●吴颖红（杭州师范大学 钱江学院图书馆，杭州 310012）

OAI协议与数字图书馆互操作性研究

[关键词] OAI协议；数据整合；网格；互操作性

[摘　要] OAI协议是近年来在数字图书馆界引起广泛关注的新技术，它具有简单性、开放性与灵活性等特点，可以很好地解决数字图书馆的互操作问题。文章通过对基本的OAI协议与数字图书馆互操作方案存在的一些问题的研究，着重论述了基于网格技术的OAI协议的数据整合方法及其与数字图书馆实现互操作框架的设想。

[中图分类号] G250.76

[文献标志码] A

[文章编号] 1005－8214(2009)01－0104－03

由于数字图书馆的“馆藏”资源具有动态性、异构性、分布性等特点，造成其主要功能即解决信息资源的有效利用在实践中难于实现。而目前，数字资源的整合方式有多种，OAI协议标准整合是其中一种重要方法，也是促进数字图书馆建设不断发展，使不同系统、不同数据拥有者实现资源的共享和互操作的检索机制。本文主要论述基于网格技术的OAI协议资源整合与数字图书馆的互操作性。

1 OAI协议简介[1]

OAI-PMH（Open Archive Initiative for Protocol Metadata Harvesting），简称OAI协议，是近几年在数字图书馆界引起广泛关注的新技术。它具有简单性、开放性与灵活性等特点，可以很好地解决数字图书馆的互操作问题。它通过定义一个标准的接口，使服务器能将其存储的元数据信息有选择地提供给外部应用程序服务器或其他服务器，也可以认为是解决不同资源的元数据互操作，有效挖掘、发布和利用互联网上数字信息资源的协议。目前，该协议可通过 http://www.openarchives.org/OAI/openarchivesporotocol.html 获取最新信息。

2 基于网格技术的OAI协议的数据整合与数字图书馆的互操作性设想

2.1 基本的OAI协议与数字图书馆互操作方案存在的问题[2]

当前，OAI协议与数字图书馆主要有两种互操作方案：分布式搜索（Distributed Search）和元数据收集（Metadata Harvesting）。

2.1.1 分布式搜索

该方法实时将用户提交的查询请求，转换成每一个DL可接受的形式，分别送往多个DLs站点执行，收集每个DL返回的结果，综合整理后交给用户。按是否遵循标准，分布式搜索分为两大类：基于标准的方法和基于数据驱动的方法。

由于分布式搜索方法依赖于实时地执行查询、处理查询结果，因此，对于数字图书馆节点比较少(一般来说是不超过20个)的情况下，该技术比较适用，但在互联网环境中，数字图书馆节点的数量都比较大(大于100)，利用分布式搜索技术来解决数字图书馆互操作问题就变得十分困难。

2.1.2 元数据采集方法

它的基本策略：从每个DL中采集并提取元数据，经过处理、合并后集中保存在一个元数据仓储中，用户对保存在元数据仓储中的元数据进行查询。该方法演变为著名的元数据采集框架OAI-PMH，为DLs的互操作问题提出了一种简单、可行的解决方案。OAI系统主要由数据提供者（Data Provider，DP）、服务提供者（Service Provider，SP）组成。DP将自己拥有的元数据用公共元数据格式（Dublin Core）表达，并通过OAI协议提供统一的标准化接口，向外部揭示自身的元数据。SP则通过OAI协议获取数据提供者的元数据，并以这些元数据为基础为用户提供进一步的信息增值服务。该方法有效地解决了各资源库在元数据格式上可能存在的异构性问题，实现跨资源库检索。

目前，基于OAI协议的Harvesting的联邦搜索是数字图书馆界研究与开发的热点，一些著名的DLs项目，如NDLTD和NSDL均采用此方法作为互操作的解决方案。由于OAI比较新，所以有些元数据收集的重要问题尚未涉及，有一定的局限性。如：没有规定如何选择数据源，没有强调如何实现服务提供者，互操作框架存在元数据的同步更新问题等。

2.2 基于网格技术的OAI数据整合与数字图书馆互操作框架的改进构想

由于数字图书馆信息资源具有分布性、异构性、自治性等特征，因此在保持各数字图书馆原有结构、标准不变的基础上，结合OAI-PMH框架，构建基于网格技术的数字图书馆互操作框架——数字图书馆网格DL Grid（Digital Library Grid），利用网格采集、组织和存储元数据，通过元数据的互操作，屏蔽数字图书馆的异构性，完成资源整合，[3] 为用户提供一个统一、透明、高效的信息检索平台，从而实现数字图书馆间的互联及信息资源的共享。

