●李广建 汪语宇 张 丽

# 数字资源整合的实现机制及关键技术*

## ——对国外数字资源整合系统的实证研究

**摘 要** 在对国外 143 个整合系统进行统计研究的基础上,探讨数字资源整合系统的实现机制。它们可以被概括为数据仓库整合机制、中介器封装器整合机制、代理整合机制和对等整合机制。每种整合机制都涉及一些主要技术。图 3。参考文献 16。

**关键词** 数字资源 整合机制 数据仓库 中介器 封装器 代理 对等网

**分类号** G250.76

**ABSTRACT** Based on a statistical analysis of 143 integration systems in foreign countries, the authors discuss the realization mechanisms of integration systems of digital resources, which can be summarized as the integration mechanism of data warehouses, the one of mediators and wrappers, the one of agents and the one of peer-to-peer. Each integration mechanism requires some major technologies. 3 figs. 16 refs.

**KEY WORDS** Digital resource. Integration mechanism. Data warehouse. Mediator. Wrapper. Agent. Peer-to-peer.

**CLASS NUMBER** G250.76

数字资源整合是近年来数字图书馆领域的研究热点,而数字资源整合实现机制又是热点中的重点。不同的整合实现机制会导致整合系统在结构、模块组成和所采用技术等方面出现差异,也会直接影响整合系统的效率。

本文采用实证的方法,对国外现有的数字资源整合系统进行分析,着重探讨它们的实现机制。

## 1 国外数字资源整合系统概况

近年来,世界各国的学者和研究机构相继开发出了一系列数字资源整合系统,瑞士苏黎世大学信息学院整理、报导了 175 个整合项目和相关系统[1],笔者又进一步查阅文献,找到了另外 35 个项目和相关系统,两项合计并除去其中 67 项局部整合技术研究,共有 143 个项目研发有完整的系统。它们基本反映了现阶段国外数字资源整合研究的最新进展,揭示了整合研究的发展方向。

从这 143 个系统的地域分布看,美国是整合研究领域的主要力量,研发的系统占到 51.7%,其次是德国、法国、英国、意大利、瑞士等欧洲国家,它们研发的系统分别在 10% ~50% 之间,加拿大、西班牙、希腊、巴西等国也有一定的研究。美国、英国、法国、瑞士等国家的一些大学和科研机构,长期致力于数字资源整合研究,形成了比较深入且连续的理论体系,研发了较多的系统,并广泛应用于实践。这些国家的主要研究团体包括:美国的斯坦福大学(开发有 IDIMS, Infomaster, MedMaker, TSIMMIS 等系统)、华盛顿大学(开发有 BioMediator, Piazza, Tukwila 等系统)、南加州大学(开发有 Ariadne, Prometheus, SIMS 等系统)、马里兰大学(如 MOCHA, WebSemantics)、英国的谢菲尔德大学(开发有 LASIE, MELITA 等系统)、法国国家信息与自动化研究院(开发有 Active XML, Agora, DISCO 等系统)、瑞士的苏黎世大学(开发有 SINGAPORE, SIRUP 等系统)等。多国间的合作也是研究的一个特色,占 4.2%。不过,跨国研发多限于美国和欧盟国家。

143 个系统中,最早的系统出现于 20 世纪 80 年代初期,主要应用于分布式异构关系数据库的整合,目的是屏蔽各个数据库结构、组织方式等方面的差别,为用户提供访问资源的统一接口,法国国家信息与自动化研究院开发的 MRDSM 系统就属于这 4 种类型。

20 世纪 90 年代以来,随着 Web 的普及,数字资源的范围逐步扩展至网络数据库、网页、文件等结构化、半结构化和非结构化信息资源,数字资源的异构

* 本文系国家社会科学基金项目"Web 整合的机制及方法研究"(05BTQ024)的研究成果之一。

性、分布性特点更加突出。网络数字资源的整合成了研究的重点和主流。143个系统中,90%以上的整合系统都属于网络数字资源整合系统。

## 2 数字资源整合系统的实现机制

根据我们的分析,目前数字资源整合系统的实现机制可以概括为数据仓库整合机制、Mediator/Wrapper整合机制(中介器封装器整合机制)、Agent整合机制(代理整合机制)以及P2P整合机制(对等网整合机制)4种类型。

