●王军 杨冬青 唐世渭

数字图书馆的研究内容和方向

摘要 在我国,数字图书馆的研究和建设,必须鼓励跨学科的合作和国际合作;国家的资助要保持数字图书馆项目(研究类、应用类和内容建设类)间的平衡;要从具体的应用环境出发,建设有特色的数字图书馆,不能盲目追求大、全、新;应重视元数据的研究。参考文献 1。表 4。

关键词 数字图书馆 美国数字图书馆启动计划 研究内容 研究方向
分类号 G250.76

ABSTRACT In research and development of digital library in China, we should encourage interdisciplinary and international cooperation, rationally allocate government supports among research, applications and content development, put emphasis on the development of characteristic digital libraries and the research of metadata. 1 ref. 4 tabs.

KEY WORDS Digital library. Digital Libraries Initiative (DLI). Research contents. Research directions.
CLASS NUMBER G250.76

1 引言

数字图书馆(digital Library, DL)的研究和建设已在全球范围内引起了广泛的关注。在我国,虽然 DL 的研究和建设呈蓬勃发展之势,但对 DL 仍然缺乏一个共同的理解。DL 是否就是图书馆的数字化,DL 的研究包括哪些内容,如何规划 DL 的研究项目,DL 的建设是否应当由国家统一资助,如何确定 DL 合理的建设目标,采取怎样的管理模式,这些问题的解决都需要厘清对 DL 概念和研究内容的理解。

DL 是一个多学科的研究领域,覆盖的范围非常广泛。涉足 DL 的研究人员来自各个学科:图书馆、计算机、网络、人工智能、经济、通信等。这导致了对 DL 众说纷纭的理解。汪冰在其《电子图书馆的理论与实践》中列举了几十种 DL 的定义。本文以美国先后两期的 DLI 计划为研究对象,从两个方面来研究 DLI 的项目:一是在 DLI 的发展过程中 DL 概念的演化,二是分析 DLI 计划的研究内容、研究范围和所涉及的学科。然后,在综合有关资料的基础上探讨数字图书馆的五大类研究内容和方向。最后针对国内 DL 建设和研究的现状,提出若干建议。

2 DLI 计划的发展和 DL 定义的演化

数字图书馆引起整个研究界的关注应归功于美国的 NSF、DARPA 和 NASA 和 1993 年联合资助的数字图书馆启动(Digital Libraries Initiative, DLI)计划。这是一组基础研究项目,发起的意图是促进网络环境下收集、存储、组织和获取数字化信息的手段和技术的进步。很快,DLI 已不能够涵盖 DL 的迅速发展,需要更宽广、更深入的指导方向。于是,NSF 和 DARPA 另外资助了 DL 的工作组,专门讨论 DL 所涉及的研究领域和范围。美国政府的信息基础设施技术与应用(IITA)工作组也成立了相应的研究组讨论 DL 的发展方向和研究范围。1995 年 IITA 在题为"互操作法、可扩展性和数字图书馆研究议程"的报告中把 DL 定义为:"一个有组织的多媒体数据的集合,加上将数据表示为信息和知识的信息管理工具"。

综合 DLI 和 IITA 对 DL 的理解,DL 等价于"网络环境 + 信息集合 + 信息管理工具"。这种理解,把 DL 看成是一个多媒体数据的存储和检索系统在网络环境下的扩展和增强,能为用户提供信息和知识的服务。其本质仍然是一个信息系统。这不足以概括 DL 全部的研究内容。

1997 年召开了 NSF 资助的 Santa Fe 分布式知识工作环境会议。它扩展了 DL 的定义:"数字图书馆不

* 本文的研究得到国家重点基础研究发展规划(973)资助(项目编号 G1999032705)。

仅仅等价于数字化的集合加上信息管理工具,数字图书馆是将信息集合、服务和人连接在一起的一个环境,以支持数据、信息和知识整个生命周期的活动,包括生成、发布、使用、保存和存档。”

