基于网格的泛在图书馆构建研究

□ 郭文丽 / 中国科学院研究生院 北京 100049
中国科学院国家科学图书馆 北京 100080

摘要：网格为泛在图书馆的构建提供了基础条件。文章通过介绍国外两个具有代表性的基于网格构建泛在图书馆的实例，分析了网格和数字图书馆结合的优势以及所面临的挑战，为国内基于网格技术在一定的地域或领域范围内构建泛在图书馆环境提供借鉴。

关键词：泛在图书馆，网格，虚拟数字图书馆，信息检索

1 引言

泛在技术给图书馆领域带来了新的理念。通过把数字化内容、有线或无线网络、标引等技术结合起来，图书馆可以为用户提供一个泛在图书馆环境[1]。这个环境能够满足用户的广泛需要，使他们可以在任何时候、从任何地方打开定制的图书馆，方便地获取信息。

网格作为一种支持网络资源共享的基础设施，为泛在图书馆的构建提供了条件。网格技术可以在不影响局部自治性的情况下，使用户跨越机构和地理位置的障碍，安全地共享计算能力、数据库与其它在线工具。网格不仅为实现数字图书馆基础设施提供了更为全面和安全的解决方案，也为实现更为有效的数字图书馆服务提供了一个全新的环境[2]。

网格系统可以分为三个基本层次：资源层、中间件层和应用层[3]。中间件层是用来屏蔽网格资源层异质特性的一系列工具，向网格应用层提供透明、一致的使用接口，以支持网格应用的开发。网格中间件也称为网格操作系统，是网格技术的核心所在。

网格技术发展至今，已经取得了很多成果，但离实用化还有很大距离。当前网格技术的局限性也对新一代数字图书馆的构建提出了挑战。例如，网格提供的文档结构无法满足数字图书馆响应环境请求的需要；再如，数字图书馆特有的复杂工作流也难以利用网格技术来处理。因此，目前基于网格构建泛在图书馆，仍是一种前瞻性的研究和探索。

本文通过研究国外两个具有代表性的基于网格构建泛在图书馆的实例，分析了网格和数字图书馆结合的优势以及所存在的问题，为国内基于网格技术在一定地域或领域范围内构建泛在图书馆环境提供借鉴。

2 基于网格的虚拟数字图书馆——DILIGENT

欧盟的DILIGENT研究计划（a Digital Library Infrastructure on Grid Enabled Technology）[4]是一个基于网格中间件gLite和Globus Toolkit的虚拟数字图书馆研究计划，其目的是要在网格环境下建立支持用户个性化需求的虚拟数字图书馆，从而为科研用户构建一个动态的泛在知识环境[5]。

2.1 虚拟数字图书馆的概念

DILIGENT项目首次在网格虚拟组织的基础上提出了虚拟数字图书馆的概念[6]。

虚拟组织是为了解决资源共享和访问控制问题而引入的一个概念。一个虚拟组织是指一个由动态用户（可能来自一个或多个组织）共享的分布式动态资源池。虚拟组织由“用户（User）”、“角色（Role）”、“许可（Permission）”以及三者之间的关系来定义。“用户”是指被授权使用资源

的实体，既包括人，也包括想使用资源的其它资源和服务。“角色”是指与授予用户的权利和责任相关的工作职责。“许可”是指允许对一个或多个对象进行操作的权利。每个“用户”可担当若干“角色”，而每个“角色”又可拥有若干“许可”。在DILIGENT中，“许可”与“角色”、动作及资源都有关系。

虚拟数字图书馆是利用虚拟组织这种机制把用户和资源绑定在一个可信任环境下的实体。虚拟数字图书馆以一种虚拟的方式把一组资源聚合在一起，从而为一个固定用户群提供相应的数字图书馆服务。每个虚拟数字图书馆对应一个虚拟组织。