数据仓库整合机制是较早提出的一种数字资源的物理集成方法,在早期的系统和当前系统中均有应用,采用这种机制的系统有21个,占14.7%。中介器封装器整合机制是目前数字资源整合系统的主流实现方式,有99个系统采用,占69.2%。代理整合机制是一种新的整合机制,能够有效增加数字资源整合系统的灵活性,提高系统效率,已有14个系统采用,占9.8%。对等网也是近年来出现的一项新兴技术,因灵活性、适用性强等优点受到数字图书馆界的广泛关注,是近年来数字资源整合研究领域的一个新热点,采用这种机制的系统有9个,占6.3%。

### 2.1 数据仓库整合机制

数据仓库整合机制是一种物理集成方式,它将不同来源的数字资源按特定的方式(通常是按主题或其他多维方式)建模并存储在同一物理位置(称为数据仓库),提供给用户一个新的、统一的目标数据模式,使得用户能够一站式地访问各种数字资源,从而达到整合目的。

数据仓库整合机制最根本的特点是在同一物理位置存放数字资源,集中管理不同来源的数字资源,简化了用户访问信息的复杂度,提高了数字资源的访问速度和整合系统的性能。而且由于实现了不同来源的数字资源的一致性存储,这种整合机制还有利于实施比信息检索更复杂、更深入的数据挖掘、知识发现等服务。采用数据仓库整合机制进行整合的前提是必须能合法地(例如通过授权)获得来源系统中的数字资源。但因为是集中存储要整合的数字资源,所以难以适应网络数字资源类型多样、变化快等特点,还会增加本地系统存储与维护的负担。

法国、德国联合开发的Xyleme、美国斯坦福大学的WHIPS、美国乔治亚大学的InfoHarness等数字资源整合系统都是使用数据仓库整合机制的代表性系统。数据仓库整合机制涉及的技术主要有:

(1)海量数据存储。解决海量数据的存储问题,除了需要相应的专门设备如磁盘阵列、光盘库、磁带库等,还需要精心设计存储结构和存储算法,既要保证数字资源有合理的物理存储结构,又要保证有较快的存取速度。

(2)ETL技术。它是对要加以整合的数字资源进行抽取(Extract)、转换(Transform)、清洗(Cleaning)、装载(Loading)的技术。对不同来源的数字资源进行物理集成,首先需要从参与整合的系统中抽取出相关数字资源,这需要使用信息抽取技术。由于信息源是异构的,还必须利用转换技术将不同结构的数字资源规范化,消除异构数字资源之间的不一致性,为来自不同系统的数字资源之间的比较、整合以及统一存储奠定基础。清洗技术主要是解决信息冗余的去重以及错误和不完整信息的修正、剔除问题。装载技术则是将清洗后的数字资源按一定的规则加载至数据仓库,形成数据仓库的物理存储结构和逻辑存储结构。

(3)信息源的监控与更新。数据仓库本身与信息源在物理上是分离的,必须要解决数据仓库与信息源的同步问题。必须要监控参与整合的信息源的变化,同步更新数据仓库,确保用户在整合系统中查询到的是各个信息源中的最新数据资源。

### 2.2 中介器封装器整合机制

这是一种虚拟整合方式。在基于这种机制的整合系统中,并不真正存储需要整合的数字资源,而是通过中介器和封装器来实现整合。它们均为软件组件,位于用户和数据源之间,中介器负责处理用户提问和查询结果的整合,封装器负责对信息源的连接和具体查询。

该整合机制的基本原理如图1所示。在基于这种机制的整合系统中,用户按全局模式(Global Schema)进行查询,中介器接收用户查询并将之转换成中间格式,然后提交给相应的封装器,封装器进一步将中间格式的查询转化为信息源模式或本地模式(Source/Local Schema)的查询,并与参加整合的相应

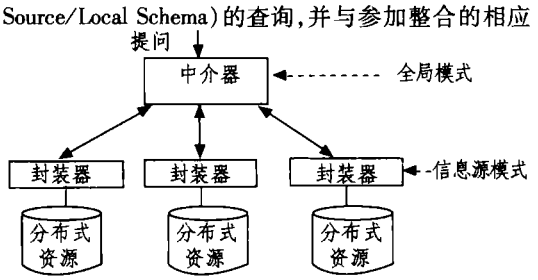


**图1 中介器封装器整合机制的基本原理**

信息源进行连接,实现对相应信息源的查询,将查询结果返回给中介器,中介器对结果进行处理,以统一的形式提供给用户。

与数据仓库整合机制相比,中介器封装器整合机制能够有效保持各个异构信息源的自治性,满足局部的应用,并且能够充分发挥中介器的作用,满足全局性应用。使用这种机制的整合系统不需要在本地储存大量的数字资源,能够适应网络环境下信息源高度自治、数量多、更新变化快等特点。在中介器中引入本体论等语义相关技术后,能够有效解决知识整合、个性化服务等问题。因而该整合机制是目前实现数字资源整合的主流方式。

