浙 江 林 学 院 学 报 2003，**20**(4)：403～407
Journal of Zhejiang Forestry College

文章编号：1000-5692(2003)04-0403-05

XML 技术在森林资源管理信息系统异构数据集成中的应用

吴达胜，方陆明，唐丽华，刘丽娟
（浙江林学院 信息工程学院，浙江 临安 311300）

摘要：森林资源管理信息系统在发展过程中积累了大量的异构数据，导致了各种系统间的数据不能进行有效共享，系统操作人员重复劳动多，差错率高，森林资源管理部门人员难以利用现有数据进行有效的决策分析。如何将这些异构数据进行有效集成，这是实现林业信息化亟待解决的一个重要问题。分析了目前森林资源管理信息的数据集成需求，对各种数据集成的常用中间件技术进行了比较，提出了森林资源异构数据集成系统的分层结构模型。这种结构模型包含信息源层、XML 中间件层、XML 接口层和表现层。论述了基于 XML 的数据集成系统的中间件层的实现。图 1 参 9

关键词：森林资源；管理信息系统；异构数据；中间件
中图分类号：S757　　文献标识码：A

1　森林资源管理信息系统异构数据集成的必要性

森林资源是一个地区和一个国家重要的再生生物资源，它与人类的物质生活和精神生活密切相关[1]。森林资源信息是一种表达和控制森林资源运动状态和方式的数据，只要有森林资源它就存在，反映了信息绝对性和普遍性[2]。在新的世纪里，森林资源信息管理要以网络为基础，信息交换平台的建设是森林资源信息管理系统的核心[3]。为了实现维护生态平衡，实现森林资源的可持续发展，我们必须以实时掌握动态森林资源信息为前提，以数据集成为基础，以数据挖掘为手段，以知识发现为依据，以决策分析、控制、评价为目标，实现对森林资源的信息化管理。进行数据挖掘，开发决策支持系统是森林资源信息系统发展有待解决的问题，而数据集成是开发决策支持系统，进行数据挖掘的必要条件。

森林资源管理信息系统在发展过程中积累了大量数据，而且森林资源管理部门为存储和管理这些数据不断投资。然而，由于实施数据库管理系统的阶段性、技术性以及其他经济和人为因素的影响，以致在不同的森林资源管理部门或是同一个森林资源管理部门内部采用的数据库管理系统也大不相同，导致以下的不一致性：①数据结构的不一致性；②处理方式的不一致性；③服务范围的不一致性。这些不一致性导致了各种系统间的数据不能有效共享，系统操作人员重复劳动多，差错率高，森林资源管理部门人员难以利用现有数据进行有效地决策分析。同时，网络的发展使林业部门逐渐从一

收稿日期：2003-05-30；修回日期：2003-09-22
基金项目：浙江省教育厅资助项目(24110050023)
作者简介：吴达胜(1972－)，男，浙江庆元人，讲师，高级程序员，硕士，从事森林资源信息管理和信息系统等研究。E-mail：WU62390710@263.net

个孤立节点发展成为不断与网络交换信息和进行有关业务活动的实体，森林资源数据集成也从林业内部集成走向了林业部门间集成。这样，一方面，现在的森林资源管理需要将各种数据进行交换和在网上发布。为了满足这种要求，我们必须将各种有关的异构数据源进行集成。另一方面，为了保护原有的信息化投资，这种数据集成又不能只是简单地将其他系统的数据一次性移植到某一个系统中。传统的数据库集成方法现在已经远远不能适应人们获取数据的需求，因此迫切需要一种新的数据集成系统[4]。