基于 DLI 所取得的成就,DLI2 期计划于 1998 年春季启动。2 期计划中,美国国家医学图书馆、国会图书馆、国家人文科学基金会等机构也加入到资助者的行列。Santa Fe 对 DL 的定义确定了 DL1 - 2 的研究方向和内容。DL1 - 2 强调互操作性和集成、内容和收藏的开发管理、应用和可运行结构,强调在领域、经济、社会和国际化的环境中理解 DL。一句话,数字图书馆被看成是一个以人为中心的系统。

将 DL 理解成一种新的信息环境,是一个较为宽泛的定义。这为 DL 的发展留出了相当大的空间。但是,这样的定义也是模糊的。DL 有哪些研究问题,应发展哪些关键技术,如何规划和资助 DL 的研究项目,诸如此类的问题很难从这样的定义中找出明确的答案。要回答这样的问题,需要考察实际的 DL 项目。

3　对 DLI - 1 及 DLI - 2 的项目的分析

3.1　资助情况

DLI - 1 于 1994 年,即项目启动的第 1 年,获得 240 万美元;1994～1999 年间共获得 680 万美元。DLI - 2 在第 1 年获得 440 万美元,几乎是 DLI - 1 第 1 年的两倍,后续资金随着项目的进展继续追加。DLI - 2 资助力度和范围较 DLI - 1 大了许多。DLI - 1 资助了 6 个项目,分别由 6 所著名大学承担,资助额基本上平均分配在 6 个项目中(见表 1)。DLI - 2 到 1999 年底共资助了 24 个项目。10 个大型项目,共计 350 万美元;6 个国际合作项目,共计 23 万美元;其余的共计 65 万美元。DLI - 1 的大多数项目在 DLI - 2 中都得以继续。

3.2　研究的覆盖范围和参与者所在的学科

DLI - 1 侧重于 DL 的研究方面,主要是计算机和信息科学,6 个项目都是由具有非常强的技术背景的专家主持。DLI - 2 极大地扩展了 DL 的研究范围,增加了对不同学科领域支持,包括艺术和人文。表 1 列出了 DLI - 2 项目的研究者所在的学科和系。虽然从总体上看 DLI - 2 的项目涉及众多的科系,但大多数项目仍然是由计算机和信息科学方面的专家主持的。

表 1　DLI - 2 所覆盖的学科

计算机科学	图书馆和信息科学	信息管理
管理信息系统	机器人技术	语言技术
电子工程	地质学	环境科学
地理学	生物医学信息	信息研究
医学信息学	经济学	政府管理
人类学	宗教学	政治学
心理学	社会学	语言学
英语	西班牙语	古典文学
历史	艺术	师范教育

3.3　研究的信息类型和所涉及的技术领域

表 2 列出了 DLI - 2 的项目所研究的信息的类型、媒体和格式,表 5 列出了 DLI - 2 的项目所涉及的技术领域。与 DLI - 1 的项目比较,DLI - 2 的项目在研究的范围、广度上都大大地增加。与 DLI - 1 比,DLI - 2 的项目更加精细和实用,大多数的研究都有较高的实用价值。

表 2　DLI - 2 所研究的信息类型

书目记录	工程教育
Eprints	民间文学
地理参考信息	健康医疗
人文	图书馆参考咨询
医学图像	混合媒体
病例记录	轻音乐
骨骼	模拟程序
社会科学数据	语音
影像	Web
X 光和 CT 照片	

3.4　项目的内容

DLI - 1 的重点在数字图书馆的基础研究方面。把数字图书馆看做是网络信息系统,研究如何从大量的信息源中查找到并显示所需的信息,设计能够有效地操作网络上海量信息的体系结构。

总的来说,DLI - 1 项目有两个重点:DL 的体系结构;海量的多媒体信息的集成和检索技术。重点研究 DL 体系结构的有 UIUC、Stanford 和 Michigan。UIUC 提出了一个基于概念抽取和概念空间构建的 DL 框架;Stanford 提出了一个 DL 服务的互操作协议 InfoBus,分散的 DL 服务和异构的信息仓储只要遵循该协议就可以互相交互;Michigan 研究基于分布式代理的 DL 体系结构。

