第 34 卷 第 1 期 计 算 机 工 程 2008 年 1 月
Vol.34 No.1 Computer Engineering January 2008
·软件技术与数据库· 文章编号：1000—3428(2008)01—0090—04 文献标识码：A 中图分类号：N945

基于网格的联邦数字图书馆

侯 骏，王永剑，钱德沛，白跃彬，王 克
（北京航空航天大学计算机学院中德软件技术联合研究所，北京 100083）

摘 要：介绍了一种联邦数字图书馆结构，它通过对网络中数字图书馆资源进行一定的服务封装形成数字图书馆资源仓储节点，在此基础上运用网格思想对仓储节点进行整合，形成数字图书馆资源联邦，使用户能够通过联邦门户透明地访问联邦中所有数据仓储节点中的图书馆数据资源。为了满足专业化、个性化数字图书馆建立的需求，联邦提供了以检索为基础的个性化数字图书馆实例的动态定制服务。
关键词：数字图书馆；网格；服务；数字图书馆定制

Grid-based Federated Digital Library

HOU Jun, WANG Yong-jian, QIAN De-pei, BAI Yue-bin, WANG Ke
(Sino-German Joint Software Institute, School of Computer Science, Beihang University, Beijing 100083)

【Abstract】This paper proposes a grid-based federated digital library solution. In the solution, digital library data resource node is created by encapsulating the digital library asset with services. A digital library asset federation is built to integrate depot nodes with grid concept to transparently access the data assets of the integrated nodes. The federation can provide a search-based dynamic customization engine for digital library instance to meet the requirements of building professional and personal digital libraries.
【Key words】digital library; grid; service; digital library customization

随着信息化建设的开展，数字图书馆的建设非常迅速。据易观国际《图书馆信息化建设综合研究报告》[1]显示，目前中国已统计的 15 437 家图书馆数中，18%的图书馆实现了业务的计算机网络化处理；6%的图书馆完全进入数字化阶段，向读者提供数字资源加工和检索服务。通过互联网提供图书馆馆藏资源查询和使用已成为趋势，经过数字化的图书馆资源也成为网络中重要的数据资源。因此，仅中国互联网中就存在数以百计的数字图书馆系统，并且还在不断增加。数字图书馆数目繁多，但数字资源仅供本馆专用，馆际相互独立，缺乏交互，因此，真正被读者了解和使用的数字图书馆为数不多。如果能够对馆藏数据资源进行整合，实现对整合资源的透明访问，读者无须改变使用方式即可访问更多的图书馆资源；各数字图书馆也能够被更多地访问，从而充分提高了资源的使用效率。

1 相关研究工作

近年来基于网格的数字图书馆研究开始起步。网格能较好地解决互联网环境下计算分布和存储分布中存在的交互、协作问题，因此，用网格的观点进行数字图书馆的研究能有效解决数字资源的分布存储、数字图书馆分布构建以及访问。

2000 年，意大利ISTI-CNR 的DLib小组着手开发用于创建和管理数字图书馆的服务系统，取名为“OpenDLib”[2-3]。其目标是开发一个可定制系统，只须进行适当的配置，就能满足不同应用程序框架的需要。在该系统建立的基础上，DLib小组和欧洲的研究机构、软件公司启动了DILIGENT[4]研究计划，目标是开发一个基于网格框架、面向服务的数字图书馆系统，推广使用新方法构造数字图书馆框架。它是欧洲网格研究项目EGEE[5]的子项目，是目前世界上规模最大的应用网格技术支持数字图书馆建设的研究计划。

BRICKS[6]是一个以在EDM[6]上建立数字图书馆的组织和技术基础为目标的综合项目。这里的数字图书馆是一个网络化的系统服务，实现全球化的可用的多媒体数字文档的收集，并针对不同用户和访问模型提供不同的知识层次。它包含了数字博物馆、数字档案馆以及其他广泛使用的数字存储系统。