2 异构数据集成技术

当前，实现异构数据源的集成一般有 2 种方法。第 1 种是将原有的数据移植到新的数据管理系统中。为了集成不同类型的数据，必须将一些非传统的数据类型转化成新的数据类型。这种集成方式的缺点是随着数据管理系统的升级，原来数据的相关应用软件，或是被废弃或是重新开发，以适应新的数据管理系统，因此，通常移植到一个新系统不是一个实际的解决方案。第 2 种方法是利用中间件集成异构数据源。该方法并不需要改变原始数据的存储和管理方式。中间件位于异构数据源（数据层）和应用程序（应用层）之间，向下协调各数据源，向上为访问集成数据的应用提供统一数据模式和数据访问的通用接口。各数据源的应用仍然完成它们的任务，中间件系统则主要为异构数据源提供一个高层次集成服务。显然，中间件系统模式是实现异构数据集成较理想的解决方案[5]。

早期的数据来源主要是各种关系型数据库，因而集成主要针对关系数据库进行。像 ODBC（open database connectivity）方法和传统的模式集成方法都是典型的对关系数据库进行集成的方法。随着信息技术的迅速发展，数据的存储超出了关系数据库的范畴，相应的也就产生了跨平台对多种类型的数据进行集成的要求。新出现的技术例如：微软的通用数据访问结构、三层集成方案、DCOM/CORBA（distributed component object model/common pbject request broker architecture）和用 XML（extensible markup language）进行集成等都可以对多种异构的数据进行集成。

ODBC 方法。ODBC 最初是由制定 UNIX 标准的 X/OPEN 集团和 SQL Access Group 提出的。Microsoft 是 ODBC 的实现者。目前，ODBC 已被其确定为 WSOA（the windows open system architecture，即 Windows 开放系统体系结构）的主要部分。ODBC 之所以得到广泛的应用，首先在于它具有良好的数据独立性。使用 ODBC 编写的应用更改起后台数据库来非常方便——只要更改相应的驱动程序就可以了，在实现上即表现为简单地装入不同的 .DLL 文件。这一点也使得利用它可以缩短开发时间。ODBC 使用层次的方法来管理数据，即在数据库通讯结构的每一层对可能出现产品依赖的地方都引入一个公共接口以解决潜在的不一致性[6]。

DCOM，CORBA 技术。20 世纪 90 年代以来，分布对象技术（DOC）得到了迅速的发展，随着研究的深入和应用的日益广泛，DOC 形成了 2 个阵营：一个是 Microsoft 公司，使用 DCOM 技术；另一个是 OMO 组织，使用 CORBA 技术。DCOM 是组件对象模型（COM）的进一步扩展。COM 定义了组件和客户之间的相互作用方式，它使得组件和客户端之间无需任何中介组件就能相互联系。客户可以通过组件对象提供的接口直接访问组件中的方法。DCOM 技术只适用于 Windows 平台，现在虽然在 UNIX 平台上有了一定的扩展，但效果仍不理想。但是，因为它和 Windows 都是微软的产品，因而可以和操作系统紧密相关，从而大大提高了它的运行效率。

CORBA 是 OMG 的对象管理体系结构中的一个关键组成部分，利用它用户可以在异种平台上开发分布式面向对象应用，而不必考虑各种平台的细节和差异。目前已经有很多家公司开发了基于 CORBA 的应用。CORBA 的跨平台能力非常优秀，但正因为此，所有与操作系统之间的交互必须通过中介代理进行。这使得它的运作效率不如 DCOM[7]。

DCOM 和 CORBA 都采用了包装的思想，以统一的接口的方式向外提供调用，并且二者也都实现了对对象的透明访问，这就给我们对数据集成提供了极大的便利。我们可以利用 DCOM 或 CORBA 将对源数据进行的部分操作进行统一的包装，而后就可以很容易地在此之上建立集成模块，对包装过的数据进行集成，再提交给用户。

利用 XML 对数据进行集成。近年来，互联网发展迅猛，WEB 实际上已经成为最大的数据来源，

而且互联网的用户数和网页的数量仍然在飞速增加。在这种情况下，对网页上的数据进行集成势在必行。但是由于现在的网页大多用 HTML 编写，它缺乏必要的结构和语义信息，给数据集成带来了很大的困难，甚至使得几乎没有可能设计出一种集成 HTML 页面上的信息的通用的方法。为了便于以后人们共享网上的数据，在 3WC（world wide web consortium）带头下，建立了正式的 XML 规范。XML 代表可扩展标识语言，它是一个定义其他语言的标准。它采用将结构、内容和表现相分离的办法，1 个 XML 源文档只写 1 次，就可以用不同的方法表现出来。

