设计原理

可互操作的元数据注册系统的实现是基于我们刚才实现的数字对象体系结构所具有的功能。数字对象体系结构非常适合用来实现元数据注册系统,主要基于以下原因:数字对象体系结构提供一种简单的机制来为资源指定一个唯一标识符,即通过把它们封装在一个唯一标志的数字对象中,在本文中即元数据标准对象。在我们的注册系统中,这种功能能把一个元数据标准和它对应的标准标识符绑定在一起。数字对象体系结构为对象的绑定提供一个统一的框架,即通过简单地把服务的标识符和某个数据对象的资源联系起来,将元数据注册服务和元数据标准绑定在一起。同时数字对象体系结构是可扩展的,它允许新的服务以一种动态的和分散的方式添加进原有的系统。这个特点让元数据注册系统能够吸收新的元数据和元数据标准。数字对象体系结构提供的内容类型公开了一些标准的接口,这些接口能够应用那些概念上相似但是内容上不同的元数据。

3.2 定义元数据标准的数字对象

将元数据标准的描述和服务封装在一个数字对象中就把元数据标准变成了一个一级类网络对象了。通过抽象化的格式和编码,这些得到的元数据标准数字对象提供了一个标准化的方法以访问它的元数据标准的定义[10]。一个新的元数据标准数字对象通过下面这些操作建立:

(1) 创建一个新的数字对象,这个对象的标识符使用元数据标准的标识符。

(2) 把已编码的元数据标准的描述如 Schema 储存在刚刚创建对象的一个数据元素中。

(3) 将元数据标准的标识符/数字对象的标识符注册进句柄系统,然后将它的值初始化为指向包含这个数字对象的元数据注册系统。任何人可以用句柄来标志和定位元数据标准的描述和相关资源。

(4) 添加标准内容类型到数字对象中去,然后将它和元数据标准的描述绑定在一起。这个标准内容类型定义了一组特定的分发请求以提供访问元数据标准的描述和服务。这个内容类型通过基于方法的接口将元数据标准的描述编码抽象化,并建立与核心元数据映射关系。

我们的元数据互操作框架被设计成提供如下两层不同的功能,而后面一个主要是为了元数据互操作考虑的。

(1) 通用元数据标准描述。它提供必要的功能来访问所有元数据标准的描述,同时抽象出一种特定的方法来描述和编码元数据标准。这类方法其中一个例子是 ListAllElements(),它返回元数据标准中所有元数据元素列表;GetElementDefini-

tion(ElementName) 返回一个特定的元数据元素描述。

(2) 特定元数据的转换。它提供从一个元数据标准到另一个的转换方法。Metadata_Schema 内容类型能够提高元数据转换的基本原理在于, 负责管理元数据标准的描述的组织最适合为他们自己的元数据标准建立转换方式。这类方法的其中一个例子是 ListConversion( ) , 它返回内容类型能够把元数据转换成元数据标准的标识符; 另一个例子是 Convert(TargetSchema, Metadata) , 它能把给定的元数据转换到用户指定的元数据标准中。

3.3　构造可互操作的元数据

通过把元数据和它的服务封装在数字对象里就能把元数据变成一个一级类的网络对象。通过抽象化标准的格式和编码方式, 有很多元数据数字对象提供的操作可以对元数据进行访问。正如刚才提到的那样, 元数据的内在问题就是它们有太多的种类, 而且都是为了不同的目的创建的。为了在数字对象中实现元数据描述的可互操作性, 就必须提供一个与元数据无关的接口, 同时还要保证客户能取得元数据的各个部分。为了解决这个问题, 我们采用了科学数据库核心元数据的内容类型, 每个标准与该类型各项建立一对一或者一对多的关系。而对于核心元数据不能完全覆盖的数据项之间的映射, 通过扩展映射表来完成。因此这种内容类型提供一种与具体元数据无关的方式来访问附属的元数据。这个功能可以分为两类: 通过手工指定或者配制文件, 而后者比较灵活, 相关元数据元素间的映射关系可以建立, 元数据标准之间的关系也随之建立。当两个标准间各个元数据元素关系建立完毕后, 它们之间的转换关系自然就建立起来, 元素关系分为一对一和一对多两类。我们的框架并不提供任何转换、标志和描述元数据的新方法, 我们也不提出全新的概念和方式来创造和映射元数据标准, 我们提出的网络体系结构只是通过一对一和一对多的映射能提高灵活性和提供可扩展性的服务。

3.4　实现架构

用户可以通过 WWW 的方式或者 Web Service 的方式访问系统, 包含元数据实例的数字对象就会产生符合编码要求格式的元数据实例。系统结构如图 3 所示。