2 基于网格的联邦式数字图书馆

数字图书馆保存了多种数据资源，包括数字化文本(如书籍、报刊等)以及视频、音频等多媒体。网络中的数字图书馆独立地提供对馆藏数据资源的检索和访问服务。随着信息化建设的开展，数字图书馆的数量越来越多，但其中用户真正能使用的资源非常少，因为用户能够了解并记住的数字图书馆有限，他们往往关注较权威的数字图书馆，而这些资源只占数字图书馆资源总量非常小的一部分；而且对于网络中众多的数字图书馆，用户只能独立访问，因此，用户使用不同数字图书馆就要不断重复类似的人工操作，给用户的查询和使用带来巨大的工作量，在很大程度上限制了用户的资源访问能力。

基于网格的联邦式数字图书馆用网格技术将网络中独立存在的数字图书馆资源在授权条件下进行整合，形成数字图书馆联邦，为用户提供统一入口，透明地访问联邦中所有数据资源，无须了解各个资源具体的存储方式和访问方法。为了能够更好地使用数据资源，联邦式数字图书馆为一类缺乏

基金项目：国家自然科学基金资助项目(90104022，90412011，90612004)；国家发改委基金资助项目“基于 CNGI 的中国国家网格应用”(CNGI-04-15-7A)
作者简介：侯 骏(1982－)，男，硕士研究生，主研方向：网格技术，数字图书馆；王永剑，博士；钱德沛，教授；白跃彬，副教授；王 克，硕士
收稿日期：2007-02-22 **E-mail：**houjun@263.net