XML 使用 DTD（data type definition）和 Schema 来定义数据的结构，利用它可以确认文档中数据是否有效，但更重要的是它们还能够定义数据的类型和数据间的关系，使 DTD 和 Schema 的功能类似于数据库的元数据。充分利用二者的相似性，就可以将传统的数据集成策略以 DTD-Schema 为桥梁移植到对 XML 文档的集成上来，从而实现对 XML 所写的 WEB 页面的集成。同时，还可以利用它对很多类型的信息进行高级集成如关系数据库、文件和多媒体信息等，我们都可以为其设计相应的包装器，将其包装成统一的 XML 格式的数据，然后对这些 XML 数据进行集成，再将集成后的结果数据以 XML 文档的形式发送到各个应用客户端或是更高级的数据集成器去[8,9]。

3　建立森林资源管理信息系统异构数据集成系统必须解决的问题

异构数据源集成是数据库领域的经典问题，并随着 XML 技术的兴起，再次成为该领域研究的一个热点。在构建森林资源管理信息系统异构数据源集成系统时，主要会面对以下几方面问题。

3.1　异构性

异构性是森林资源管理信息系统异构数据集成必须面临的首要问题。其主要表现在 2 方面：系统异构，数据源所依赖的应用系统、数据库管理系统乃至操作系统之间的不同构成了系统异构；模式异构，数据源在存储模式上的不同。一般的存储模式包括关系模式、对象模式、对象关系模式和文档嵌套模式等几种，其中关系模式为主流存储模式。需要注意的是，即便是同一类存储模式，它们的模式结构可能也存在着差异。例如 Oracle 所采用的数据类型与 SQL Server 所采用的数据类型并不完全一致。

3.2　完整性

异构数据源数据集成的目的是为应用提供统一的访问支持。为了满足各种应用处理（包括发布）数据的条件，集成后的数据必须保证一定的完整性，包括数据完整性和约束完整性两方面。数据完整性是指完整提取数据本身。一般来说，这一点较容易达到。约束完整性中，约束是指数据与数据之间的关联关系，它惟一表征数据间逻辑的特征。保证约束的完整性是良好的数据发布和交换的前提，可以方便数据处理过程，提高效率。

3.3　性能

网络时代的数据应用对传统数据集成方法提出了更高的标准。一般说来，当前负责集成的应用必须满足：轻量快速部署，即系统可以快速适应数据源改变和低投入的特性。

3.4　语义冲突

信息资源之间存在着语义上的区别。这些语义上的不同可能引起各种矛盾，从简单的名字语义冲突（不同的名字代表相同的概念），到复杂的结构语义冲突（不同的模型表达同样的信息）。语义冲突会带来数据集成结果的冗余，干扰数据处理、发布和交换。

3.5　权限瓶颈

由于数据库资源可能归属不同的单位，所以如何在访问异构数据源数据基础上保证原有数据库的权限不被侵犯，实现对原有数据源访问权限的隔离和控制，就成为连接异构数据源必须解决的问题。

