

●谈春梅 叶继元 汪令全

图书馆电子信息资源数据库系统的开发研究^{*}

摘要 电子信息资源数据库系统是指把有价值的电子出版物信息资源和非电子出版物的电子信息资源,进行有效搜集,采用数据处理技术和数据库接口技术,以完全基于 Internet 应用的高质量数据库,供用户检索和利用。数据库建设过程中,必须把握好资源选择和数据格式标准化两个重要环节。参考文献 5。

关键词 电子信息资源数据库 信息组织 网络应用

分类号 G250.74

ABSTRACT Database system of electronic information resources is a kind of high-quality Internet-based database that effectively collects valuable information resources of electronic publications and electronic information resources of non-electronic publications with the application of data processing and database interface technologies. The authors think that resource selection and format standardization are two important factors in database development. 5 refs.

KEY WORDS Electronic information resource (EIR). Information organization. Network applications.

CLASS NUMBER G250.74

电子信息资源数据库系统是指紧紧围绕用户的需求,把有价值的电子出版物信息资源和非电子出版物的电子信息资源(主要是指 Internet 上的电子期刊、电子论坛、电子报告等)进行有效搜集,采用数据处理技术和数据库接口技术,以完全基于 Internet 应用的新的质量数据库,供用户检索和利用。电子信息资源数据库是电子信息资源组织的一种有效方式,已成为图书馆界开发研究的重要课题。

1 加强图书馆电子信息资源数据库系统的开发研究势在必行

1.1 满足用户需求的需要

网络环境下,用户对图书馆电子信息资源服务的需求发生了很大变化,形成了综合化、社会化、网络化、集成化、高效化的特点。这就对图书馆提出了如下要求:(1)能够提供经过图书馆鉴别、选择、加工、分类和组织的有价值的电子信息资源;(2)能够为用户提供高质量的电子信息资源数据库检索系统,以提高检索效率,节约检索时间,降低检索成本。

1.2 提高图书馆信息技术开发研究能力的需要

面对图书馆软硬件的更新换代,网络系统的应用和升级以及数据库的开发设计等繁重而艰巨的任务,国内图书馆的信息技术开发研究能力很不适应。瞄准国际图书馆界先进技术,在科研实践中大力提高图书馆的信息技术开发研究能力

已成当务之急。只要每个图书馆都能根据用户的需求及本馆的技术力量和物质条件,有组织有计划有步骤地围绕电子信息资源数据库系统的相关技术,分课题、分层次、分阶段地开展研究,不断探索和实践,不仅能取得丰硕的科研成果,而且能使图书馆专业人员的计算机操作技能、网络技术应用及软件设计能力和标引水平等也得到相应提高。

1.3 图书馆自身发展的需要

随着网络技术的不断发展和电子信息资源的不断涌现,图书馆如何进一步提高自身生存能力、竞争能力,有所作为?美国国会图书馆的思路对我们极有参考价值。他们要求图书馆专业人员对数字图书馆技术进行开发研究,通过对一切可获得的电子信息资源的理解和组织,以新的集成的方法实现数据库在网络上的动态查询和利用。这意味着,图书馆紧紧围绕电子信息资源数据库系统相关技术展开研究,符合图书馆的发展方向,使图书馆的服务对象切实感受到,无论过去、现在还是将来,图书馆都能够为他们的科学研究和知识创新提供信息保障。惟有如此,图书馆才会有广阔的发展空间和发展前景,才可能进一步走向繁荣。

2 开发研究电子信息资源数据库系统必须把握两个重要环节

2.1 开发研究对象的选择

在明确主题的基础上,电子信息资源开发研究对象应重

^{*} 本文系南京邮电学院科技计划项目(99院02号)的研究成果之一。

点选择具有权威性和实用性的光盘数据库以及电子期刊、电子论坛、电子报告等。这些电子信息资源不仅用户使用范围广、频率高,而且容易收集、加工和组织。

2.2 数据格式的标准化

标准化是任何数据库系统赖以生存的基础,也是基于 Internet 应用和发展的主线。图书馆对于电子信息资源数据库系统开发研究,必须坚持数据格式标准化原则。数据格式是信息基础结构的描述。数据格式只有符合为大家所公认和遵循的统一标准,才有可能在不同的计算机系统、不同的网络平台间交换数据。图书馆馆藏书目文献数据库的建设,因为采用了由头标区、目次区和数据区三部分组成的 MARC 数据格式,才得以达到书目数据资源共享的目的。电子信息资源数据库的数据格式,应采用 MARC 数据格式。因为只有采用 MARC 数据格式,电子信息资源才能被完整地提示和描述,实现馆际间的通讯交流,达到电子数据资源共享的目的。