2.2.1 网格技术解决OAI协议的数据整合方法[4]

目前，数字资源的整合方法主要有以下几种：门户整合、数据库整合、系统整合、检索方式的整合、协议标准整合。所谓的协议标准整合是针对各种不同的数据组织方式和网络通信协议而言的，即通过以上的中间技术手段或者完全对数据进行重组的手段，对采用不同访问协议和不同数据标准的数据库在同一界面内实现集成检索或者整合检索，从而达到资源整合的目的。主要目的是为检索提供数据，也是整个检索系统的基础。OAI数据整合与数字图书馆互操作性最终通过索引部分和检索部分来实现。利用OAI协议标准整合数据资源从而实现与数字图书馆的互操作是数字图书馆的发展趋势。OAI数据资源整合系统可以将需整合数据统一在检索平台中，与其他数据库检索以及非标准数据库检索处在并列的位置并接受集成平台的查询要求，负责对所属数据库进行查询，将结果返回给集成平台，集成平台将这部分数据进行整合后最终结果返回给用户。互操作性问题是数字图书馆的一个关键性问题，为了更好解决DLs互操作中出现的问题，需要建立新的框架体系结构。OAI-PMH是利用Harvesting概念建立的典型的元数据采集框架，通过元数据的互操作，实现数字图书馆的互操作，克服了分布式搜索无法解决的规模问题。而网格技术显著之处在于关注大规模的资源共享与数据整合，强调多机构之间大规模的资源共享和合作使用，提供了资源共享的基本方法。将数据整合系统与元数据采集harvesting方法相结合是基于OAI协议与数字图书馆实现互操作的新构想，本文试图提出一种增强数字图书馆互操作的新框架，利用网格技术更好地解决DLs资源发现、整合、跨仓储检索、安全等问题，克服传统DLs互操作方案的局限性，支持大规模DLs信息资源共享。

2.2.2 基于网格技术的OAI数据整合系统与数字图书馆互操作框架

3层体系框架的实现设计如下：[5]

（1）数据资源层：即信息资源提供层，由广域分布的DLs组成，构成整个数字图书馆的信息提供者。它位于整个框架的最底层，其体系结构采用分布式对等结构。它的网络结构在网格技术（如p2p技术）支持下可实现一系列优化：提高资源利用率；可充分使用闲置宽带资源、信息资源等；提高搜索深度；克服单点失效，实现负载均衡；克服时滞问题。P2P在搜索中，各对等点之间动态而又实时互联的关系，使得搜索可以在对等点间实时进行，提高可扩展能力。含超级节点的对等网络结构借鉴了集中式对等网络和完全分布式对等网络的优点，克服了由中央服务器和网络流量的限制造成的可扩张性差异。

（2）OAI中间层：即信息资源整合层，它是利用开放的网格技术和OAI-PHM协议，屏蔽资源层中DL的分布、异构特性，实现元数据的发现、采集、组织、存储等功能，通过资源整合提供本层透明、一致的接口。也即完成元数据整合并实现检索的过程。基于网格技术的OAI层子框架信息集成平台是用户与数据源间的中间件，由数据提供者、服务提供者和注册模块几部分组成，它至少完成以下三个任务：① 将用户的查询按具体查询协议请求转换成不同的格式发送到数据源。② 将数据源返回的结果标准化。③ 合并/筛选标准化后的结果。信息集成平台接收用户的查询请求，组织成XML格式的查询条件进行查询并集成查询结果，同时进行查重、分页等一系列处理后返回给用户。数据交互采用的文档格式是基于XML的，因为采用此格式所包含的内容可由用户自定义，并且可扩展性好，这样不同子系统检索返回的结果样式可以实现统一，便于信息集成平台的处理。

（3）应用服务层：即信息资源互操作层，在集成DLs元数据的基础上，通过单一的服务接口，为用户提供增值服务，如文献检索、个性化服务、参考咨询等。它是信息资源互操作的实现过程，分索引部分和检索部分。索引的过程，其实是应用服务层的形成过程，实现的途径是由数据提供者通过元数据创建模块将要分布的信息转换成数字对象入数据仓储，将之转换成符合OAI协议规范的元数据格式，并进行结构化处理，完成元数据收集过程，生成用于检索的应用层。检索的过程则是应用服务层的使用过程，也是整合资源与数字图书馆的互操作的实现过程。某节点将