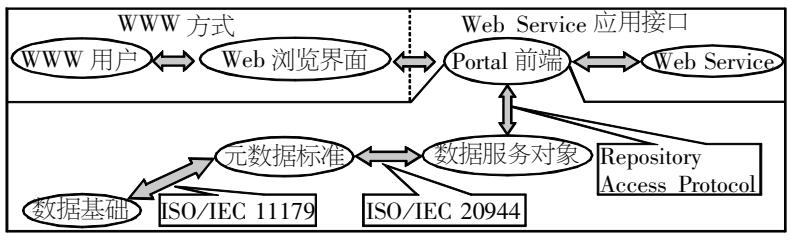

图 3　系统结构

元数据构造者也可以为了得到一个不同的表现形式选择一个自定义的标准对象的实现。注册系统在一个搜索数字对象(Search Digital Object) 中保存有一个所有元数据对象内容的反向索引表。通过新元数据对象的句柄作为参数, 这个元数据对象就被注册进刚才提到的搜索对象中, 搜索对象就会对这个元数据对象建立一个索引[11]。搜索对象还支持关键字检索, 一个包含关键字的元数据对象的句柄用网页或者服务列表形式返回, 添加一个新的元数据标准不需要管理员的任何额外工作。因为注册系统知道元数据标准的细节, 所以它还能提供元数据实例和它们的标准之间的对照索引。因此当一个用户看到一个元数据实例时, 他可以跟踪到实例的标准, 据此可以知道这个元数据字段更多的含义和上下文关系。

4　总结

元数据注册系统的基本实现说明了我们方法的可行性和灵活性。虽然我们试验的元数据标准和元数据集合比较小, 但是目前系统的实现应可以适当地加大规模, 并能处理新的元数据以及元数据标准。当需要新的元数据转换工具时, 不需要更新任何数字对象, 元数据标准转换模块能动态地添加进底层的构造中(Infrastructure) 。这个框架能为需要元数据移植的系统提供一个吸引人的解决方案, 此时, 要实现元数据标准内容类型要求有软件模型的开发。虽然这个方法很不错, 但最好还是能提供一个转换元数据标准的不需要额外编程工作的解决方案。实际上, 使用简单的等价关系或 XML 的表示能表达简单的元数据标准之间的映射, 使添加一个新的转换操作变得很简单。未来的工作就考虑只要简单地把一个 XML 文档附加到一个通用的元数据标准转换内容类型。

科学数据库元数据注册系统的建设是一个长期积累、不断完善、不断发展的过程。由于国际上元数据注册系统的相关技术和理论还处于研究摸索阶段, 并没有成熟、可借鉴的实例, 因此科学数据库元数据注册系统更应该以满足用户需求为最根本的原则, 不断进行探索和完善。例如, 将 RDF 技术应用于元数据注册系统就是一个目前虽未实现, 但很值得进一步研究的热点问题。

参考文献:

[1] ohn Cowan, Reuters. Extensible Markup Language (XML) 1.1 [EB/OL]. http://www.w3.org/TR/XML11/.

[2] Ora Lassila, Ralph R. Swick. Resource Description Framework (RDF) Model and Syntax Specification[EB/OL]. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/, 1999-02-22.

[3] Dave Beckett. RDF/XML Syntax Specification (Revised) [EB/OL]. http://www.w3.org/TR/rdf-syntax-grammar/, 2003-10-10.

[4] Blanchi C, Petrone J. Distributed Interoperable Metadata Registry[J/OL]. D-Lib Magazine, 2001, 7(12). http://www.dlib.org/dlib/december01/blanchi/12blanchi.html.

[5] Heery R, Wagner H. A Metadata Registry for the Semantic Web[J/OL]. D-Lib Magazine, 2002, 8(5). http://www.dlib.org/dlib/may02/wagner/05wagner.html.

[6] ISO/IEC 11179-3. Information Technology-Metadata Registries(MDR) ——Part3: Registry Metamodel and Basic Attributes[S]. 2003.

[7] 科学数据库核心元数据标准规范 v1.2[EB/OL]. http://md.sdb.ac.cn/md_paper/index.html.

[8] 科学数据库标准规范项目[EB/OL]. http://md.sdb.ac.cn.

[9] 科学数据库及其应用系统标准规范建设课题任务书[R]. 北京: 中国科学院计算机网络信息中心.

[10] 刘飞, 黎建辉, 阎保平. XML 和 RDF 在科学数据库元数据标准建设中的应用与展望[J]. 微电子学与计算机, 2004, (7): 128-132.

[11] 刘飞, 黎建辉, 阎保平. XML Schema 在科学数据库元数据互操作中的应用[J]. 计算机应用研究, 2005, 22(5): 199-201.

作者简介:

曾炜(1980-), 硕士研究生, 主要研究方向为数据库及其应用; 黎建辉(1973-), 硕士生导师, 主要研究方向为元数据技术与标准。