3.6　附加约束

集成 2 个或多个数据源的时候，数据源的数据之间可能存在着某种联系，把这种逻辑联系附加到集成结果中的过程就称为附加约束。

3.7　集成内容限定

多个数据源之间的数据集成，并不是要将所有的数据进行集成，那么如何定义要集成的范围，就

构成了集成内容的限定问题。

3.8 数据访问的透明性

用户对于数据集成过程中的数据访问是透明的，这样用户不必了解太多的底层数据源管理系统的知识就能够容易地实现各种类型数据的集成。

3.9 可扩展性

为了实现数据源的“即插即用”，在构建森林资源管理信息系统异构数据源集成系统时应该具有可扩展性。

4 森林资源管理信息系统异构数据集成中间件的实现

在集成系统中为了实现数据库的“即插即用”，可以建立一个通用数据库中间件，通过在系统业务逻辑层、通用构件服务层（如 ADO，ADO.NET，或是用户自定义控件等）和数据源层之间建立一个中间层，对服务层屏蔽数据源的差异。中间件向服务层提供一致的数据视图，完成从实际数据源到用户数据视图的转换，并在中间充当数据总线。当选用了中间件作为森林资源管理信息系统异构数据源集成的解决方案后，必须为中间件系统选择一种全局的数据模式。负责集成的中间件系统必须提供一种全局数据模式来统一异构的源数据模式。通过引入 XML 技术，将 XML 技术与全局数据模式相结合可以使异构数据源集成中间件系统能更好地适应于开放、发展环境中的数据集成。

4.1 集成系统的分层结构模型

系统分为 4 层结构（图 1），由下至上各层的基本服务功能如下：①信息源层，处于最低层，是系统的数据提供者，在此应该包括森林资源管理中用到的各种类型的数据库、文件和多媒体等信息。②XML 中间件层，提供必要的数据转换功能或工具，进行数据与 XML 格式的相互转换，将数据存储到 XML 数据空间中，并维持 XML 数据空间与各异构数据源之间的映射关系。③XML 接口层，依据特定的协议或协作模型，负责不同应用组件请求格式的信息发布。不同的组件可以在这层被表示，不同的应用组件需要从应用级别访问 XML 数据空间。例如包括 XML 浏览器的组件，通过 CORBA 接口使分布式对象应用程序与 XML 数据空间进行交互。一方面，实现必要的策略保持 XML 数据的一致性，从简单的读/写策略到复杂的事务操作。另一方面，接口级必须实现必要的访问控制策略，防止非法访问。④表现层，即用户界面层，根据具体的应用和用户计算环境，采用合适的信息访问技术或应用软件。

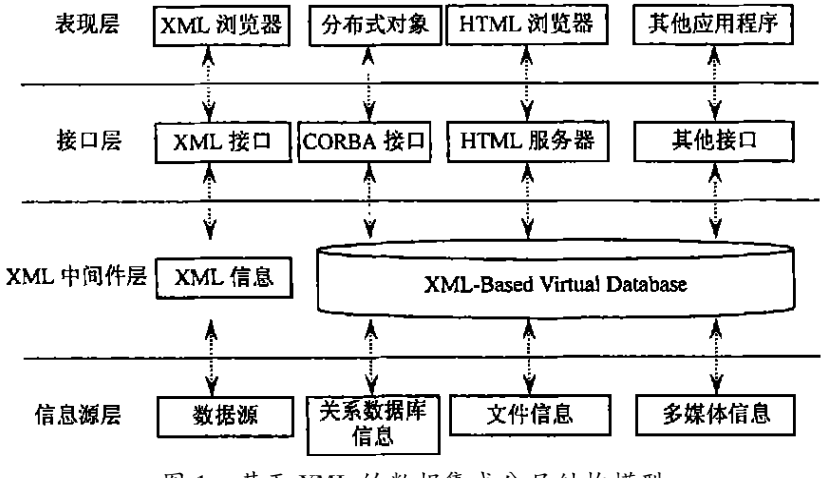

图 1 基于 XML 的数据集成分层结构模型

Figure 1 The layer structure model for data integration based on XML

4.2 基于 XML 的中间件层的实现

可以看出，不同的应用需求，可以采用不同的实现技术。以上所述的数据集成系统的 4 层结构中，从实现角度来看表现层和信息源层相对比较简单，而接口层，根据具体应用的需求，可以采用多