CMU、UCB 和 UCSB 重点研究新的媒体的集成

和检索技术。CMU 致力于建设一个大型的联机数字影像图书馆,研究视频资料的基于内容的处理和检索技术;UCSB 重点研究海量的空间信息和地图的存储及检索技术;UCB 侧重于对多种类型信息及其服务的集成,以提供一个以工作为中心的数字化信息服务环境(见表 3)。

表 3　DLI-1 项目概要

单位	项目名称	信息类型	主要内容	关键技术
UCB	UCDL	环境数据	环境规划和地理信息系统	文档图像分析、多层文档模型
UCSB	Alexandria	空间信息、地图等	海量的空间和地理信息的集成和检索	图像的并行处理、用户接口、元数据模型
CMU	Informedia	影像	数字影像图书馆	视频信息的基于知识的全内容的处理
UIUC	InterSpace	SGML 科技文献	科技文献的联邦化仓储	概念抽取、自动索引、概念空间生成
Stanford	InfoBus	计算机技术文献	各类信息服务的互操作机制	分布式对象技术、服务交互模型
Michigan	UMDL	地球和空间科学资源	智能代理框架	分布式智能代理技术

表 4　DLI-2 项目所涉及的的技术领域

3-D 模型	访问控制
代理技术	档案和保存
声频信息检索	自动分类和聚类
数据(访问)服务	数字化影像
DL 的经济模型	DL 的协议
DL 的联邦	地理信息系统
图像处理	信息过滤
信息可视化	学习内容
链接技术	文献踪迹分析
移动计算	多媒体整合
自然语言处理	OCR 技术
平行处理	电子化笔记本
个人化	电子文献出处管理
手稿恢复	语音处理
视频信息概要化	文本分析
影像编辑	

表 4 列出 DLI-2 的项目所涉及的技术领域。其中有对新技术的研究项目,如 Cornell 大学和 Stanford 大学的互操作性和安全性的研究,Arizona 大学的自动分类,Indiana 大学的信息过滤,Pennsylvania 大学的数据出处研究;有对新的媒体的研究,如 Michigan 大学的人声录音,Johns Hopkins 的音乐,Harvard 大学的政治和经济数据;有对新的信息内容和新的领域的研究:Texas 大学的人类模型和图像,Kentucky 大学的文学手稿,Columbia 大学的病例,California 大学的民间文学等等。DL 的研究和应用已全面铺开,涉及到众多学科领域和技术。

数字图书馆正在形成一个新的学科社区,其成员来自不同的学科,是学科综合化的典型代表。因此,很难界定 DL 的研究范围。DL 已经超出了计算机和通信的专业领域,扩展到了广泛的人类知识和研究领域。DLI-2 认识到技术的关键进步将来自于非科学领域的审视、方法和应用,而非计算机科学的实验室。因为,它们将是计算机和信息科学新的重要的研究问题的来源和解决的场所。DLI 项目没有去限定 DL 的研究内容和领域,而是吸引了一大批来自各个领域的研究人员,由他们来不断地拓展和深入 DL 的研究领域,这可算是 DLI 的项目的一大成功之处。

4　DL 的研究内容和方向

DL 的研究内容有 5 个方面:知识产权和经济问题,全球分布式 DL 环境下资源的索引和发现,互操作性,元数据,多语种信息检索。这 5 个方面包括了实现全球 DL 的关键研究内容。

4.1　知识产权和经济

知识产权和经济的问题是拥有高质量内容的 DL 大范围应用的最大障碍,合理的经济模型是在分布的、自治的 DL 中获得资源、传播资源和激励高质量资源生产的基础。知识产权和经济的问题是 DL 研究中相当复杂的问题。

4.1.1　法律和社会政策

直接套用为传统的纸质媒体制定的法律体系产生了很多问题,要研究制定适用于数字资源和网络环境特点的知识产权和经济管理法规。在法律法规缺乏或尚不完善的情况下,信息商要制定自己的信息传播政策,实施证书管理等措施以保证自己的权益。