资源、但希望拥有个性化专业数字图书馆系统的用户提供基于检索的数字图书馆实例的动态定制服务，使这类用户仅须花费较小的代价就能拥有数字图书馆。

联邦式数字图书馆结构包括 3 部分：数字图书馆数据资源节点（简称数据资源节点），数字图书馆资源联邦以及数字图书馆实例，如图 1 所示。

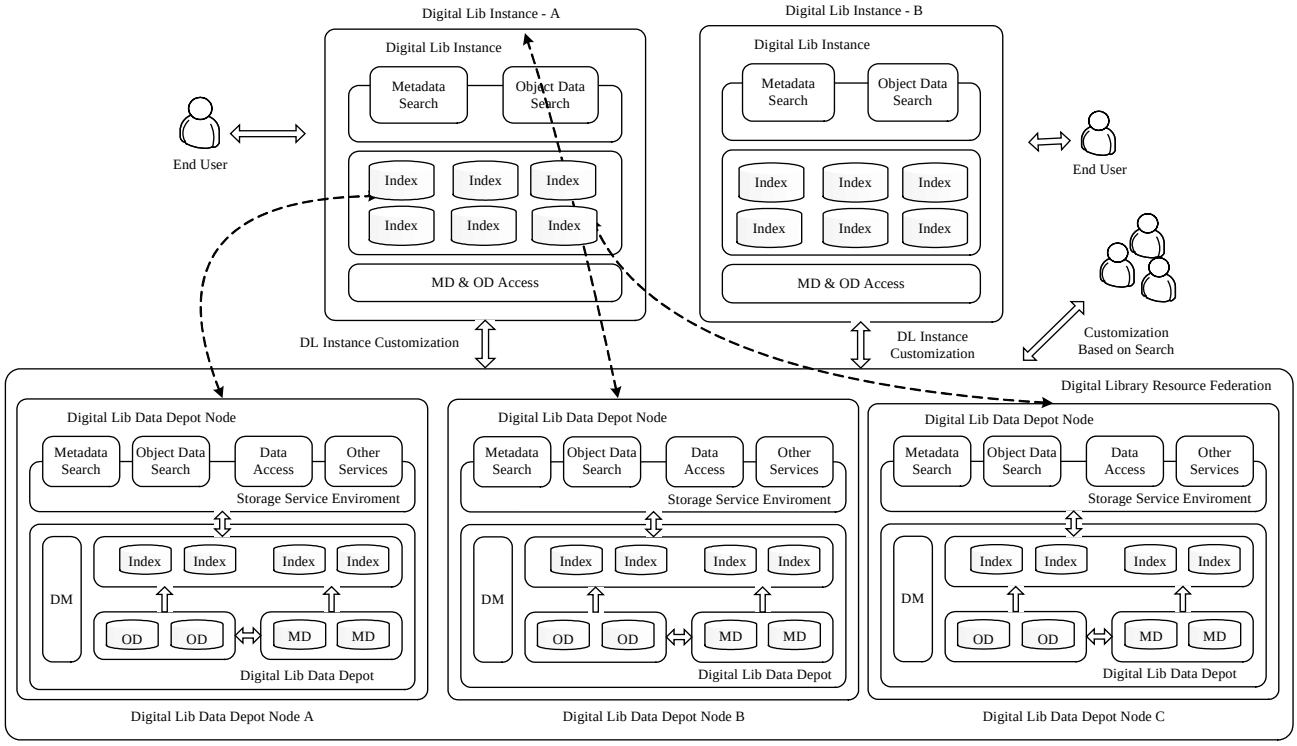

图 1　联邦式数字图书馆体系结构

数据资源节点是一个地理位置中通过服务方式封装数字图书馆数据资源。系统首先使用安全统一的服务接口对馆藏数据资源进行封装，构成能够提供外界检索和访问数据资源节点；在此之上将节点注册到数字图书馆资源联邦，由联邦负责对节点及节点服务的统一管理。数字图书馆资源联邦是整个联邦式数字图书馆的中枢，它负责接入并管理数据资源节点，整合节点服务并暴露给用户，使用户在不改变使用方式的条件下透明地访问联邦中各个节点包含的数据资源。资源联邦还包含一个数字图书馆实例定制引擎，为用户提供数字图书馆实例的定制。数字图书馆实例是使用资源联邦的定制引擎，通过检索联邦中各数据资源节点获取定制数据并配合检索引擎和访问门户动态生成的数字图书馆系统。它向数字图书馆的最终用户提供实例数据资源的检索和访问服务。

3　数据资源节点的存储和服务封装

3.1　数据资源节点中的数据资源

数字图书馆的数据包含对象数据(object data)和元数据(metadata)。对象数据是数字化的数据对象，包括文本、视频和音频等；元数据是对象数据的描述数据，如书籍的书名、作者和出版社等。对象数据以文件形式保存；元数据以记录的形式保存在元数据文件中，一个元数据文件包含若干元数据记录。每一个对象数据都存在一条描述它的元数据记录。为了对数据进行检索，系统为元数据和全文对象数据建立索引。数据和索引就构成了一个数据资源节点中的数据资源。

不同来源的元数据具有不同的格式，如 MARC, DC，因此，系统定义了如下基于 XML 元数据格式，所有元数据记录都转换为该格式保存：

```
<Metadata [namespace]>
    <yy:aa qualifier=”zzzzz”>value1</yy:xxx>
    …
    <yy:aa qualifier=”zzzzz”>valuen</yy:xxx>
</Metadata>
```

对象数据文件和元数据文件都拥有全局唯一 ID，对象数据和元数据的访问都通过 ID 进行。

3.2　数据资源节点的仓储结构

数字图书馆拥有大量的数据资源，并且不断有新的资源加入。因此，这些资源的仓储、检索和数据访问成为广受关注的问题。系统需要一种可扩展、高效的仓储结构进行资源的存储并提供高效的数据检索和访问机制。联邦式数字图书馆的数据资源节点采用了图 2 所示的分布式结构进行构建。

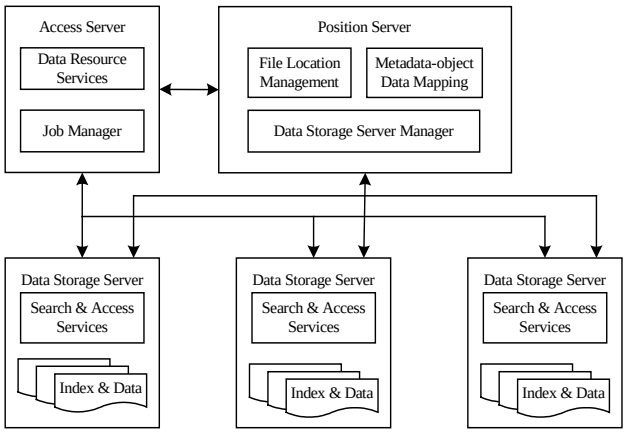

图 2　数据仓储节点仓储结构

这种结构包括若干存储服务器、一个位置服务器以及若干访问服务器。存储服务器用于存储元数据文件、对象数据

文件和索引，并提供服务器存储数据的检索和文件访问。索引除了包括元数据和对象数据的全文索引外，为了能够定位元数据文件中的每一条记录，每个元数据文件都拥有一个以元数据记录 ID 和记录在文件中起始位置(字节)为索引域的定位索引。存储服务器启动后根据配置连接位置服务器，由位置服务器统一管理。