事实上，虚拟数字图书馆就是为用户定制的数字图书馆，是用户根据自己的个性化需求而创建的数字图书馆虚拟视图。这些视图隐藏了内容和服务空间的异质性，因而可以为用户提供高效的虚拟工作环境[7]。

2.2 虚拟数字图书馆的基础架构

作为一个建立在网格之上的数字图书馆实验平台，DILIGENT的目标之一就是动态地创建和管理数字图书馆[8]。为此，DILIGENT提出了如下图所示的虚拟数字图书馆基础架构。

DILIGENT基础架构包含了三类服务组件[9]：基础服务组件、数字图书馆内容服务组件和数字图书馆网格服务组件。基础服务组件是一组用来操作基础设施的具体服务组件，提供数字图书馆的配置和管理、过程管理、高级恢复代理、索引和搜索管理等服务。数字图书馆内容服务组件提供对内容和

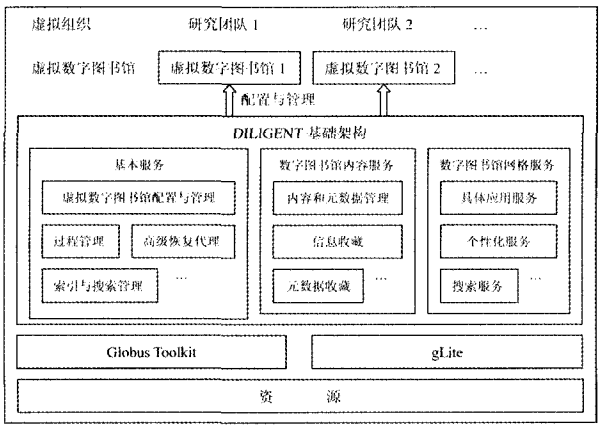

图1　DILIGENT的虚拟数字图书馆基础架构

元数据的管理服务，以及信息收藏与元数据收藏服务。数字图书馆网格服务组件则提供一组具体的数字图书馆服务，既包括对现有数字图书馆服务的网格封装形式，也包括为适应网格环境而实现的新服务（如分布式检索服务），还包括与具体应用有关的服务。

DILIGENT基础架构用来为虚拟组织创建虚拟数字图书馆。虚拟数字图书馆的定制不仅包括对内容和元数据的定制，也包括对信息处理工作流的定制。对工作流的定制必须在可用服务组件的基础上以动态方式建立。这需要两个面向任务的支持：一是为处理任务选择足够的服务组件，二是基于所选任务制定流程。后者可利用Web服务方面的已有成果来实现，前者则要求有适当的服务组件协调过程。

DILIGENT使用代表任务属性的分类体系来建立任务模型，以支持服务组件的协调。任务属性与一定的服务组件相关。用户可对每个任务指定其输入/输出需求、前置/后置条件和执行任务所需服务能力等属性。DILIGENT把这些属性转换为对应的分类体系。服务组件被描述为语法、语义和整体三个维度上的属性。对分类体系应用推理技术，就可为处理任务选择必要的服务组件。

2.3 DILIGENT的功能定位

DILIGENT作为数字图书馆的基础架构，可以看作是数字图书馆资源的提供者和消费者之间的服务代理[10]。数字图书馆资源的提供者是指决定要公布其资源的个人或组织。他们接受服务代理依照一定的访问和使用策略而进行的管理。数字图书馆资源的消费者是指想要建立自己的数字图书馆的用户团体。这个服务代理所管理的资源包括内容资源（通过单一入口可搜索和访问的信息仓库）、服务（实现某种特定功能且其描述、接口和约束均被定义并可公开获得的软件工具）和宿主节点（即具备计算和存储能力并为内容来源和服务提供环境的网络实体）。数字图书馆资源的提供者利用DILIGENT的适当机制对其资源进行注册和描述。DILIGENT还可从资源描述中自动推导出资源的其它属性，还可通过支持资源的发现、监控和使用以及实现其它一些控制共享和服务质量的功能，对注册资源进行管理。一个用户团体可以通过指定需求来创建一个或多个数字图书馆。这些用户需求指定