4.1.2　体系结构和机制

知识产权和经济的问题不是 DL 设计中一个孤

立的方面,它要求整个 DL 体系结构从底向上的有效支持。如果把 DL 的信息服务看成是一种商务活动,那么对 DL 体系结构最基本的要求就是支持这种商务活动。这不仅是指收费管理。收费管理只占整个商务活动的最后阶段的一小部分。它还包括:发现感兴趣的内容和服务(包括与之相关联的代理),协商交易条件、实际的交易操作等。相关的机制包括:描述信息商品和服务的通用语言,用以匹配服务的寻求者和提供者的广告和查找工具,促进交换条件讨论的协商工具,个体身份、团体成员的认证机制,加密服务,交换付费等。

4.1.3　内容和服务

体系结构仅是建立 DL 的一个壳子。DL 的核心是内容和服务。DL 服务的提供者面临着诸如此类的难题:开发什么样的服务,怎样传递这些服务,以什么样的条件提供这些服务。用户同样需要做出选择,使用什么样的服务,遵循什么样的条件。其他的问题还有用户行为研究,适用于信息消费者和提供者的经济模型等。

4.2　全球资源发现

全球资源环境有三个发展趋势:网络的连通性和计算能力将显著增加;越来越多的文献将带有安全性和价格的控制;限制信息访问的法律和法规将增加。在这样的背景下,全球资源发现所研究问题可概括为 3 个方面。

4.2.1　系统

开发一个基础的系统体系结构。该体系结构能够方便导航和检索,为网上大量、多样的信息提供中间层的支持;能够识别、访问和检索数字资源;能为尽可能多的信息仓储提供一个统一的视图。主要的研究方向有:查询路由(routing)、数据库交互和一致性管理。

4.2.2　内容

(1)数据库选择:选择合适的数据库完成用户的查询。影响数据库选择的主要因素有:描述数据库的元数据,性能和价格等。

(2)表示和查询语言:DL 中包含各种类型的多媒体和超媒体文档,如何描述和表示这些复杂的信息是一个需要研究的问题。描述文档的视图有 3 类:逻辑视图,表示(layout)视图和内容视图(如 XML)。未来的查询语言应当包含能够指定查询结果的结构、表示和内容的操作符。需要制定相应的标准以支持互操作性。

(3)语义的异构性:数据库的模式、文档类型、查询语言各不相同。系统要提供处理这类异构性的机制以方便用户。

(4)分级:为了把用户引导到合适的资源上选择最合适的答案,需要对数据库和文档按照内容和质量分级,这对未来的 DL 至关重要。

4.2.3　人机界面

(1)查询构造和指导:帮助和指导用户构造更优的查询。可能的技术包括 QBE、自动查询扩展和参考。

(2)查询结果的表示和可视化:在理解文档内容的基础上用更成熟的方式来表示检索结果。

(3)用户任务的理解:理解用户的任务能够大大地改善查询的执行质量。

(4)用户的自动培训和执行过程显示:引导用户高效地使用系统以发挥系统的潜能,显示查询执行的过程将增加界面的友好性。

4.3　互操作性

互操作研究的主要问题有:

(1)信息模型。主要的研究问题有两个:中间层元数据和文档的数据库视图。中间层收集下层资源的结构、语义、服务等方面的信息。文档视图对 DL 中的文档创建数据库视图以便利查询。

(2)协调和控制。DL 的协调和控制机制主要有两个方面:精化资源描述语言,应付动态的变化;一致性问题,即如何为不同的自治的服务提供事务执行的保证。

(3)查询处理。用于异构的信息资源的查询的规范化;查询估价和查询路由;相关性反馈。

(4)实现机制:互操作的实现实际上就是分布式计算要研究的问题。分布式计算的标准是一个快速发展的领域,如 CORBA,DCOM 等。需要密切跟踪分布式计算领域的发展,为互操作选择实现机制。

4.4　元数据

元数据的主要研究问题有:

(1)元数据和资源联系的模型:元数据可以嵌入到资源中,通过协议与资源相关联或存储到分离的数据库中。

(2)关于服务中介的元数据:越来越多的网上信息是通过专门的服务接口提供的,不能公开访问。需要描述这类信息资源的服务、政策、交易手段的元数据。

(3)元数据的生成和管理:供非专业人员使用的

元数据的生成和管理工具。这些工具常集成到用户的系统和环境中。

(4)与信息体系结构标准的集成:W3C 已经提出了一个表示元数据的建议标准——RDF。基于 RDF 的通用元数据体系结构以及其他标准,如 ISO11179,都需要进一步的细化和应用评价。

