第 16 卷增刊　　湖　泊　科　学　　Vol.16, Suppl.
2004 年 12 月　　JOURNAL OF LAKE SCIENCES　　Dec. , 2004

数字流域环境数据集成与共享研究*

张 鹏[1,2]　李士进[1]　叶 健[2]　王志坚[1]

(1:河海大学计算机及信息工程学院, 南京 210098; 2:江苏省水利厅规划办, 南京 210029)

提　要　作为数字地球的有机组成部分,数字流域是一个异构的、分布式的巨型信息系统. 数据集成与信息共享是数字流域的基本要求和技术难点. 没有良好的数据集成与共享机制,各地区、各部门的数字流域建设只能停留于“信息孤岛”的水平,不能充分发挥数字化建设的整体效益. 本文提出了基于 XML 技术的三层架构解决数字流域数据集成与共享问题,即应用层,中间层和 XML 包装器. 其中,应用层向中间层发出数据请求,按给定协议接受并处理来自中间层的 XML 文档;XML 包装器是各种异构数据的提供者,中间层是连接应用层和 XML 包装器(wrapper)的桥梁. 文章最后结合我国水资源和防洪管理模式,给出了符合我国实际的一个数字流域数据集成与共享实现框架.

关键词　数字流域,信息共享,XML,三层体系结构

分类号

尽管目前数字流域的严格定义尚不统一,但其基本共识是:它是一个数据和服务应用系统,应用包括“3S”技术、海量数据管理技术、多媒体数据传输等现代信息技术,对流域基础信息进行自动采集、动态监测和虚拟再现,实现流域各种信息的全数字化,从而为水事管理和决策提供科学、全面和综合的决策支持服务. 数字流域是一个数据密集的巨系统,数据集成与共享是数字流域研究与建设的核心问题之一.

1　数字流域及数据互操作

1.1　数字流域

数字流域的构成可大致划分为三大系统,即:基础信息服务系统、专业应用系统和综合管理和决策系统. 三大系统相互联系又相对独立,自下而上,分层构成整个数字流域系统[1](图 1).

其中,基础信息服务系统服务于上层的专业应用和综合决策系统. 基础信息系统主要对各种流域基础信息进行采集、整理、清洁、存贮、管理和融合等,提供面向人、专业应用系统及综合管理与决策的数据服务,同时还可以来自上层专业应用和管理决策的分析数据进行持久性存贮和管理.

* 江苏水利科技推广项目资助.

2003-08-09 收稿;2004-09-12 收修改稿.张鹏,男, 1973 年生,工程师.Email:　pierrezhang@hotmail.com,

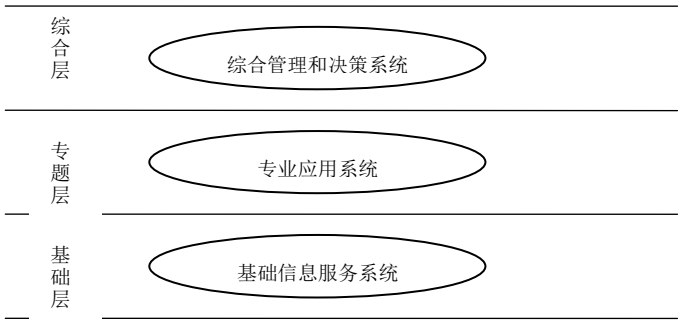

图 1 数字流域的主要内容及多层结构

Fig.1 The architecture of digital basin

1.2 数字流域中的数据及特点

流域专业应用和管理决策的内容十分广泛，其所需的基础数据也极为丰富. 数字流域系统中的数据按其内容可分为：空间地理数据、水文气象数据、水资源数据、水利工程及运行管理数据、资源环境数据、土地利用及社会经济数据、法律法规数据等.

这些数据常由不同数据库系统或不同操作系统下文件系统管理. 在空间分布上相互各异，分别由不同的部门管理和维护. 其数据形式也十分多样化，有结构化数据，如由关系数据库管理的水文、水情数据等，也有半结构化与非结构化数据，如文字图表数据以及音频视频等多媒体数据等. 总之，数字流域数据体现了分布性、时效性、海量性、异构性以及数据利用的综合性等特点.

1.3 数字流域中的数据组织的要求

系统观是数字流域的一个基本出发点. 数字流域要求系统全面地解决各种流域问题，相应数字流域的系统观对其数据组织和管理方式提出了新的要求：

可扩展性和模块性. 数字流域的应用包括各种专题数据和服务，各数据和服务之间相对独立. 数字流域建设的长期性要求系统允许在系统发展过程中新服务和应用可以方便加入，因此可扩展性和模块性是数字流域数据服务的基本要求之一. 这种特点也正 WWW 得以广泛应用的主要原因.