位置服务器是分布式存储结构的核心。它负责监控存储服务器的状态管理，保存了存储服务器中元数据文件和对象数据文件 ID 与存放的映射信息以及对应元数据记录到对象数据的映射信息。文件存放信息由文件所在服务器 IP 和存放的绝对存储路径所确定；对应的元数据记录和对象数据通过 ID 相互映射。在数据检索过程中，位置服务器负责提供可用的存储服务器列表；在数据访问过程中，位置服务器负责提供待访问数据的定位信息。

访问服务器负责提供外界检索和访问数据资源节点数据所需的服务。服务器由资源节点服务及其运行环境以及负责任务划分和执行的任务管理器 2 个部分组成。当访问服务器接到数据检索和访问请求时，服务运行环境调用所请求的服务进行响应，根据被调服务和请求参数向位置服务器查询执行请求任务的存储服务器列表(不同服务的查询内容不同，如数据检索查询服务器状态信息，对象数据或元数据的访问则查询数据文件的位置信息)，任务管理器根据列表将请求任务划分成子任务，并交付给存储服务器执行；访问服务器将合并所有子任务执行结果，统一返回请求任务的结果。

数据资源节点是由这类服务器协同构成的有机整体。运行时位置服务器首先启动，加载存储服务器管理模块和文件定位服务。存储服务器和访问服务器相继启动主动连接位置服务器，并使用心跳连接保持它们和位置服务器间的连接状态。资源节点还包括一个管理工具，负责数据添加和服务器信息维护和管理工作。资源节点内部各服务器都采用 TCP/IP 进行通信；访问服务器可通过多种方式对外提供服务，如 HTTP, TCP/IP 和 Web Service 等。存储服务器能够随着资源节点数据量的增大而增加，新的存储服务器在确保服务器数据的定位和映射信息被正确添加到位置服务器中后，连接位置服务器即可使用。同时，访问服务器可根据访问量的要求灵活增减，进行负载均衡，提高仓储节点的总体响应能力。

3.3 数据仓储节点的服务封装

为了能够透明地访问各数据仓储节点，对各节点进行统一的服务封装，并采用基于 XML 统一的服务调用和返回消息格式：

(1)Service Request Message

```
<service-request>
<service-name>xxx</service-name>
<params>
<param name=”xxx” valuexml=”true/false”>xxx</param>
…
</params>
</service-request>
```

(2)Service Response Message

```
<service-response>
<retcode>xxx</retcode>
<results>
<result name=”xxx” valuexml=”true/false”>xxx</result>
…
</results>
</service-response>
```

服务请求被响应时，节点根据 service-name 调用服务并根据 params 获得服务参数；调用返回时，retcode 表明服务调用成功与否，调用成功时返回结果包含在 results 字段中。每次服务调用可包括若干服务请求和服务响应。

每个数据仓储节点必须提供 4 类用于数据检索和访问的基础服务，包括权限认证服务、检索服务、元数据记录访问服务以及对象数据访问服务：权限认证服务负责界定服务访问的合法性；检索服务负责在节点索引中检索用户的检索条件；元数据记录访问服务和对象数据访问服务是根据 ID 访问元数据记录和对象数据。进行服务封装时，数字图书馆资源联邦负责统一界定基础服务的内容以及请求和响应消息，确保各数据仓储节点访问的一致性。同时，各资源节点能够开发增值服务，满足用户其他数据访问要求。

4 数据仓储节点的整合

数字图书馆资源联邦负责对分布在网络的数据资源节点进行整合，提供被整合节点数据资源的透明访问，如图 3 所示。数据资源节点通过服务封装便可注册加入数字图书馆资源联邦。注册时，数据仓储节点提供节点描述和服务访问信息，4 类基础服务必须根据联邦所要求的服务内容、请求和响应消息格式进行封装，注册后联邦会测试这 4 类基础服务的可用性。联邦中每个数据仓储节点都被赋予唯一的节点 ID 进行标识。