了信息空间的条件。

作为这样一个管理资源的服务代理，DILIGENT需要具备以下5个方面的功能：数字图书馆的创建和管理、内容和元数据管理、过程管理、索引和搜索管理以及具体应用方面的功能[11]。

数字图书馆的创建和管理功能组件负责为用户动态地构建和维护临时数字图书馆，并为用户自动识别和安排所需资源池。

内容和元数据管理功能组件实现对数字图书馆内容和相关元数据的处理、对标注的一致和分布的管理以及外部内容和元数据源的集成。

过程管理功能组件负责创建用户的信息处理过程：验证其正确性、按照可用资源和服务特点自动优化其定义并可靠地执行。可通过附加工作流来满足新用户的要求，从而方便地扩充DILIGENT系统的内容。

索引和搜索管理功能组件负责以合理的代价对数字图书馆中的信息进行检索。

具体应用功能组件提供支持用户具体场景（如门户、文档可视化等）所需的功能。

DILIGENT在eLite和GlobusToolkit基础设施之上集成数字图书馆服务，通过增加建立、操作和维护临时的虚拟数字图书馆这一功能来提升现有的网格服务。DILIGENT架构不仅把计算资源和存储资源连接起来，而且把构成数字图书馆的信息库、叙词表、本体、工具等所有资源都连接在一起，将网格技术所提出的共享概念进一步推广。DILIGENT采用面向服务的体系结构，其组件可以为其它eScience 应用所重用。

3 基于网格的分布式检索系统——Cheshire3

分布式信息搜索引擎Cheshire3是网格技术与数字图书馆技术结合的又一个具有代表性的实例。Cheshire3建立在SRB(Storage Resource Broker)网格基础之上，是由UC Berkeley和University of Liverpool合作开发的一个开源软件系统。该软件目前还不能算是真正的产品，其最新推出的版本是0.9版，可从其网站[12]上下载。

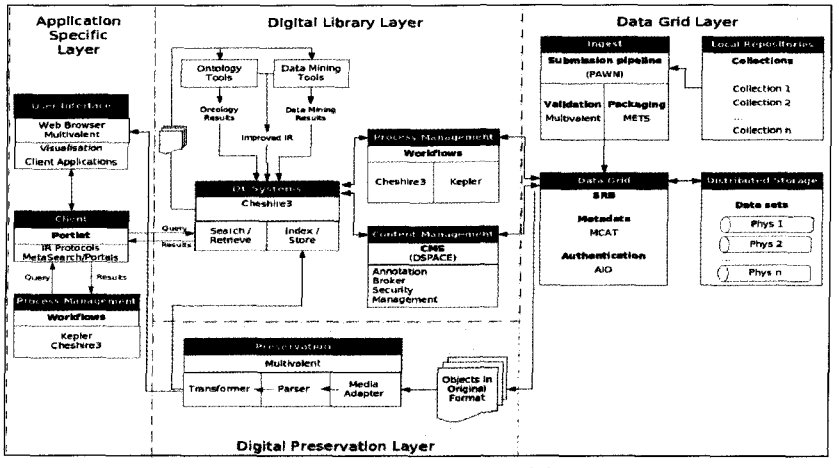

图2 Cheshire3的直接环境[13]

3.1 Cheshire3的环境

Cheshire3系统可以看作是基于网格的泛在图书馆环境中的一个元素[13]。如图2所示，Cheshire3的直接环境由数据网格层、数字保存层、数字图书馆层和具体应用层构成。

数据网格层利用San Diego Supercomputer Center（SDSC）开发的数据网格SRB来存储大量的分布式数据。

数字保存层主要由多格式文档分析器(Multivalent)构成。多格式文档分析器[14]用来把旧格式的文档引入到一个灵活的环境中，通过一个可插入媒体适配系统把原文档格式转变为内部模型，利用Web浏览器与文档进行交互，并允许内部模型转换为XML格式。这样就避免了旧格式文件的模仿（Emulate）和迁徙（Migrate）问题，既可以使用户直接看到原来的文件，同时又能使系统利用分析器从原来的文件中抽取所需信息。

