元数据技术在数据共享平台中的应用

熊建斌,李振坤,陈平华,刘怡俊,林瑞峰
(广东工业大学 计算机学院,广东 广州 510006)

摘 要:以科技厅数据共享规范与接口为标准,把元数据技术充分应用在数据共享平台中,构成一个安全、可靠、高效、稳定的信息交换渠道,为跨部门的信息共享和信息交换提供服务,促进信息资源的开发利用。

关键词:分布式数据;元数据;数据共享;信息畅通;信息交换
中图分类号:TP311 **文献标识码**:A

Design and implementation of metadata-based technology platform for data sharing

XIONG Jian Bin, LI Zhen Kun, CHEN Ping Hua, LIU Yi Jun, LIN Rui Feng
(Faculty of Computer, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Taking measure of science and technology agency data sharing standards and interface standards, the metadata technology is used in the data sharing platform, to constitute a safe, reliable, efficient, and stable channels of information exchange, which can cross-sectoral information sharing and exchange of information to provide services to promote the development and utilization of information resources.

Key words: distributed data; metadata; data sharing; information flow; information exchange

随着信息技术的不断发展以及人们对信息共享的迫切需求,元数据技术被应用于更多的领域。为了适应网络环境下信息资源共建共享的需求,元数据的研究成为一个热点。国外关于元数据研究已经很成熟,国内的研究正处于起步发展的过渡时期[1]。如何低代价、方便地将企业内部或企业间异构数据进行交换,实现大范围的跨企业实体的商务应用系统的对接,是当前互联网环境下每个企业发展所面临的一个大问题。由于系统的开发语言、运行平台和通信协议不同,对外数据交换的数据格式也存在很大的差异,因此如何解决语言差异、平台差异、协议差异和数据差异所造成的高代价的系统集成和信息资源共享成为问题的关键。目前大多数数据交换系统仍使用传统方式,显而易见这种设计缺乏通用性和扩展性。在数据共享上无疑是繁杂低效的,而且不可避免地会产生许多漏洞,不利于数据的安全。建立一个通用的、可扩展性的数据交换系统,对这些异构系统进行有效的信息集成已是当务之急。

1 元数据技术

1.1 元数据定义

元数据是关于数据的组织、数据域以及关系的信息,也就是“关于数据的数据”[2]。

1.2 元数据标准

元数据标准是经过标准化组织认可的元数据方案。在不同的科学数据共享领域中,都会有各自的元数据标准。为了便于实现数据的定位、共享、减少重复以及促进其合理使用,1994 年,美国联邦地球空间数据委员会便开始了元数据的研究,并制定了一种以元数据为核心的标准。

英国 Dublin 元数据核心元素标准适用于各种网络资源。它定义了 65 个元数据,包括 15 个 DC 核心元数据、26 个限定元数据、21 个编码体系元数据和 3 个其他元数据。该标准按照信息的类型和范围将 15 个核心元素分为 3 个子集:数据资源内容、数据知识产权和数据实体。Dublin 元数据的每一个核心元素都是可选的和可以重复使用的[3]。

1.3 分布式元数据的组织管理

科技管理元数据[4]可分为3个层次:元数据元素、元数据实体和元数据子集。元数据元素是元数据最基本的信息单元;元数据实体是同类元数据元素的集合;元数据子集是相互关联的元数据实体和元素的集合。在同一个元数据子集中,实体可以有2类,即简单实体和复合实体。简单实体只包含元素,复合实体既包含简单实体又包含元素,同时复合实体与简单实体及构成这2种实体的元素之间具有继承关系。科技管理元数据内容如图1所示。

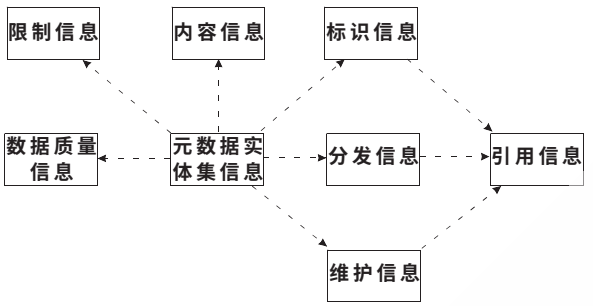

图1 元数据内容

元数据实体集信息包含必选的和可选的元数据实体和元数据元素信息,是标识信息、内容信息、分发信息、数据质量信息、限制信息、维护信息、引用信息的聚集。标识信息包含唯一标识数据的信息,包括有关资源的引用,数据集摘要、目的、可信度、状态和联系办法以及数据集维护信息等实体信息;内容信息提供数据内容特征的描述信息,是必选的,其"资源域"属性用于表明数据集所在的资源范围;分发信息包含有关资源分发者的信息以及用户获取资源的途径;数据质量信息包含数据集质量的评价信息;限制信息包含访问和使用资源的限制信息;维护信息包含有关资源的更新频率及更新范围的信息,如引用、负责方、地址、联系信息、日期等。