(5)建构注册系统:需要一个注册系统的体系结构,能够用人和机器都可以理解的方式来表达模式,为元数据的应用、局部扩展、与其他模式的映射等提供权威的指导。

(6)核心元数据集:已有的核心元数据集有 Warwick 框架、Dublin Core 等。还需要其他方面的核心元数据集,如结构和导航,数字对象的管理,认证、证书和出处,许可条件等。

(7)互操作性和复杂性:类似于 Dublin Core 的核心元数据集是可以通过增加元素或局部细化来扩展的。但是复杂性的增加必然会损害互操作性。需要这类问题的系统化解决方案。

(8)复杂的数字资源的元数据:现有的大多数元素数据都是关于文字信息的,需要开发关于复杂的数据对象的元数据,如声频、视频资源,动态变化的对象等。

(9)评测和衡量:关于各种使用元数据标准和提案的可应用性、成本、效益等的评测。

(10)政策问题:元数据的应用势必跨越国家和文化的边界,这将造成对元数据不同的理解。需要制定有关的政策进行规范。

4.5　多语种信息访问

这方面的研究大致可以分为 3 类:

(1)用户需求。建构必需的体系结构,研究用户与多语种信息交互的方式,研究如何帮助用户克服语言障碍。特别是在多语种环境中如何容纳用户在语言理解和运用上的差异,如何获得用户的反馈以提高查询的质量、改善查询结果的表示、提高翻译的质量等。

(2)技术。多语种信息访问不是一个简单的"信息检索 + 机器翻译 + 资源"构成的通道。它的主要研究领域包括:多语种索引工具,用户查询处理,文档聚类,检索结果的自动总结,可视化工具,多语种的多媒体访问等。

(3)资源。开发系列的高质量的多语种资源对于实现实用的系统至关重要。应重点研究的领域包括:建立新的标准的多语言信息仓储;对已有资源扩展以包含新的功能和语言;独立于语言的资源的识别和开发,如本体论(ontologies)和主题词表;真正的多语种而非双语的资源的建议;开发评价新资源的标准和流程。

5　结语

通过对 DLI 项目的考察,结合对 DL 研究内容发展方向的综述,有如下几点值得强调。

5.1　鼓励跨学科的合作和国际合作

对于我国这样的发展中国家,合作与共享具有特别意义。一方面,国内的 DL 研究和建设力量要联合,取长补短,避免重复;另一方面通过国际化的合作,及时了解国际的发展动态。以资源共享为条件来换取技术,降低研发成本和建设周期。组建有关的数字图书馆建设协调委员会,与有关的国际组织和机构联系协调,促进合作,强调共享,推进长期项目。

5.2　支持 DL 的研究项目

DL 属于国家信息基础设施。国际上重要的 DL 项目都是由政府资助的。要发展我国的 DL 事业,离不开国家的支持。对于我国这样的发展中国家,政府支持更应得到重视,要研究资金的投入方式和管理制度,充分地利用每一分钱,真正地推动 DL 事业的发展。

5.3　以内容和服务为中心来组织 DL 的建设

目前成功的 DL 项目都是基于可操作和可应用的系统,能够提供真实的、对用户有意义的内容的访问。DL 的真正贡献仍然是信息的内容和服务。图书馆必须高举服务的旗帜,以用户需求为中心来组织资源和改进工作模式。计算机的研究人员要了解用户的真实需求,以此作为技术的出发点。必须意识到最终能使数字图书馆得到资助与发展的是高质量的信息内容和便利的访问与使用,真正吸引和鼓励大多数人使用 Internet 和 DL 的是其内容,技术是为内容服务的。