信息查询发送给其所在区域的超级节点，应用服务层面根据用户权限和输入内容与用户进行交互，选择恰当内容或主题检索。如无内容无资源信息，则通过定向广度优先方式转发搜索请求，这样可降低对网络宽带的消耗，又有效节约检索时间。如找到资源信息，则将所有满足需求的节点传给发出查询请求的节点，并对各节点进行有效的延时最小选择，与其直接建立链接，进行点对点下载，从而提高信息传递的速度；如交互结果不能使之满意，则再类推选择延时其次小的节点，建立链接并下载。

OAI协议作为元数据提供与采集协议，从语法上保证不同的数字图书馆的元数据交换与共享，并可以从语义上保证同一领域的语义可互操作，各种资源通过网格互联，利用高性能的Grid计算节点，增强收集和索引的动态性能，加快元数据的更新速度，提高服务提供者的质量，也克服了原有OAI-PMH框架的局限性。因此，基于网格技术的DLs互操作体系结构——DLGrid，在支持互操作上更加有效。

3　结语

OAI协议支持对各种数据库进行整合检索。该协议是近几年来引起广泛关注的新技术，很多国外数据库都支持该协议，针对它的研究对于国内数字图书馆建设具有现实意义。其资源整合部分的研究，也是目前数字图书馆研究中的热门课题。

本文结合国内外研究现状，通过对OAI协议标准整合方法的调查和参考，分析研究了目前广泛应用于数字图书馆的OAI协议，并由此提出了实现互联网上大规模的数字图书馆互操作的可行性，同时也试图用现代化网络技术与OAI-PMH框架相结合，提出两者的框架体系结构，并对网格环境下元数据的发现、采集、组织和传送等关键技术进行了分析研究，通过原型系统初步实现了在集成的元数据基础上数字图书馆信息的共享，为解决数字图书馆的互操作问题提供了一种新的设想和方法。

[参考文献]

[1] The Open Archibes Initatibe Protocol for Metadata Harvesting [EB/OL]. [2008－01－17]. http://opernarchives.org.

[2] 郑志蕴，等. 网格环境下基于OAI的数字图书馆互操作机制 [J]. 计算机工程，2006 (5)：37—39.

[3] 董慧，丁波涛. 用OAI—PMH协议解决数字图书馆互操作问题 [J]. 情报科学，2004 (6)：609—702.

[4] 王权良. 数字图书馆OAI数据资源整合系统的研究与实现 [D]. 北京：北京交通大学，2006：24—26.

[5] 夏立新，王忠义. 基于OAI合主题图的分布式数字图书馆体系框架 [J]. 现代图书情报技术，2007 (12)：11—15.

[作者简介] 吴颖红（1971—），女，杭州师范大学钱江学院图书馆副馆长，副研究馆员，已发论文16篇。

[收稿日期] 2008－03－17　　[责任编辑] 陈永平

(上接第54页)

[参考文献]

[1] 孟小峰. Web信息集成技术研究 [J]. 计算机应用与软件，2003 (11)：33.

[2] T Berners-Lee，J Hendler，O Lassila. The Semantic web [J]. Scientific American，2001.

[3] Gruber T R. A translation approach to portable ontology specifications [J]. Knowledge Acquisition，1993(5)：199—220.

[4] 岳昆，王晓玲，周傲英. Web服务核心支撑技术：研究综述 [J]. 软件学报，2004 (3)：435.

[5] David Martin，et al. OWL-S：Semantic Markup for Web Services[EB/OL].[2007－12－18]. http://www.w3.org/Submission/2004/SUBM-OWL-S-20041122/.

[6] 管强，张申生，杜涛. 基于Web服务的电子政务应用集成研究 [J]. 计算机工程，2005 (6)：34.

[7] 赵新力，等. 电子政务主题词表编制规则 GB/T19486—2004 [S]. 北京：中国标准出版社，2004.

[8] 杜小勇，李曼，王大治. 语义Web与本体研究综述 [J]. 计算机应用，2004 (10)：15—16.

[9] Web Service Composer [EB/OL]. [2007－12－25]. http://www.mindswap.org/2005/composer.

[作者简介] 孟祥宏（1971—）男，呼伦贝尔学院副教授，中国人民大学信息资源管理学院2006级博士生，研究方向为电子政务、网络安全。

[收稿日期] 2008－03－22　　[责任编辑] 陈永平