Cheshire3作为搜索引擎，是数字图书馆层的一个组成部分。在数字图书馆层，扩展后的DSpace将以SRB作为底层存储来完成内容管理功能。这会使DSpace具有虚拟的无限收藏能力，而且通过统一的网格技术可对这些收藏进行存储、复制和访问。在SRB数据网格环境下，Cheshire可与DSpace等系统集成在一起，更为方便地发现资源，摄取数字内容，并完成内容的管理、分发和保存。

数字图书馆层的过程管理系统是多所大学或科研机构开发的工作流处理环境，以SDSC

的Kepler[15]为主体。在这样一个过程管理系统中，采用了主管/参与者模型，即参与者按照主管的要求一起执行任务。研究人员可在诸如Kepler之类的工作流环境中设计和执行灵活的处理流程以完成复杂的数据分析。工作流环境提供了图形化的用户界面，使得具有不同学科背景的不同层次的用户都可以通过拖放的方式来设计工作流。过程管理系统旨在提供一个可以把文本挖掘技术与方法论结合起来的平台。

3.2 Cheshire3的对象模型

Cheshire3系统的特点在于它定义并实现了一系列对象，这些对象具有明确的分工，使得数字图书馆的操作可以分布在网络中的多个节点上[16]。Cheshire3的对象模型如图3所示。

Cheshire3中将对象分为四类：

（1）数据对象：封装数据和相关元数据的对象，包括文档组（Document Group）、文档（Document）、记录（Record）、索引（Index）、用户（User）、查询（Query）、项（Term）等。在图3中，该类对象用灰色的矩形框表示。

（2）处理对象：对数据对象进行处理的对象，包括预分析器（PreParser）、分析器（Parser）、转换器（Transformer）、抽取器（Extracter）、标准转换器（Normaliser）等。见图3中的灰色椭圆形框。

（3）存储对象：对象的存储形式，包括配置、用户、记录、文档和索引存储。在图3中用磁盘状图形表示。

（4）抽象对象：对象的逻辑集合，包括服务器（Server）、数据库（Database）和工作流（Workflow）。见图3中的白色矩形框。

在上述对象中，除了用户给出的数据和处理得到的结果之外，其它对象都通过XML记录加以配置。这些配置信息包括对象的类型、标识符、参数（或其它允许对象元素完成其功能的必要信息）等，它们可以和普通的记录一样按照统一的方式被存储、获取和操作，并可通过OAI-PMH和