5.4　从具体的应用环境出发建设各具特色的 DL

我国各单位、各地区的发展很不平衡,资金、技术、人员素质、信息储量等条件差别很大。应该根据具体的应用环境建设各具特色的 DL,不应强求一致和集中管理。在承认多样性和分布性的基础上,推荐一些共同遵循的标准,如元数据、体系结构框架、接口等,使得各个 DL 的结构是开放的和灵活的,能够进一步联合和集成,组成更大规模的 DL。总之,

只有在尊重应用环境的前提下开发的 DL 才是具有生命力的。

5.5　脚踏实地地发展 DL 事业,不能盲目追求大、全、新

不同的应用环境决定了我国 DL 的建设不能完全照搬国外的发展模式。我国在自动化程度、数字资源丰富度,以及 DL 起步阶段都不同于西方国家,因此,应在充分考虑国内实际情况的基础上实事求是地发展我们的 DL 事业,不能盲目地追求大、全、新。

5.6　重视元数据的研究

元数据的重要性再强调也不过分。Web 的发展初期极大地受益于几乎任何人都可以灵活地生成和发布信息。但是,也是这种灵活性使得今天在 Web 上查询和利用信息如此困难。对元数据的重视实际上反映了 Internet 社区力图通过增加标准化、提高一致性来提高网络信息的可访问性,同时不以牺牲 Web 的灵活性为代价。国内的 DL 研究要重视中文元数据的开发,注意与有关国际规范的兼容和互操作。

5.7　大力发展应用

要充分挖掘 DL 的利用价值,显示 DL 研究和建设的效益,为 DL 的深入发展提供素材。特别要重视远程教学的应用,特对是大学教育。DL 本身就是一个良好的网络教学和虚拟教学的平台。

参考文献

1　Progress Toward Digital Libraries, eds. Gary Marchionini and Edward A. Fox, Special Issue, Information Processing & Management, 35(3), May 1999

2　The Digital Libraries Initiative: Update and Discussion, Edward A. Fox, Bulletin of the American Society for Information Science, Vol. 26, No. 1, 1999

3　Social Aspects Of Digital Libraries —Final Report to the national Science Foundation, UCLA - NSF Social Aspects of Digital Libraries Workshop, Invitational workshop held at UCLA, February 15 - 17, 1996. http://www - lis. gseis. ucla. edu/DL/UCLA DL Report. html

4　Perspectives on DLI - 2 - Growing the Field, by Michael Lesk, Bulletin of the American Society for Information Science, Volume 26, No. 1, 1999

5　NSF/DARPA/NASA Digital Libraries Initiative - A Program Manager's Perspective, Stephen M. Griffin, D - Lib Magazine, July/August 1998

6　Digital Libraries Initiative - Phase 2 Fiscal Year 1999 Awards, Stephen M. Griffin, D - Lib Magazine, July/August 1998

7　Interoperability, Scaling, and the Digital - Libraries Research Agenda, IITA Digital Libraries Workshop, Clifford Lynch, Hector Garcia - Molina, August 22, 1995, http://www - diglib. stanford. edu/diglib/pub/reports/iita - dlw/main. html

8　An International Research Agenda for Digital Libraries - Summary Report of the Series of Joint NSF - EU Working Groups on Future Directions for Digital Libraries Research, October 12, 1998, http://www. ercim. org/

9　"Digital Libraries: Issues and Architecures", Peter J. Nurnberg. etc. http://csdl. tamu. edu/dl95/papers/nuernberg/nuernberg. html

王　军　北京大学信息管理系讲师,北京大学计算机系博士生。通讯地址:北京市海淀区。邮编 100871。

杨冬青　北京大学计算机系教授、博士生导师。通讯地址同上。

唐世渭　北京大学计算机系、信息中心教授,博士生导师。通讯地址同上。

(来稿时间:2001-05-18)

《企业战略信息管理》出版

霍国庆博士编著的《企业战略信息管理》一书,已由科学出版社出版发行。该书从信息主管(CIO)的角度去审视一个组织的信息战略管理问题,对战略信息管理和 CIO 的关系,以及其三个组成部分(信息技术、信息资源、电子商务管理)作了深入论述。该书是各类社会组织,尤其是企业 CIO 的必读专业书,可作为 MBA 教学及高校信息管理、图书情报、企业管理专业师生的教材或教学参考书。(里边)