中国科学院文献情报系统数字图书馆资源组织体系研究

张智雄 宋 文 孙 坦 李广建 黄永文

(中国科学院文献情报中心 北京 100080)

[摘要] 在分析中国科学院知识创新工程对文献情报的需求和中国科学院文献情报系统已有资源的基础上,提出中国科学院文献情报系统数字图书馆建设中数字资源的建设内容、资源组织模式和资源描述方法。

[关键词] 数字图书馆 数字资源 元数据

[分类号] G254

A Study on the Resources Organization System of the Digital Library of the Documentation and Information System of Chinese Academy of Sciences

Zhang Zhixiong Song Wen Sun Tan Li Guangjian Huang Yongwen

(The Documentation and Information Center of Chinese Academy of Sciences, Beijing)

[**Abstract**] Based on an analysis of the documentation and information needs of the knowledge innovation and existing digital resources , this paper puts forward the content, organizing mode and description method of the digital resources in the construction of the digital library of the Documentation and Information System of Chinese Academy of Sciences.

[**Keywords**] digital library digital resources metadata

1 引 言

根据国家科技创新体系的总体规划,中国科学院正在全面推进知识创新试点工程。为了保障知识创新工程对科技文献信息的需求,为知识创新工程提供强有力的支撑,中国科学院于2000年9月召开了全院第五次文献情报系统工作会议,提出了建设中国科学院数字图书馆的目标。

中国科学院文献情报系统(DISCAS)数字图书馆建设并不是简单地数字化馆藏,也不是建立起网上资源的简单目录,亦非以展示某一先进技术为目的,更不是将图书馆工作流程自动化或将部分或全部藏书数字化,而且是广泛借鉴国内外数字图书馆发展的经验与教训,充分利用中国科学院文献情报系统三级文献保障体系、丰富的文献资源及成熟的技术等方面的优势,构建起支持普遍存取、分布式管理和集成化服务的信息环境。

收稿日期:2001—05—22

DISCAS数字图书馆建设的核心是以统一标准和规范为基础,以数字化的各种信息资源为底层,以分布式海量资源库群为支撑,以智能检索技术为手段,以电子商务为管理方式,以宽带高速网络为传输通道,将信息准确、快速地传递给用户,促进网络环境下的信息传递和知识流动,加速信息的知识增值和社会的知识更新。

2 中国科学院系统数字资源组织体系

数字图书馆的基础是海量的信息资源,但海量信息资源并不等同于馆藏资源的数字化,也不等同于单纯的网络资源目录,或各种数字资源的简单拼凑。中国科学院数字图书馆的资源建设将吸取国内外的经验,建设一个包含不同层次、不同类型,相互联系、密切配合的资源库群,通过先进的技术手段,随时、快速地向读者提供信息服务。

2.1 中国科学院系统数字化资源组织的内容

配合中国科学院知识创新工程,保障知识创新工程对文献信息的需求,建立从书目到目次文摘,最终到全文的全方位文献查询和提供服务。

- 在文献内容上:作为国家科学图书馆,重点搜集整理基础学科,如数理化天地生及高新技术学科的文献;
- 在文献类型上:重点是期刊、会议录、科技报告等;
- 在文献载体上:包括印刷型、光盘型、网络型等各种载体类型的文献;
- 在文献来源上:包括中国科学院文献情报系统各单位收藏的传统文献、国内兄弟图书馆收藏的文献、购买的光盘数据库、网络电子信息资源、通过专用搜索引擎搜集的免费网络信息资源等。

2.2 数字信息资源的组织模式

中国科学院数字图书馆将建设3个层次的数字资源群:书目库群、目次文摘库群和全文库群。目次文摘库群可保障用户更深层次的查询需求,书目库群和全文库则是为用户提供全文服务的基础。

2.2.1 书目数据库群 中国科学院文献情报系统是由中国科学院文献情报中心、地区文献情报中心和所级文献情报机构共100余个单位组成的,将这100多个单位收藏的文献通过全院联合目录数据库集中统一揭示,采用高效的联机采编系统实现全院范围内的联合采购和联机联合编目,这是数字图书馆建设的最基础的工作。