兼容性. 目前 WEB 协议和标准已在世界范围内得到普遍的接受. 作为数字地球的重要应用和实现，只有在现有 WEB 协议和标准的基础上构建数字流域系统才能保持与其它数字系统的相互兼容，并尽可能利用现有网络传输基础设施，降低系统建设的费用.

一致性. 数字流域必须解决不同数据集和服务之间语义冲突，包括格式/平台的差异等.

支持元数据的描述. 元数据包括数据集编码数据表达方式数据质量状况、数据存储介质、数据存储格式、数据量、数据来源等对数据集自身的描述. 元数据可以用于数字流域系统中数据与服务资源的搜索代理、目录服务及空间查询等.

显然，数字流域系统中数据内容的丰富性、数据形式的多源异构等特征与数字流域系统服务与应用的统一、无缝要求存在矛盾. 随着我国信息化建设进程的加快，各地各部门已逐步建设并积累了越来越多的基础信息库，但由于数据的集成和共享的困难，“信

息孤岛”效应正日益成为制约我国信息化发展和数字流域建设的瓶颈.

1.4 数据共享与互操作

目前，解决数据共享的模式大致有三种：数据格式转换模式、直接数据访问模式和数据互操作模式[2]. 对于数据格式转换模式，由于缺乏对数据对象统一的描述方法，不同数据格式描述空间对象时采用的数据模型不同，因而转换后不能完全准确地表达原数据的信息，经常性地造成一些信息丢失，且各转换方案还没有为数据的集中和分布式处理提供解决方案，不能自动同步更新. 直接数据访问提供了一种更为经济实用的多源数据共享模式，但直接数据访问同样要建立在对要访问的数据格式的充分了解的基础上，如果要访问的数据的格式不公开，就难以实现了.

数据互操作模式 数据互操作为多源数据集成提供了崭新的思路和规范，为数据集中式管理、分布式存储与共享提供了操作的依据. 数据互操作模式指在异构数据库和分布计算的情况下，用户在相互理解的基础上，能透明地获取所需的信息. 数据互操作是当前数据共享与集成研究的热点问题之一.

数据互操作模式主有两种实现途径：数据仓库技术（Data Warehouse）和基于中间层系统（Mediator-based System）. 两种技术的差别主要在于各组件之间关系的紧密程度和可扩展性上. 数据仓库技术将基础数据交给少数设计良好、联系紧密的数据中心管理. 数据中心内数据的集成由事先的“预计算”完成. 该模式优点在于少数核心数据中心的运行效率高，但缺点在于不利于数据中心以外不断增加的数据源的集成，特别是大量半结构化数据的集成. 而基于中间层系统则是建立在大量相对自治的数据和服务源的基础上，各数据源间通过统一的标准协议相互通信，在需要时完成的数据集成.

显然，基于中间层系统的体系结构更为适应数字流域系统可扩展性和模块化的要求. 本文将根据数字流域体系的特点，提出基于中间层系统的数字流域数据集成与共享三层结构框架.

2 基于 XML 技术的数字流域数据集成与共享

2.1 数据集成与共享框架

XML 即可扩展标记语言（Extensible Markup Language），是由 W3C 组织提出的一种独立于软、硬件的跨平台信息交互工具. XML 定义了结构化表达数据的标准格式，是一种自描述的标志语言. XML 的主要特点为结构性、可扩展性及平台独立性[3].

根据前文讨论，数字流域系统中基础信息服务层为上层的专题服务和决策管理提供所需的数据. 显然，XML 可以对不同来源异构的数据进行描述，对服务层屏蔽数据源间的差异，形成一致的、可以被不同服务理解的 XML 文档，向上层提供透明无缝的数据服务. 因此，可以将 XML 作为异构数据源间集成与访问的中间层. 同时 XML 的可扩展性允许方便地拓展新的服务和应用，满足了数字流域建设统一规划，分步实施的要求. XML 是纯文本文件，可在现有网络传输协议的基础上实现远程传输，十分方便.

数字流域数据集成与共享框架按三层构建：(1) 经 XML 包装器（wrapper）屏蔽的基础数据及服务网络层. 该层处于底层，是数据的提供者，包括所有空间地理数据管理系统、水资源水文数据库及文件多媒体数据管理系统等；(2) 支持数字流域系统内资源发现和查询的中间层. 该层利用 XML 技术，将应用层发出的数据要求和数据服务层反馈结果按给定的协议，进行双向交互；(3) 各终端应用及决策管理层. 该层向中间层发出数据

请求，按给定协议接受并处理来自中间层的 XML 文档框架构造如图 2 所示.