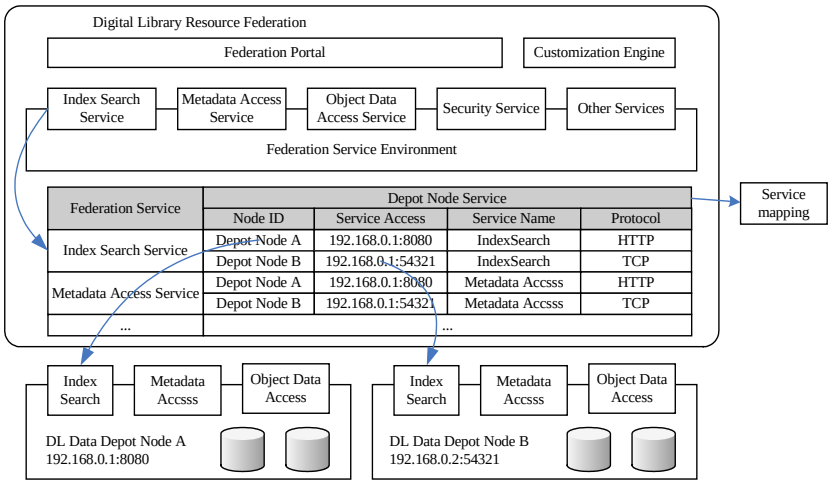

图 3 基于数据资源联邦整合基础结构

数字图书馆资源联邦提供了统一的服务访问联邦整合的数据资源节点，这些服务被称为联邦服务，是对联邦中资源节点提供的同类服务的整合。联邦通过服务映射的方式将联邦服务和数据资源节点服务联系起来，每个数据资源节点服务都会映射到该类服务所对应的联邦服务上。映射还包括资源节点服务的访问信息，如访问入口、服务名称以及网络协议等。数字图书馆资源联邦 Portal 和定制引擎通过使用联邦服务为用户提供数据检索和访问以及数据图书馆实例的定制服务。当用户访问联邦服务时，该服务首先查询服务映射表，获得联邦中提供该类服务的数据资源节点列表及服务访问方式，并针对列表中的节点生成相应的服务访问请求(如

图 3)，因此，用户无须了解节点细节便能对其中的数据资源进行透明访问。

5 数字图书馆实例定制

人们在面对领域问题时常常希望针对领域信息查询。传统数字图书馆通常覆盖所有领域，数字图书馆资源联邦也聚合了各领域的数据资源，难以按领域进行详细的分类。为了满足用户需求，数字图书馆资源联邦在资源整合的基础上提供数字图书馆实例的定制服务，即根据需要个性化地定制数据、系统门户和功能，生成新的数字图书馆系统。

定制包括数据定制和门户定制 2 部分。数据定制是根据检索条件从资源联邦中的数据资源节点中检索获得满足需求的数据资源，检索条件根据用户需求产生。这种方法完全以数据内容的符合程度为标准，用户根据自己的需要构建检索条件，定制引擎根据条件在联邦的资源节点中检索，返回符合条件的结果。图书馆数据资源通常具有严格的版权限制，不能随意复制和传播。因此，数据定制仅从各数据资源节点定制索引，当需要访问具体数据时，系统在授权允许的条件下直接访问保存数据的数据资源节点。

数字图书馆门户是数字图书馆实例的入口，它配合数据定制专业化和个性化的特点，从主题、风格和功能方面提供定制服务。主题定制针对实例面向的领域，定制实例的名称、领域、内容介绍、版权等信息；风格定制是针对门户图片、颜色的定制，使门户风格符合实例主题要求。门户采用组件化方式构建，每个组件实现一个功能。简单检索、高级检索、结果显示以及数据访问等是实例的核心组件；用户还可以选择添加可选功能组件。同时，各组件能够按照用户需求进行布局。

6 模型实现

“CNGI 中国国家网格”的数字图书馆应用是基于联邦式数字图书馆结构建立的。系统包括中国国家图书馆、北京航空航天大学以及西安交通大学 3 个数据资源节点，每个节点包括 1 000 万条元数据记录，600 GB 对象数据，它们存储在 3 台~4 台存储容量为 200 GB 的存储服务器中。数据资源节点通过 Web 方式封装数据的检索和访问，并在中国国家网格中整合，形成数字图书馆资源联邦。联邦提供了门户以及数字图书馆实例定制引擎。前者使用户能够访问 3 个节点中的数据资源，后者与定制客户端工具配合，动态定制数字图书馆实例。