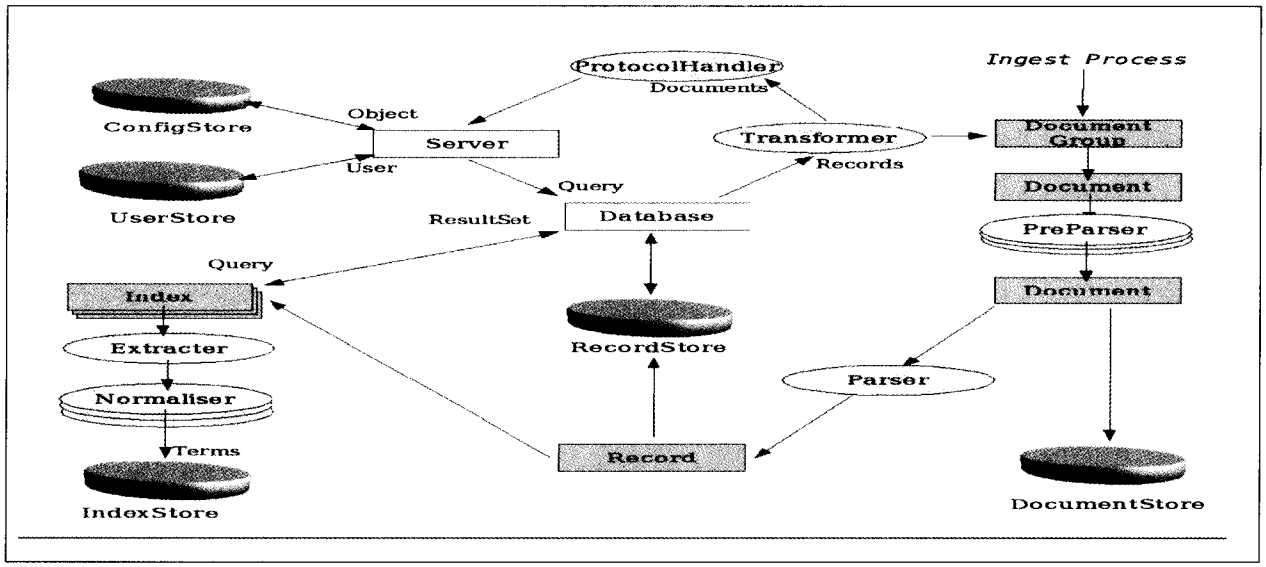

图3 Cheshire3的对象模型[16]

SRW/U之类的信息检索协议进行分发。这就使得Cheshire3网格可以方便地处理配置信息，从而无缝地共享资源。

DILIGENT和Cheshire3虽然从网格基础到实现目标都有着很大的差异，但它们之间又存在着一定的联系。Cheshire3作为一个基于网格的信息检索系统，实现了超越一般信息检索的低层功能，可以作为DILIGENT体系结构中的索引与检索部分。反过来，DILIGENT体系结构中的许多对象也可被看作是Cheshire3对象的混合体。由于有一个较低层次的标准，我们可以很容易地配置Cheshire3体系结构的实例以复制其它系统或实现与其它系统的互操作[17]，这使数字图书馆网格的扩展成为可能。

目前Cheshire3系统已在英国国家文本挖

掘中心(the UK National Text Mining Centre，NaCTeM)、NARA(The National Archives and Records Administration)、NSDL (National Science Digital Library)等机构得到初步的应用。

4 结语

本文通过研究国外两个基于网格构建泛在图书馆的典型实例，分析了网格和数字图书馆结合的优势以及所存在的问题。

网格技术为泛在图书馆的构建提供了一定的基础条件。在网格之上构建数字图书馆，可以提高数字图书馆的可复用性、灵活性和动态性，并能在一定程度上满足用户的个性化需求、为用户提供方便的虚拟工作环境。同时我们也应该注意到，网格提供的文档结构无法满足数字图书馆响应环境请求的需要，网格技术也难以处理数字图书馆特有的复杂工作流。本文中重点分析的DILIGENT和Cheshire3在解决这些问题方面做出了有益的探索。

DILIGENT试图在网格的基础上构建一个动态的管理资源和定制信息处理流程的支撑环境，Cheshire3则试图利用网格超越一般的信息检索功能，提高系统的灵活性和可扩充性。DILIGENT侧重宏观架构的搭建，Cheshire3则关注具体功能的实现。这两个系统从不同的角度为数字图书馆与网格技术的结合提供了范例，并为基于网格技术在一定的地域或领域范围内构建泛在图书馆环境提供了可供借鉴的模式和技术。

参考文献

[1] BAE K J, JEONG Y S, et al. The Ubiquitous Library for the Blind and Physically Handicapped—A Case Study of the LG Sangnam Library[C/OL]// Korea WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL, August 20—24 2006, Seoul, Korea. [2007—07—12]. http://www.ifla.org/IV/ifla72/papers/140—Bae—en.pdf.

[2] PATKAR V. e—Research and the Ubiquitous Open Grid Digital Libraries of the Future[OL]. [2007—07—12]. http://www.ifla.org/IV/ifla72/papers/140—Patkar_Chandra—en.pdf.