种手段实现用户对数据的访问。关键部分就是 XML 中间件层。在此我们只对模型中 XML 中间件层的实现进行探讨。在数据管理上我们采用“虚拟集中”的方式进行，也即所有数据的变化请求均需通过 XML 中间件层进行存取访问，而中间件层并不存储具体的数据，只存储所有数据的集成模式。具体说，每一个异构信息由一个传统数据源和一个“外套（wrapper）”构成，它通过一个 XML 接口（引擎）作为低层 DBMS 或其他数据源的包装。在不改变服务器中原 DBMS 或其他数据源的前提下，只需用一种统一的可扩展语言 XML 为各种服务器做一件外观统一的“外套（wrapper）”，就可构成一个“虚拟数据库服务器”，为系统提供虚拟数据库服务。对 XML 中间件层，主要涉及 2 个问题：一个是针对每个数据源的 wrapper，即完成某种类型数据源与虚拟数据库之间的双向映射；另一个是 XML-enabled 的集成数据（虚拟数据库）公共模型的建立及管理。

5　结论与讨论

从目前森林资源管理信息的数据集成需求出发，分析了森林资源信息管理中异构数据集成的必要性及比较了各种数据集成的常用中间件技术，提出了森林资源异构数据集成系统的分层结构模型并探讨了基于 XML 的数据集成系统的中间件层的实现。今后需要选择某个市（县）作为实验地点，利用已建立并在实际中使用的各种类型的数据库，开发 XML 组件层、XML 接口层和客户端程序，并进行集成调试和实际使用。

参考文献：

[1] 张志耀，陈立军. 森林资源经营管理决策支持系统[J]. 系统工程理论与实践，1998，(10)：119－125.
[2] 方陆明. 我国森林资源信息管理的发展[J]. 浙江林学院学报，2001，**18**(3)：322－328.
[3] 方陆明. 我国森林资源信息管理网络系统解决方案的探讨[J]. 北京林业大学学报，2003，**25**(3)：127－130.
[4] 靳强勇，李冠宇，张俊. 异构数据集成技术的发展和现状[J]. 计算机工程与应用，2002，**20**(11)：112－114.
[5] 周竞涛. 企业异构数据集成[OL]. Available from http://www.e-works.net.cn/ewkArtiCles/Category13/ArtiCle11525.htm，2002-10-28.
[6] 谢鸿强，董逸生. 异构数据源的集成技术[J]. 工业控制计算机，2001，**14**(6)：1－6.
[7] 李冠宇，靳强勇，张俊. 一个改进的基于 CORBA 的异构数据集成系统体系结构[J]. 交通与计算机，2001，**19**(4)：45－47.
[8] 李军怀，周明全，耿国华，等. XML 在异构数据集成中的应用研究[J]. 计算机应用，2002，**22**(9)：10－12.
[9] 董玉萍，邹承明，钟珞. 基于 XML 的异构数据源共享技术的研究[J]. 武汉理工大学学报，2002，**24**(9)：90－92.

Discussion of XML in heterogeneous data integration of forest resources management information systems

WU Da-sheng，FANG Lu-ming，TANG Li-hua，LIU Li-juan
(School of Information Engineering，Zhejiang Forestry College，Lin'an 311300，Zhejiang，China)

Abstract：Large amount of heterogeneous data have been accumulated along with the development of the forest resources management information systems. The result of this is that data are not effective shared among various systems and the operators repeat work with high error rate. Therefore，the decision-making is difficult made by the manager through the existing data. The emergent issue is how we integrate the heterogeneous data. First，this paper analyzes the demand of data integration of the current forest resources management information system and compares the various middleware technology. Secondly，this paper proposes the model with layer structure，which servers for the integration of the heterogeneous data of forest resources management systems. This model consists of：(1) information source layer；(2) XML middleware layer；(3) XML interface layer；(4) XML presentation layer. Finally，the paper discusses the implementation of data integration system based on XML，using XML middleware layer. [Ch，1 fig. 9 ref.]

Key words：forest resources；management information system；heterogeneity data；middle components