中介器封装器整合机制中,主要通过 GAV,LAV,GLAV 和 BAV 等 4 种方式实现全局模式与信息源模式或本地模式之间的映射。

GAV(Global as view)映射方式根据信息源模式(本地模式)来定义全局模式,以全局模式为中心,全局模式中的关系被定义为信息源模式(本地模式)中关系的视图。采用 GAV 映射方式的整合系统在提问式转换的过程中,只需将用全局模式表示的用户提问进行视图展开,就可以得到用信息源模式(本地模式)表示的子提问,整个过程比较简单。GAV 映射方式是目前应用最为广泛的一种映射方式[2]。美国斯坦福大学开发的 TSIMMIS 系统、英国曼彻斯特大学开发的 TAMBIS 系统、德国汉堡大学开发的 SQXML 系统、意大利罗马大学开发的 IBIS 系统等均使用了 GAV 映射方式。

LAV(Local as View)映射方式与 GAV 恰好相反,它根据全局模式来定义信息源模式(本地模式),以信息源模式(本地模式)为中心。全局模式涵盖了参与整合的所有信息源模式(本地模式),一个具体的信息源模式(本地模式)是全局模式的一组视图。采用 LAV 映射的整合系统在提问式转换的过程中,必须将根据全局模式定义的查询式进行重组,称为"利用视图重写查询式"(rewriting queries using views),整个过程比较复杂[3]。使用 LAV 映射方式的数字资源整合系统主要有法国国家信息与自动化研究院的 Agora、法国巴黎第四大学的 PICSEL、美国华盛顿大学的 Razor 等。

GLAV(global - local - as - view)映射方式混合使用 GAV 和 LAV,通过定义全局模式与信息源模式(本地模式)之间的语义映射,能够在与信息源模式(本地模式)无关的情况下灵活定义全局模式[4]。一般地说,GLAV 可看做是 LAV 的变体,它实现了表达能力和查询的易处理性之间的最佳折衷,并保留了 LAV 在可扩展性方面的优势,近年逐渐受到研究人员的重视,越来越多地应用于整合系统的研发。美国加州大学的 MARS 系统和意大利罗马大学的 DIS@DIS 系统均使用 GLAV 映射方式。

BAV(Both as View)映射方式使用双向模式变换,建立全局模式与信息源模式(本地模式)之间的转换规则,通过转换规则,既能够从信息模式(本地模式)中抽取出全局模式的定义,也能够从后者中抽取出前者的定义。BAV 可以同时支持信息源模式(本地模式)和全局模式的动态变化[5]。BAV 是近年来提出的新的整合模式,理论上优点明显,但实现技术较复杂,还未得到广泛应用。目前只有伦敦皇家学院开发的 AutoMed 采用 BAV 映射方式。

中介器封装器整合机制涉及的主要技术有:

(1)信息源选择技术。中介器封装器整合机制是一种虚拟整合方式,整合系统本身并不存储被整合的数字资源。如果将用户查询不加区别地发送给参与整合的所有信息源,必然占用较多的带宽并增加系统的负担。这就需要利用信息源选择技术来确定相关度高的信息源,以提高整合系统的效率。信息源选择技术主要包括信息源描述和信息源选择两个方面,前者是按一定的算法建立对各信息源的描述模型,后者是在信息源描述模型的基础上,根据用户查询,按一定算法选出相关度高的信息源作为查询对象[6-7]。

(2)信息抽取技术。它应用于整合系统的目的是将参与整合的半结构化、非结构化信息源中的数字资源转化成结构性更强、语义更清晰的格式,以提高查询速度[8]。信息抽取技术已成为生成封装器的关键技术之一,广泛应用于面向网络数字资源的整合系统中。

(3)查询处理技术。这是对查询进行检验、重构、优化的技术。用户对整合系统查询时,整合系统需首先对用户查询进行语法分析和检验,确保查询符合系统全局模式的要求,这部分工作主要由查询检验技术完成。此后,整合系统再将经过检验的查询按一定的规则转换为面向不同信息源模式(本地模式)的多个查询,这个过程称为查询重构。整合系统面对的是不同的信息源,由于各个信息源有其自身的特点,例如有不同的传输带宽和传输延时,加之当前运行情况有不确定性,如当前的信息源访问量是大还是小、信息源是否能很快与整合系统建立连接等等,这就要