中国科学院文献情报中心从上世纪80年代初就开始建设全院期刊联合目录数据库,现在该数据库已发展成为全国期刊联合目录数据库。数据库收录国内300余家主要图书馆和情报机构收藏的中西日俄文期刊10万余种,报导馆藏35万条,是目前国内也是亚洲规模最大的期刊联合目录数据库。数据库同时提供光盘和网络版(http://159.226.100.6/catalog/)两种服务方式。网络服务系统提供了世界3 000余种网上期刊的目次文摘和部分全文的文献链接,已成为网络期刊的导引。数据库成员馆包括大型公共图书馆、中国科学院系统的图书馆、中国社会科学院系统的图书馆、各大部委的情报所、科研系统的图书馆、重点高校的图书馆和全军卫生系统的图书馆等各类型的图书馆。数据库在推进全国文献信息资源的共建共享、协调外文期刊的订购、共享标准编目数据、促进我国数字化图书馆的建设等方面发挥了不可替代的作用。

在期刊联合目录数据库的基础上,建设全院各文种图书、会议录、学位论文、电子资源、声像资料等的联合目录数据库,达到全院各单位收藏的所有文种、所有类型文献的集中统一揭示,配合二次文献数据库,为全院范围的原文传递做好准备。

2.2.2 目次文摘数据库群 在联合目录数据库的基础上,根据中国科学院数字图书馆的定位和服务对象,对重点学科的重点文献进行目次文摘级的数字化加工,满足读者深层次的文献查询需求。目次文摘库群将包括中国科学院系统多年来合作建设的13个中文文献数据库,含中国数学文献数据库、中国物理文献数据库、中国化学文献数据库、科技成果数据库等;承担国家科技文献信息系统中自然科学和高技术馆藏期刊的文摘数据加工工作,形成西文期刊会议录二次文献数据库;购买EBSCO、PQDDB,CSA等期刊目次文摘库以及通过专题搜索引擎或人工收集的网络文献库等等。根据中国科学院数字图书馆的规划,将组织全院以分工合作的方式建设一批二次文献数据库,扩大二次文献数据库的规模和覆盖范围。

2.2.3 全文数据库 中国科学院数字图书馆全文数据库主要包括两部分内容:①中国科学院文献情报中心收藏有大量珍贵的古籍文献,其中有一部分是孤本,为保存珍贵的历史文化遗产,在数字图书馆建设中,要对这些特色馆藏进行全文加工。②互连网是一个无限庞大的信息资源库,这个信息资源库中存在着大量有价值的科学文献,也有大量的所谓"垃圾"信息,中国科学院数字图书馆全文数据库建设的一个重要方面是对网上有价值的科学文献进行收集、整理,为用户提供更好的服务。

在上述3种类型数据库群的基础上建设统一的集成用户信息服务系统,通过这些数据库群的协同服务,向读者提供全方位的文献查询和原文获取服务。

读者在查找期刊、会议录或图书等指定的文献时,系统可自动进入相应的书目库进行查询,并显示读者所需文献的收藏单位和收藏地,通过自动原文请求服务系统和全院文献保障系统,文献可在最短的时间内送达读者手中。

读者也可以随时浏览指定的期刊、会议录等文献的目次文摘,看到需要的文章可随时发出原文请求。

读者还可根据研究方向和内容,进行专题文献检索。系统会自动对目次文摘库和全文库进行查询操作,如数据库中有全文,则显示全文供读者阅读和下载,若库中无全文,系统会显示读者所需的文章在哪份文献上,此文献被哪家单位收藏等情况,通过自动原文请求服务系统和全院文献保障系统,读者同样能得到全文。