2 科技管理数据共享平台设计与实现

2.1 技术体系分析

数据共享平台采用的核心技术是Web Services技术、XML技术,J2EE技术及中间件技术。采用J2EE体系架构,充分运用Web Services的应用技术和XML的数据交换技术,设计开发功能强大、可扩展性好的数据共享和交换平台,以及基于Browser/AppServer/DBServer三层架构的数据交换体系,三层的技术架构图如图2。

表示层	..	Browser
应用层	..	AppServer
持久层	..	DBServer

图2 三层的技术架构图

(1)表示层主要负责:提供发布和搜索信息的门户网页界面;提供一个Controller,委派调用业务逻辑和其他上层处理;处理异常,抛给Struts Action*为显示提供一个模型;UI验证。

(2)持久层主要负责:用于执行数据的CREATE、RETRIVE、UPDATE、DELETE等操作;用于管理数据库连接池,增强数据库性能;为将来数据库迁移做准备(一般持久层支持大多数数据库,并且迁移时改动特别小)。

(3)应用层主要负责:处理发布和搜索服务的请求,即利用Web Service和中间件技术处理这些请求;提供与表示层及持久层交互的接口;管理业务层级别的对象依赖;在显示层和持久层之间增加了一个灵活机制,使得它们不直接联系在一起;管理程序的执行。

2.2 数据共享平台中的元数据

元数据分布于数据共享平台所连接的各共享节点上,元数据管理系统对不同层次、地域分布的众多节点的元数据进行统一组织、管理,集成在统一的平台框架内,为用户提供全局数据导航和获取接口,实现特征级数据元转换[5]。元数据管理系统部署在平台的各节点上,是一个分布式的信息管理软件,由元数据网关、元数据服务器和元数据库组成,如图3所示。

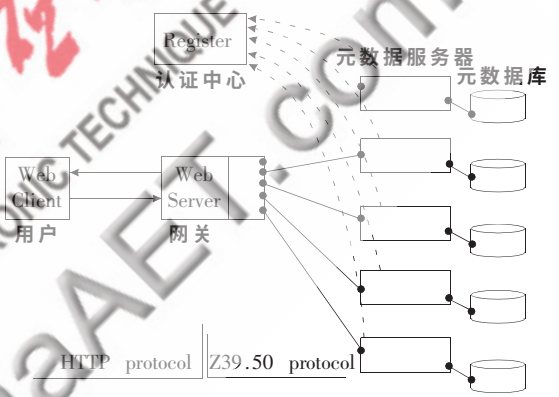

图3 元数据管理系统构成

元数据网关是支持元数据服务的中心枢纽,具有服务器代理、注册管理、网络客户管理等功能。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平台中,另一方面又接收Web服务器对本节点的元数据和数据搜索指令,这样,用户通过平台就可以透明访问任一节点上的元数据和数据信息。元数据库是元数据信息管理系统的核心内容,各种元数据信息按照统一的元数据标准进行处理,利用元数据编辑器或其他自动方式上载到元数据库中。

2.3 元数据共享平台的总体框架

在统一的元数据交换平台上构建的一站式数据交换和共享服务整体框架,本平台可以将现有的政府部门的信息系统联系起来,以统一的门户协同为各级政府及政府各部门提供服务,实现数据交换和共享服务的集中式协调调度和分布式管理运作,采用常用的多层分布式J2EE软件架构,应用Web Services和中间件技术来搭建这个平台,平台采用B/S模式。数据共享平台的软件架构设计如图4所示。

在数据共享平台中,各个应用主体都是独立的,包含诸多功能的系统,主体内部功能之间、主体之间都存

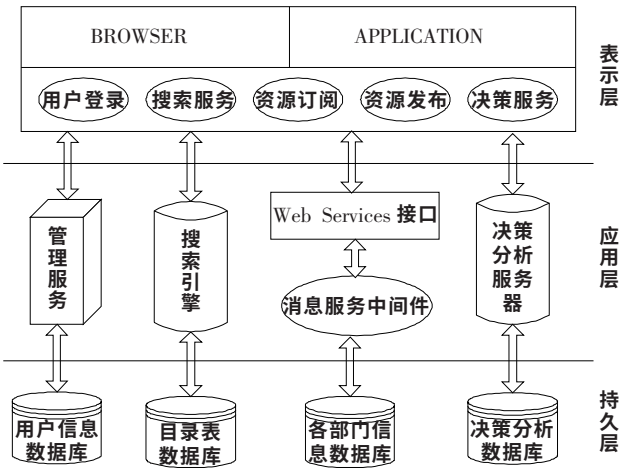

图 4 元数据共享平台的软件架构

在复杂的相互联系,因此在总体设计中采用数据交换中心 DEC(Data Exchange Center)和应用主体节点的前置机处理系统 FPS(Front-end Processing System)的结构来简化这些关系,并在应用主体上为应用主体提供相应的服务,提供一致的访问行为和接口。