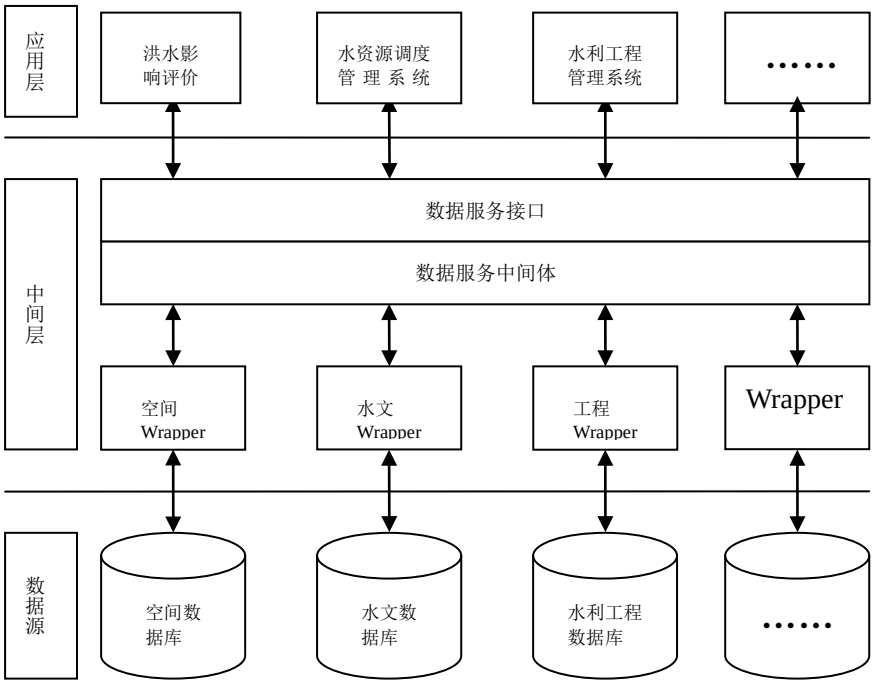

图 2 数字流域基础数据集成与共享框架

Fig.2 The framework of data integration in digital basin

2.2 数字流域系统框架中间层

中间层由三部分的中间件（Middleware）构成，分别为数据服务接口、数据服务中间体（Data Server Mediator）和包装器（Wrapper）[4]. 各中间件均以 XML 文件格式进行相互之间的消息传递，以解决数字流域环境下数据格式和平台的差异性.

数据服务接口. 其主要作用为根据不同的协议，如 HTTP、CORBA、RMI、应用客户等提供接口.

数据服务中间体（Data Server Mediator）. 数据服务中间体接受来自应用层的数据请求，并根据数据源层中各数据源的可提供的服务，将其分割成相应数据请求片，分别发送至各数据源包装器. 各数据源完成各自的数据处理任务，结果返回后，再由数据服务中间体将各结果片集成至统一的单个结果，返回至应用层，完成整个多源异构数据的集成.

包装器. 包装器的任务主要是对来自中间体的请求转换成各数据源的宿主语言或 API，并将来自数据源的请求结果转换成中间体语言. 因此，包装器的作用相当于数据源和中间体间的代理. 这样，中间层便可实现对不同数据格式和多种协议标准下的分布式数据进行统一管理.

2.3 中间层的实现策略

数字流域的构建与实现必须充分考虑管理模式的制约. 流域由各级子流域（分区）

组成，同时流域也由各级行政区域交叉构成. 2002 年 10 月 1 日颁布实施的新《水法》规定“国家对水资源实行流域管理与行政管理相结合的管理体制. ……”. 1998 年 1 月 1 日颁布实施《防洪法》又规定：“…流域管理机构在所管辖的范围内行使法律、行政法规规定和国务院水行政主管部门授权的防洪协调和监督管理职责. ……县级以上地方人民政府水行政主管部门在本级人民政府的领导下，负责本行政区域内防洪的组织、协调、监督、指导等日常工作. ……”. 可见，我国防洪和水资源管理为以流域管理为纵线，行政管理为横线的矩阵式管理模式. 无疑这种管理模式增加了数字流域实现的难度. 图三以一个省级数字流域构建为例，说明数字流域中间层的实现策略.