3 资源描述体系

为了描述书目、馆藏、目次、文摘、全文、古籍、多媒体等

多种数据类型，确保数据的一致性、完整性和有效性，中国科学院提出了建设数字图书馆资源描述体系的概念，希望通过"资源描述体系"的建设，推动数据标准化，提高数据质量，同时通过引进开放、简便的新的数据描述方法，提高信息资源的加工效率，实现系统间的相互操作。资源描述体系如下：

- 所有数据格式遵循 ISO2709 和 XML 的标准；
- 书目数据的描述依据 USMARC、CNMARC 和 UNIMARC；
- 网络资源、目次文摘依据 DC CORE 标准；
- 档案，手稿依据 EAD 标准；
- 推送频道依据 CDF 标准；
- 多媒体依据 SMIL 标准；
- 政府出版物依据 GILS 标准；
- 其他类型文献可自定义元数据集及 DTD 进行描述。

3.1 以 XML 和 MARC 为基础的元数据应用体系

中国科学院文献情报系统数字图书馆的建设，并不是利用现有的几种单纯的元数据标准能够支撑的。制定有限的元数据集，不能满足一个系统建设数字图书馆的需要。为实现系统互操作提供支持，应优先确定元数据描述格式。

上世纪 90 年代以来，出现了众多种的 matadata 规范，如 Dublin Core，ROADS Template，PICS(plateform for internet content selection)，Web collections，CDF(channel definition format)，MCF(meta content framework)，EAD(encoding archival description)等等，这些元数据规范各自定义了自己的元数据集和元数据描述格式，造成的结果自然是各种元数据的不兼容。而只有在 XML 描述语言出现之后，通过 RDF 的应用，才将众多的元数据集纳入到一个开放式的框架之中。这说明，构建可互操作的元数据，仅仅定义元数据集是不够的，还应充分重视元数据的描述格式。

同样是 Dublin Core 元素集，在一般格式(element ="value")、HTML 格式和 XML 格式中会需要不同的应用系统，以达到完全不同的效果。

一般格式的示例如下：

DC.Title ="Song of the Open Road"

这种格式是非常简单的，但是它不能表示层次、嵌套等复杂结构。

Dublin Core 在 HTML 格式中采用"name ="、"content ="组对方式，如下所示：

< META name ="description" content = "Everything you wanted to know about stamps, from prices to history." >

< META name = "keywords" content = "stamps, stamp collecting, stamp history, prices, stamps for sale" >

它同样不能表示复杂的结构。而且至少到目前为止，这种格式还没有十分通用的解析器，在设计层次、嵌套的元素集等方面同样存在着缺陷。只有在 XML 之中(即纳入 RDF 框架之中)，Dublin Core 才能实现真正意义上的互操作。XML 采用结构化的描述方式，因此，在元数据集中，可以定义层次、嵌套等结构复杂的元素；由于 XML 允许定义自己的文件类型(DTD)，因此可在通用的(如 DC、TEI)元数据集之外定义自己的元数据元素；由于 XML 得到 IT 产业界的巨大支持，包括 Microsoft、Oracle、SUN、Sybase 在内的产业巨人都在开发它的解析器、检索引擎、编辑、显示软件，所以今后人们能够实现真正的大规模操作。

笔者认为 MARC(更确切些是 ISO 2709)是与 XML 相对应的描述格式，而只有 USMARC(或 CNMARC)这样一个个的具体应用，才能与 Dublin Core 等具体的元数据集相对应。

尽管 MARC 有着种种缺陷，如它是磁带时代的一种流式文件体系，是为顺序读取而不是为随机读取设计的；只能做数据交换格式，在数据检索方面没有太多的意义；作为标记语言，在反映结构化数据方面缺少可扩充性；只能描述文本而且长度有所限制(全记录长不超过 99999，字段长不超过 9999，不能描述大字段内容如全文)；另外它过于繁琐，自我封闭，得不到 IT 产业界的广泛支持。但 MARC 毕竟已有一套成熟的体系，中国科学院联机联合编目数据和各研究所图书馆馆藏数据的建设，还是需要利用 MARC 描述方式的。