求整合系统根据各信息源的当前运行情况制定最优查询计划和查询调配方案,并据此对各信息源进行查询,这个过程称为查询优化[9~11]。

(4)结果整合技术。整合系统提供对各信息源的一站式访问,因而对信息源访问完毕后需要用中介器对来自不同信息源的结果信息进行整合及合成,以统一的形式呈现给用户。结果整合主要是对不同信息源的结果做并操作,并且重新计算结果的相关度。一般地说,对一个信息源进行查询会形成一个相应的查询子视图,结果整合的目的就是将这些子视图连接起来,形成一个完整的视图,提供给用户。

(5)语义整合技术。随着整合研究理论和实践不断深入,Ontology、语言建模、机器学习等语义相关技术逐渐应用于数字资源整合。语义整合技术被用来在系统运行期间获取和处理数字资源的意义及其之间的关联,使整合系统能够建立用户提问与各信息源之间的语义联系,消除各种数字资源的异构性,并能够将分散存储、表现形式不同的信息源中的有用资源进行再组织,真正满足用户的信息需求,从而提高数字资源整合的质量[12]。

2.3 代理整合机制

代理整合机制的基本原理如图2所示。这种整合机制中,使用了三类基本的Agent:用户Agent、资源Agent和代理方Agent。用户Agent负责维护用户信息,并提供系统接口,以方便用户与整合系统进行交互。资源Agent负责对分布式资源进行处理,将数字资源按照整合系统的表示形式进行描述和转换。代理方Agent负责将从用户Agent发出的查询请求与所要查询的资源Agent进行匹配[13]。

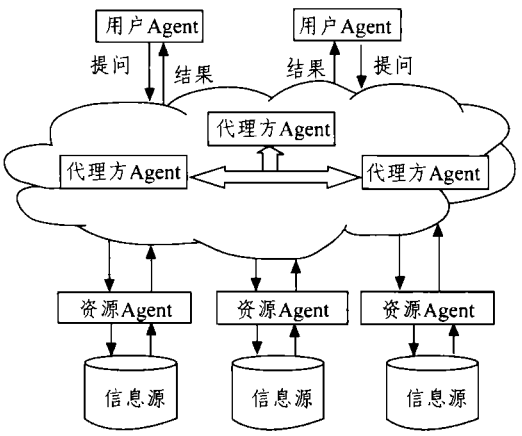


**图2 Agent机制的基本原理**

代理整合机制的优点在于能够有效利用Agent的特性来提高系统的整合效率。首先,Agent的自主性和移动性使得整合系统能够主动适应网络环境的变化,增强了整合系统的灵活性,整合系统能更加适应数字资源分布性及异构性的特点。其次,Agent能够在非连续运行的网络环境中运行,因此Agent还可处于移动计算环境中,这使得各种移动设备(如PDA等)也能加入使用整合系统当中[14]。Agent机制的这些特点,使它成为近年来整合研究的热点之一。

目前采用Agent机制的典型整合系统包括美国德克萨斯州奥斯汀微电子和计算机技术公司(MCC)开发的InfoSleuth、美国德克萨斯技术大学开发的AgentRAIDER、意大利摩德纳大学开发的MIKS等系统。在Agent整合机制中,如何使多个Agent协调工作,是采用这种机制的整合系统要解决的关键技术,具体地说,包括:

(1)Agent间的通信技术。为了达到整合目的,提高整合效率,需要通过Agent通信技术来实现Agent间的"会话"。一般来说,Agent间的通信是通过Agent通信语言(ACL)来实现的,Agent通信语言用于描述相应Agent的状态和属性,定义Agent可以交换的语法和语义消息。这种包含语义信息的通信语言不仅有利于协助Agent之间进行互操作,还有利于进行语义层次的整合[15]。

(2)Agent协调技术。在Agent整合机制中,多个Agent作为一个整体而存在,虽然每个Agent的任务有所区别,但它们作为一个整体,具有共同目标,就是实现整合。在任务执行的过程中,需要应用Agent协调技术来管理一个或多个Agent行为之间的从属关系,避免执行时发生冲突,所要解决的问题包括组织结构、任务分解、资源分配、群组决策、冲突发现与解决等。