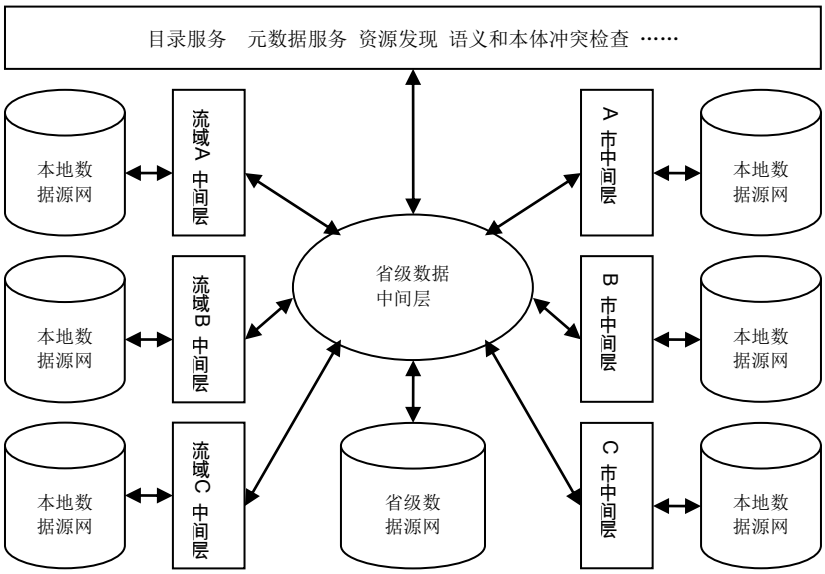

图 3 一个典型省级数字水利中间层实现策略

Fig.3 An implementation strategy of province-level digital basin mediator

在图 3 所示中间层实现策略中，当某种应用需要时，首先向本地中间层发出数据请求，如本地数据源网没有所要求的资源，则本地中间层向省级数据中间层转发数据请求，省级数据中间层则根据应用中间层明确要求的数据源向相应的流域或市级中间层转发数据请求，由该中间层完成该数据请求任务后，返回结果. 在上述数据源选择方式中，目标数据源选择方式有两种. 一种由数据服务中间体根据各数据源包装器提供的 Schema 信息，动态地决定最优选择方式；另一种数字流域的实现过程中，合理地规划各数据源的内容，而在应用发出数据请求时根据数据的地区、专题等属性等确定目标数据源. 前者有较大的灵活性，而后者则有更好的运行效率. 考虑数字流域建设主体与应用对象的相对专业性，显然后种方式更适合于数字流域系统.

实际上，在图 3 中，流域下可能还会有子流域，省级数字水利系统下可能还有各专业部门级中间层，这样更增加系统的复杂性.

3　结论

3S 技术和 Internet 技术的飞速发展给以水资源为研究对象流域的规划、管理、决策和建设带来新的契机. 数字流域建设正是以数字化和信息化手段，将人类对流域的认识

提升至一个新的高度，可以全面、系统地开发利用和保护水资源，促进人与自然的协调发展.

数据集成与共享是数字流域的基本要求和技术难点. 没有良好的数据集成与共享机制，各地区、各部门推行的数字流域建设只能停留于“信息孤岛”的水平，不能充分发挥数字化建设的整体效益. 本文提出了基本 XML 技术的中间层技术是解决数字流域数据集成与共享的有效途径，并结合我国水资源和防洪管理模式，给出了符合我国实际的实现策略.

参 考 文 献

1 张秋文,张勇传,王 乘等. 数字流域整体构架及实现策略.水电能源科学,2001,**19**(3):4-7

2 宋国民,贾奋励. 地理空间数据共享机制研究.测绘学院学报,2002,**19**(2):134-136

3 李军怀,周明全,耿国华等. XML 在异构数据集成中的应用研究.计算机应用,2002,**22**(9):10-12

4 Ilya Zaslavsky, Richard Marciano, Amarnath Gupta, *et al*. XML-based Spatial Data Mediation Infrastructure for Global Interoperability. In: 4th Global Spatial Data Infrastructure Conference. Cape Town, South Africa 13-15 March 2000.At: http://www.npaci.edu/DICE/Pubs/gsdi4-mar00/gsdi_iz.html

Research on the Data Integration and Information Sharing in Digital Basins

ZHANG Peng[1,2], LI Shijin[1], YE jian[2] & WANG Zhijian[1]

(1: School of Computers and information engineering, Hohai University, Nanjing 210098, P.R.China;

2:Water Resource Department of Jiangsu Province, Nanjing 210029, P.R.China)

Abstract

The Digital Basin comes from the Digital Earth, which is a huge, complex information system with heterogeneous and distributed components. Data integration and information sharing are two basic requirements and the key technologies to success, without which the information systems of all level management of the Digital Basin will be an “isolated information island”. This paper proposed a new XML-based architecture to tackle this problem, which consists of three tiers, i.e., the application tier, the mediator tier, and the XML wrapper. The application tier sends out data and information requests and receives the required results as XML documents from the mediator. The XML wrapper provides all kinds of heterogeneous data requested while the mediator is the “bridge” between the application tier and the XML wrapper. At last, according to the realities of the management of China water resources and flood-control, an implementation framework of the Digital Basin based on the aforementioned XML-based architecture is provided, which proved that the presented methodology is viable.

Keywords：Digital Basin, information sharing, XML, three-tier architecture