2.4 技术方案描述

2.4.1 J2EE 架构

本系统采用 J2EE 架构实现应用体系结构,系统设计采用基于 J2EE 的技术,完全采用 MVC+DAO(Model+View+Control+DAO)应用设计模式,使得层之间相对松散耦合,具有良好的扩展性和稳定性,应用设计结构如图 5 所示。

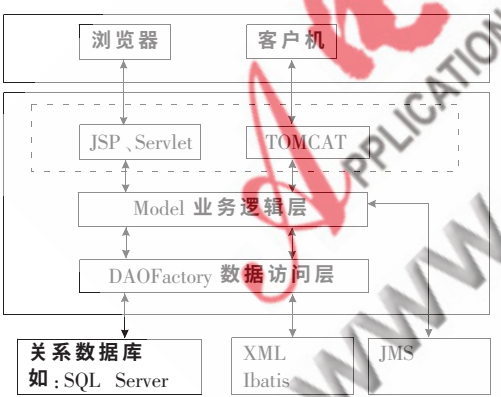

图 5 应用设计结构

2.4.2 IBATIS 架构

IBATIS 是以 SQL 为中心的持久化层框架,能支持懒加载、关联查询、继承等特性。IBATIS 不同于一般的 OR 映射框架。OR 映射框架是将数据库表、字段等映射到类、属性,这是一种元数据(meta-data)映射;IBATIS 则是将 SQL 查询的参数和结果集映射到类。具体来说,IBATIS 做的是 SQL Mapping 的工作,它把 SQL 语句看成输入以及输出,结果集就是输出,而 where 后面的条件参数则是输入;IBATIS 能将输入的普通 POJO 对象、Map、XML 等映射到 SQL 的条件参数上,同时也可以将查询结果映射到普通 POJO 对象(集合)、Map、XML 中。

2.4.3 XML 与 Web Service

可扩展标记语言 XML 是 Web 上表示结构化信息的一种标准文本格式,它没有复杂的语法和包罗万象的数据定义。XML 同 HTML 一样,都来自 SGML(标准通用标记语言)。SGML 是一种在 Web 发明之前就早已存在的用标记来描述文档资料的通用语言。但 SGML 十分庞大且难于学习和使用,鉴于此,人们提出了 HTML 语言。但近年来,随着 Web 应用的不断深入,HTML 在需求广泛的应用中已显得捉襟见肘,有人建议直接使用 SGML 作为 Web 语言。但 SGML 太庞大了,学用两难尚且不说,就是全面实现 SGML 的浏览器也非常困难。于是 Web 标准化组织 W3C 建议使用一种精简的 SGML 版本——XML。XML 与 SGML 一样,是一个用来定义其他语言的元语言。与 SGML 相比,XML 规范不到 SGML 规范的十分之一,简单易懂,是一门既无标签集也无语法的新一代标记语言。

由于各类应用主体节点在应用范围、构建方式、系统结构、数据资源等方面存在一定的差异,对整个电子政务平台的平稳、高效、安全的运行存在较大的影响;电子政务平台的数据共享要求异国在异构平台、异构环境、异构网络中实现数据交换,这些必然要求共享的数据、文档格式和公文的标准化、统一化,实现有效的数据共享环境多数据源选择[6]。因此需要借助一个能够描述数据交换和业务处理流程的规范标准,以减少数据在处理过程中因标准不统一而引起的诸多问题。

数据交换平台中采用的核心技术是 XML 技术和 Web Service 技术。这两方面的技术已经较为成熟,并在各种场合被广泛应用。

目前 XML 技术通常应用于企业和政府间系统连接、企业和政府内系统连接和文档管理等方面,并有着一系列的标准来支持这些应用的开发,如用于电子商务的 ebXML 及行业数据交换标准 aceXML、MML、DSML 等,用于文档表示的 XHTML、SMIL、MathML 等。这些标准的制定,极大地支持了 XML 应用的普及,使其成为目前大多数软件产品和项目开发必不可少的技术支撑。

关于 Web Service 技术,目前同样已经有一整套标准协议供产品开发使用,包括简单对象访问协议(SOAP)、Web 服务描述语言(WSDL)、Web 服务发现协议(UDDI)等。SOAP 协议提供了在无中心分布环境中使用 XML 交换结构化有类型数据的简单轻量的机制。WSDL 协议定义了服务描述文档的结构,如类型、消息、端口类型、端口和服务本身。DISCO 协议定义了如何从资源或者资源集合中提取服务描述文档、相关服务发现算法等。相对于 XML 而言,Web Service 的应用正在推广普及阶段,部分新项目开始使用 Web Service 技术来实现系统间互操作。