3.2 根据特定的应用，参考国际通用规范，制定中国科学院文献情报系统相应的元素集。

事实证明，像 CNMARC 一样，目前再定义一个完全统一的元数据标准是不现实，也是不可能的。处理文本和处理图像、音频、视频、多媒体需要完全不同的方式方法，多种元数据集并存既是历史的产物，同时又是现实的必然。

Dublin Core、EAD、GILS、PICS、SMIL、CDF 既然已广泛应用，同时又逐步采用了以 XML 为编码的描述方式，因此，字面"汉化"或自己定义一个迥然不同的元素集就没有必要。中国科学院文献资源的建设，将在研究现有元数据标准的基础上，选取国际上通用的规范加以利用。

对于中国科学院特殊需要的元数据集，将在参考这些规范的前提下，通过定义标准的 DTD 来自行定义。每一个元素集都将按相应的文件类型定义，从而保证在 XML 环境下，系统仍然有着较高的兼容性和互操作性。

3.3 研究开发相关软件，促进新型元数据的应用

与 MARC(ISO 2709)格式相比，XML 在任何方面都不逊色。XML 能够像 MARC 一样处理变长字段，能够非常方便地实现字段重复，具有链接功能，能够更好地实现连接字段

……,这些都是严格的关系(对象)数据库所不具备的特性。

自从XML问世后,包括Microsoft、Oralce、SUN、Sybase等在内的IT界巨头都纷纷推出了有关XML的解析器、可视化编辑器、XML检索引擎、XML浏览器等新产品,目前Oracle8i(自8.1.6之后)、Sybase 12、MS SQL Server2000等数据库都提供了对XML的支持。在Microsoft的.Net项目,Sun的Java2、J2EE、JSP规范中,XML都占了较大的份额。还有许多IT厂商正在从各个方面促进XML的应用,致力于创造基于XML的网络服务。而许多出版商如OVID、ISI、EBSCO等实际上都早已利用XML(SGML)来存储目次、文摘、全文等文献数据,提供基于XML的文献的检索,并按XML格式向客户输送数据。

中国科学院数字图书馆的建设,不走将元数据简单等同于Dublin Core的道路,更不会走图像扫描的道路。

我们将在基于MARC和XML的环境下,充分利用IT界的技术成果,研究和开发元数据的解析和校验系统、映射及转换系统、编辑浏览系统、检索发布系统,形成一个开放的、符合国际数字图书馆主流的资源组织体系。

参考文献:

1 张福学.网上虚拟资源建设论纲.四川图书馆学报,2001(2):33~38
2 史卫国.数字图书馆的信息组织.四川图书馆学报,2001(2):43~46
3 苗　凌,李人厚.数字图书馆中特色数据库的研究与实践.大学图书馆学报,2001(2):13~15
4 郑巧英,杨宗英.图书馆数字化建设与研究新进展.上海高校图书情报学刊,2001(1):15~19
5 王　纯.中国数字图书馆建设、数字信息资源开发及网络建设的现状.中国图书馆学报,2000(4):79~80
6 林海青.数字图书馆的元数据体系.中国图书馆学报,2000(4):59~64
7 孙晓菲.XML与数字图书馆.现代图书情报技术,2000(4):14~15
8 肖　珑.元数据格式在数字图书馆中的应用.大学图书馆学报,1999(4):18~24
9 索传军.论网络化图书馆的信息资源建设.图书馆,1999(1):22~25
10 赵　伟.数字图书馆研究的历史和现状.情报科学,1999(2):193~195,199
11 http://www.ifla.org/II/diglib.htm
12 http://lcweb.loc.gov/loc/ndlf/digital.html
13 http://www.libnet.sh.cn/istis/dlib/
14 http://www.dlib.org/

[作者简介] 张智雄,男,1971年生,博士,发表论文多篇;宋　文,女,1961年生,副研究馆员,发表论文数篇;孙　坦,男,1971年生,博士,发表论文多篇;李广建,男,1963年生,教授,发表论文多篇;黄永文,女,1975年生,硕士,发表论文数篇。