2.4 P2P整合机制

P2P(Peer-to-peer)是近年来兴起的一种新的计算模式,它能够使PC和其他非服务器计算实体以对等的方式联网,彼此共享对方的资源。其主要特点是支持互连主机的动态变化。

P2P整合机制的基本原理如图3所示,其中,存在有多个分布式的对等点(peer),每一个对等点都拥有一套自己的数据模式(对等点模式)。整合过程中,通过对等点模式与本地资源模式的映射,实现对本地资源的访问,同时依靠P2P映射来完成对等点之间的模式转换,实现对等点间的通信。通过这种方式,在任何一个对等点中执行的查询也均可以在其他

相连的对等点中执行,从而达到有效访问各分布信息源的目的。

P2P 整合机制不仅能够实现大规模数字资源的集成,而且可以实现 Web 资源的动态整合,使整合系统具有强大的扩展性,是一种比较有生命力的整合机制。但由于学术界对 P2P 整合机制的研究处于起步阶段,目前理论研究较多,实际应用系统数量还不太多。在我们调研的 143 个系统中,有 9 个系统采用了 P2P 整合机制,其中较有影响的系统是美国加州大学开发的 RACCOON 系统,可以在加州大学网站上免费获取该系统的源代码。

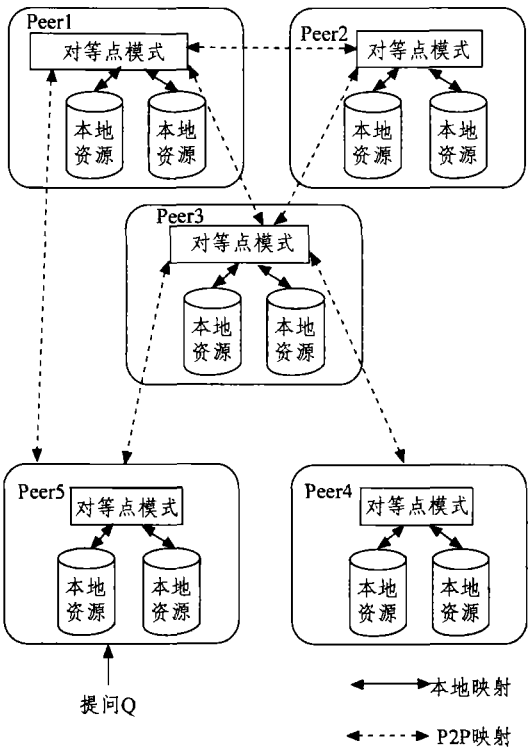


**图 3　P2P 整合机制的基本原理**

P2P 整合机制的关键技术是 P2P 映射以及对等点的发现与搜索。

(1)P2P 映射建立技术。在 P2P 整合机制中,由于每个对等点的模式不同,需要在对等点模式之间建立映射。在对等点中,以对等点模式作为处理对象,无需建立和维护单一的全局模式。相对于 GAV,LAV,GLAV 和 BAV 中的模式映射而言,P2P 映射是比较简单的,容易从系统中增加和删除,并且不影响整合的效率。

P2P 映射的建立一般包括两个步骤[16]。第一步为模式匹配,即在需匹配的模式间寻找能够标识出模式中的相同或相似元素的对应关系,这种对应关系一般是指对元素相似性的描述,基本上不包含语义信息。第二步,通过对应关系,利用一系列自动化技术,在人工干预下,建立精确的 P2P 映射。

(2)P2P 对等点的发现与搜索技术。在整合过程中,由于 P2P 网络中存在多个对等点,每个对等点存储有不同的数字资源,因此需要针对具体的用户需求,利用发现策略、搜索算法等相关技术,对 P2P 资源进行搜索,找出合适的对等点,并通过多个对等点的协作来集成资源。目前在 P2P 对等点的发现与搜索中应用较多的是分布式哈希列表(DHT)技术。这种技术使用分布式哈希算法来解决结构化的分布式存储问题,DHT 中存储有每个对等点的相关信息,通过 DHT 可针对具体需求获取所需对等点的信息,从而解决了对等点的发现问题,然后,再根据基于 DHT 的路由算法完成对等点的搜索。

## 3　结语

整合机制是数字资源整合的核心问题。随着整合研究理论和实践的深入,数字资源整合的机制不断发展;新技术的应用,导致了新的整合机制的出现,同时,每一种整合机制自身也在持续地发展和完善,通过引入新的技术来提高自身的效能,这些都值得我们关注。