7 结束语

本文针对当前网络中大量独立的数字图书馆系统存在的数据资源使用效率低下问题，提出了一种运用网格思想的联邦式数字图书馆结构，它采用服务的方式对数字图书馆数据资源进行封装，暴露资源的检索和访问而形成数字图书馆数据仓储节点。通过运用网格思想对数据仓储节点进行整合，形成数字图书馆资源联邦，使用户能够透明地访问联邦所整合的图书馆资源而无须了解访问细节，提高了资源的访问效率。为了满足建立专业化、个性化数字图书馆的需求，联邦还提供了一种基于检索的数字图书馆实例的动态构建服务，通过对数字图书馆门户和动态数据的定制动态建立新的数字图书馆供人们使用。

未来的工作包括：(1)着重于单个数据资源仓储节点承载更大数据量(TB 级)资源时的稳定性和性能。(2)伴随着数据资源仓储的加入，数据资源联邦的规模会更加庞大。模型需要保证在这种情况下数据资源仓储能为数字图书馆实例定制提供稳定的服务。(3)进一步丰富数字图书馆实例定制功能，为用户提供更加灵活个性化的服务。

参考文献

[1] 易观国际．中国图书馆行业IT信息化建设研究报告[Z]. (2004-10-23). http://report.analysys.com.cn/?action=showContent&ID=2564&TID=1&WID=13.

[2] Castelli D, Pagano P. A Flexible Repository Service: the OpenDLib Solution[C]//Proc. of the 6th International ICCC/IFIP Conference on Electronic Publishing. [S. l.]: ICCC Press, 2002-05.

[3] Castelli D, Pagano P. OpenDLib: A Digital Library Service System[C]//Proceedings of the 6th European Conference on Digital Libraries. [S. l.]: Springer-Verlag, 2002-04.

[4] DILIGENT: A Digital Library Infrastructure on Grid Enabled Technology[Z]. (2006-09-10). http://public.eu-egee.org/conferences/kickoff/programme/slot-dc.html.

[5] EGEE Team. EGEE: Enabling Grids for E-science in Europe[Z]. (2006-08-15). http://public.euegee.org.

[6] BRICKS Consortium. BRICKS — — Building Resources for Integrated Cultural Knowledge Services[Z]. (2004-05-22). http://www.brickscommunity.org/.

（上接第 89 页）

6 结束语

本文介绍了如何用 UML 的扩展机制构造新的 UML 模型，以及建立并行程序的软件模型。这个软件模型包括了处理单元拓扑、应用程序建模和两者间的映射方法。给出了一个实际工程应用程序的简要建模过程。该过程将并行程序的设计与开发纳入 UML 模型中，提供了一种良好和可信的虚拟建模语言。这种建模语言可以图形化并行程序设计，帮助程序设计人员理解程序、指导实现。它支持分布式内存程序设计，也支持共享内存程序设计，而且可以实现混合开发，解决了其他并行建模语言面向语言单一、与串行虚拟建模语言没有较好的接口等问题。最后给出了一个并行程序设计的例子，证明了它对由粗到细的设计模式有着很好的支持。

下一步的工作包括实现这样一个实用的图形化建模工具，研究对并行 I/O 的建模和对面向对象的并行语言建模。

参考文献

[1] Newton P, Browne J C. The CODE 2.0 Graphical Parallel Programming Language[Z]. (1992-01-05). http://citeseer.ist.psu.edu/63058.html.

[2] Stankovic N, Zhang K. Towards Visual Development of Message-passing Programs[C]//Proceedings of the 13th IEEE Int'l Symposium on Visual Languages. Isle of Capri, Italy: [s. n.], 1997.

[3] Resnick M. StarLogo: An Environment for Decentralized Modeling and Decentralized Thinking[C]//Proc. of Conference on Human Factors in Computing Systems. [S. l.]: ACM Press, 1996.

[4] Pllana S, Fahringer T. Performance Prophet: A Performance Modeling and Prediction Tool for Parallel and Distributed Programs[C]//Proc. of ICPP'05. [S. l.]: IEEE Computer Society Press, 2005.