[3] 李伟.万丈高楼平地起——浅谈网格计算基础[OL]. [2007—07—12]. http://www.iipl.fudan.edu.cn/research/gc.html.

[4] DILIGENT Project Website. [2007—07—12]. http://www.diligentproject.org/index.php.

[5] CASTELLI D.Virtual digital libraries:The DILIGENT Project[OL]. [2007—07—12].http://www2.garr.it/conf_05_slides/d_castelli—DILIGENT.pdf.

[6] AVANCINI H, CANDELA L, PAGANO P, et al. DILIGENT,A Digital Library Infrastructure on Grid ENabled Technology: Test—bed functional specification[EB/OL]. [2007—07—12]. http://dlib.sns.it/pub/bscw.cgi/d17088/D1.1.1%20Test—Bed%20Functional%20Specification.pdf.

[7] CASTELLI D. Digital libraries [OL]. [2007—07—12]. ftp://ftp.cordis.lu/pub/ist/docs/rn/castelli.pdf.

[8] FORMMHOLZ I, et al. Supporting Information Access in Next Generation Digital Library Architectures [M]//Peer—to—Peer, Grid, and Service—Orientation in Digital Library Architectures. Heidelberg:SpringerLink, 2005: 207—222.

[9] NIEDEREE C, STEWART A, et al. Understanding and Tailoring Your Scientific Information Environment: A Context—Oriented View on E—Science Support [M/OL]// From Integrated Publication and Information Systems to Information and Knowledge Environments. Heidelberg: Springer, 2005: 289—298. [2007—07—12]. http://www.springerlink.com/content/qivb3yq7qud8yegj/.

[10] Castelli D . Digital libraries of the future and the role of libraries. [J/OL] Library Hi Tech Year: 2006, 24(4): 496—503.

[11] CASTELLI D, et al. DILIGENT: A Digital Library Infrastructure for Supporting Joint Research [C]// Local to Global Data Interoperability — Challenges and Technologies, June 20—24, 2005: 56—59.

[12] Cheshire3 Website. [2007—07—12]. http://www.cheshire3.org/.

[13] LARSON R R. Grid—Based Digital Libraries and Cheshire3 [OL]//[2007—07—12]. http://pnclink.org:8080/pnc2006/Presentation%20material/e—Science%20——%20Ray%20Larson.pdf.

[14] Multivalent Home Page. [2007—07—12]. http://multivalent.sourceforge.net/.

[15] Kepler Project Website. [2007—07—12]. http://kepler—project.org/.

[16] LARSON R R. Grid—Based Digital Libraries: Cheshire3 and Distributed Retrieval[C]//The 5th ACM/IEEE—CS Joint Conference on Digital Libraries (JCDL 2005). New York: ACM Press, 2005.112—113.

[17] SANDERSON R, LARSON R R. Indexing and Searching Tera—Scale Grid—Based Digital Libraries [C/OL]//The 1st international conference on Scalable information systems. May 29 — June 1, 2006, Hong Kong, Article No.3. [2007—07—12]. http://www.csc.liv.ac.uk/~azaroth/papers/infoscale2006.pdf.

作者简介

郭文丽(1968—)，中国科学院国家科学图书馆博士生，研究方向是“数字图书馆技术与系统”，已发文章两篇。通讯地址：北京市中关村北四环西路33号中国科学院国家科学图书馆 100080

(收稿日期：2007—06—18)

Study on Construction of Grid-based Ubiquitous Libraries

Guo Wenli / Graduation School of Chinese Academy of Sciences, Beijing, 100049 & National Science Library Chinese Academy of Sciences, Beijing, 100080

Abstract: Grid provides an infrastructure of ubiquitous libraries. The paper introduces two grid—based ubiquitous library prototypes, and analyzes the strengths and challenges of the integration of grid and digital library technologies, aiming to provide inspiration for construction of the grid—based ubiquitous libraries in certain geographical areas or in certain environment.

Keywords: Ubiquitous libraries, Grid, Virtual digital libraries, Information retrieve