基于以上分析,在数据交换平台的开发中应用元数

据技术，结合全新的XML技术和SOA技术，并制定电子文档的XML交换的数据共享规范和标准，对数据源采用统一接口转化成XML格式以便与不同的信息系统实现便捷的数据交换。

2.4.4 元数据消息服务机制

消息服务的主要功能是保证数据交换的安全可靠，在数据交换的过程中，数据交换的参与方以及数据交换平台需要通过消息的传递实现对数据交换的过程控制，包括通过消息机制实现数据更新的通知、数据交换的请求、数据接收的确认及数据传输错误的纠错等。因此数据交换平台要制定并实现统一规范的数据交换消息协议，应用系统必须通过标准的消息协议和数据交换平台以及其他应用系统进行通信，以控制数据交换的整个过程。

2.4.5 元数据传输服务机制

数据传输服务的主要功能是实现高速的数据传输通路，保证交换数据的时效性、可靠性和一致性，并支持多种数据传输的模式。数据交换平台通过统一规范的数据传输协议，在应用主体和数据交换中心之间传输规范化的交换数据。数据传输服务将根据传输数据量的大小采用不同的传输模式，从而实现数据流的高效传输。

2.4.6 元数据交换引擎

数据交换引擎由XML-RDBMS中间件、数据模式管理、数据访问服务、数据交换服务组成。

XML-RDBMS中间件是协同平台最重要的核心部件，它实现了由XML数据到关系数据库的双向映射，即数据从关系数据库中生成并转换为XML，或将XML数据转换到关系数据库中。

数据模式管理服务是各应用主体和数据交换中心进行数据交换操作时表明要请求和操作的数据的格式和含义，由数据模式的XML Schema定义。数据交换中心收集各应用主体发布的Schema，并按照提供者和类型进行存储。通过映射工具将各子系统的关系型Schema合成为一个全局的关系模式，并通过XML Schema-RDBMS的映射在数据交换中心数据库自动生成相应的表结构，以后传递过来的数据也能够自动根据该映射存放到中心数据库的表中。数据交换中心可以根据所请求的Schema自动路由到提供该Schema的子系统中去。元数据交换服务模式主要包括“发布—订阅”和“请求—应答”2种。

“发布—订阅”模式是由元数据交换服务的提供方提供交换元数据的相关服务发布到数据交换中心，而由元数据交换服务的需求方订阅数据交换中心的相应服务，服务提供方会自动将发生改动的源数据发送给订阅相关服务的服务需求方。该模式是服务提供方主动发起的元数据交换模式。

“请求—应答”模式是元数据交换服务的需求方向数据交换中心请求执行相关获取交换数据服务，数据交换中心通过与元数据交换服务提供方的交互获取相关结果，以应答方式反馈给数据交换的需求方。该模式是服务需求方主动发起的数据交换模式。

本系统实现的主要运行环境myeclipse6.0.1、JDK1.7、TOMCAT6.0、数据库ORACLE10g、SSH构架。

科技管理数据共享平台实现了政府减少重复建设、减少投资浪费的号召；同一数据在多个部门的多个业务系统中共享，实现了科技数据集约化管理，避免产生多个数据出口、多头上报、数据冗余等问题；数据及时整合，实现了对全局数据灵活的多维度分析和多样式展示，满足了管理层监控和决策的需要。

参考文献

[1] 王媛媛.国内政府信息资源元数据研究综述[J].现代情报，2008(3):89-91.
[2] 林瑞峰，陈平华，林锦川.面向科技管理的数据共享平台关键技术研究[J].现代计算机，2009(9):104-106.
[3] 张英俊，谢斌红，郭勇义.元数据技术在科学数据共享平台中的应用[J].太原理工大学学报，2009，40(4):341-344.
[4] WANG Juan Le, ZHU Yun Qiang, SONG Jia, et al. Study on resource and environment scientific research data archiving[C]. 2009 International Conference on Environmental Science and Information Application Technology, 2009.
[5] YING Su, LEI Yang. Assuring image quality in spatial data sharing platform for disaster management[C]. 2008 International Workshop on Education Technology and Training & 2008 International Workshop on Geoscience and Remote Sensing, 2008.
[6] 汪晓庆，郑彦兴，史美林.一种有效的数据共享环境多数据源选择算法[J].软件学报，2008，19(2):314-322.

(收稿日期：2009-12-05)

作者简介：

熊建斌，男，1976年生，硕士，主要研究方向：数据挖掘，信息安全。

李振坤，男，1949年生，教授，研究生导师，主要研究方向：网络安全，分布式网络等。