**参考文献**

1　Data integration projects world - wide. [2005-07-30]. http://www.ifi.unizh.ch/dbtg/Staff/Ziegler/IntegrationProjects.html

2,3　A. Y Levy. Logic-based techniques in data integration. Logic-based artificial intelligence. Kluwer Academic Publishers,2000:575 ~ 595

4　M. Friedman. A. Y Levy, T D. Millstein. Navigational plans for data integration. In: Proc. of the National Conference on Artificial Intelligence,1999:67 ~ 73

5　P. McBrien. A. Poulovassilis. Data Integration by Bi - Directional Schema Transformation Rules. In: Proc. of ICDE '03, March 2003

6　张丽,汪语宇. Web 整合中的资源描述技术. 图书情报工作,2005(10)

7　汪语宇,张丽. 集成检索系统中资源选择技术及算法. 图书情报工作,2005(10)

8　李保利,陈玉忠,俞士汶. 信息抽取研究综述. 计算机工程与应用,2003(10)

9 Buneman P. Semistructured data. In proceeding of the Sixteenth ACM SIGACT - SIGMOD - SIGART Symposium on Principles of Database Systems, Tucson, Arizona. 1997: 117 -121

10 陈义. 面向数据集成的数据复制和查询优化. 中国科学院研究生院博士学位论文,2004

11 Mourad Ouzzani, Athman Bouguettaya. Query Processing and Optimization on the Web. Distributed and Parallel Databases. 2004. 15. DD:187,218

12 林岳等. 现代语义技术及其应用. 计算机应用研究, 2005(6)

13 Anastasiya Sotnykova. Design and Implementation of Federation of Spatio - temporal Databases: methods and tools. [2006-06-02]. http://lbdwww. epfl. ch/e/research/amber/BFR99__057_RepVl. pdf.

14 Domenico Beneventano, Sonia Bergamaschi, Gionata Gelati, Francesco Guerra, Maurizio Vincini. MIKS: An Agents Framework Supporting Information Access and Integration. [2006-05-21]. http://www. dbgroup. unimo. it/prototipo/paper/miks. pdf

15 Matthias Klusch. Information Agents technology for the Internet: A survey. Data&Knowledge Engineering,36(2001)

16 Igor Tatarinov, Zachary Ives, Jayant Madhavan, Alon Halevy, Dan Suciu, Nilesh Dalvi, Xin (Luna) Dong, Yana Kadiyska, Gerome Miklau, Peter Mork. The Piazza Peer Data Management Project. [2006-04-08]. http://www. cis. upenn. edu/~zives/research/piazza-sigmod-record. pdf

**李广建** 北京师范大学管理学院教授,博士生导师。通信地址:北京师范大学管理学院。邮编100875。

**汪语宇 张丽** 北京师范大学管理学院硕士研究生。通信地址同上。 (来稿时间:2006-07-07)

# 国家图书馆重大科研项目招标的公告

国家图书馆基于"人才兴馆"、"科技强馆"和"服务立馆"三大发展战略的需求,特设立"国家图书馆重大科研项目",面向馆内外招标。国家图书馆将遵循"公开、公平、公正"的原则,组织专家进行评审,择优立项支持。

## 一、招标内容

1. 国家图书馆数字战略研究;

2. 社会公共服务体系中图书馆的发展趋势、定位与服务研究。

内容详见《国家图书馆重大科研项目申报指南》。

## 二、投标要求

凡具有独立法人资格的高等院校、科研单位和企事业单位均可成为重大科研项目的课题承担单位,课题的申报单位既可以是独立法人单位单独申报,也鼓励多家单位联合申报。

申报所需的各种材料(包括《国家图书馆重大科研项目申报指南》、《国家图书馆科研项目申请书》等)可从国家图书馆科研支撑平台(网址:http://srsp. nlc. gov. cn)下载。申报者要如实填写申请材料。申请书要求一律用计算机填写、A4纸印制,经所在单位审查盖章后,于2007年4月30日前将申报材料一式三份报送国家图书馆科研处。

## 三、联系方式

申报材料报送地址:北京市海淀区中关村南大街33号国家图书馆科研处

邮政编码:100081

联系人:姚迎

联系电话:010 - 88545186

E-mail:yaoying@ nlc. gov. cn

**国家图书馆**

**2006年12月20日**