3 开发研究电子信息资源数据库系统的 3 个关键技术

3.1 建立电子信息资源简要式 MARC 数据结构体系

电子信息资源数据库的数据结构是决定其信息含量、标引深度和检索途径的基础。OCLC“都柏林核心”(Dublin core)数据结构体系,虽然没有突破关系型数据库非常严格的表结构,但它使电子信息资源数据库在建立数据结构体系时有了很好的参照体系。笔者在这个参照体系下,结合研究和实践,提出建立电子信息资源简要式 MARC 数据结构体系,该体系分为 CNMARC 和 USMARC 数据结构体系两类,它们所确定的 MARC 数据字段分别是:

(1) 电子信息资源简要式 CNMARC 数据结构体系的基本字段:

0 号字段,标识块:001,记录标识号;005,资源最后处理时间和日期;010,国际标准书号 ISBN;011,国际标准连续出版物号 ISSN;013,其他(如标识 CD、DVD、MTV 等光盘国际录制代码号 ISRC);091,统一书刊号;092,订购号。

1 号字段,代码信息块:101,资源语种;102,出版或制作国别;135,计算机文档数据类型(如文本、数字、程序等)。

2 号字段,著录信息块:200,题名与责任者说明;205,版本项(如网络版、单用户版、多用户版、学习版、演示版等);207,资源特殊细节项:连续出版物卷期编号;210,出版发行项;215,载体形态项(资源特定类型标识,如计算机光盘、计算机软盘等);225,丛编;230,资源特殊细节项(标识计算机文件的文件特征,如数据文件或程序文件,分别注明它们的文件个数、记录条数或语句行数等)。

3 号字段,附注块:300,一般性附注;326,出版周期附注(标识电子期刊、光盘数据库等当前出版周期与时间及较早的出版周期与时间);327,内容附注;336,计算机文档类型附注。

4 号字段,款目连接块:410,丛编;456,复制为;488,其他相关作品。

6 号字段,主题分析块:606,论题性主题词;610,非控主题词;626,技术细节检索点(标识计算机文档所用的计算机类型、操作系统、程序语言等);690,中图法分类号。

7 号字段,知识责任者块:700,个人名称—主要知识责任者;701,个人名称—等同知识责任者;710,团体名称—主要知识责任者;711,团体名称—等同知识责任者。

8 号字段,国际使用块:拟增设 856 字段,电子信息定位与检索(包括资源的网址、格式,如 ASCII、HTML 等格式、使用方式如文件传输 FTP、远程登录 Telnet 等、主机名称、文件名等)。

(2) 电子信息资源简要式 USMARC 数据结构体系的基本字段:

0 号字段,标识:001,记录控制号;005,资源最后处理时间和日期;020,国际标准书号 ISBN;022,国际标准连续出版物号 ISSN;041,语种代码;044,出版或生产国别;082,杜威十进分类号;092,中图法分类号;093,订购号。

2 号字段,题名与责任者说明等:210,缩略识别题名;222,识别题名;245,题名说明;246,题名的变异形式;250,版本项;256,计算机文件特征;260,出版发行项。

3 号字段,载体形态项:300,载体形态;310,当前出版频率;321,先前出版频率;362,出版日期/卷册标识。

4 号字段,丛编说明:440,丛编说明。

5 号字段,附注:500,一般附注;516,计算机文件类型或数据附注;520,提要等附注;538,系统细节附注;540,资源使用范围、复制的条件;546,语种附注。

6 号字段,主题标目:650,论题性标目;653,非控主题词;655,资源类型或形式。

7 号字段,附加款目和连接款目:700,附加款目—个人名称;710,附加款目—团体名称;720,附加款目—其他名称;786,资源来源款目;787,其他相关款目。

8 号字段,丛编附加款目:856,电子信息定位与检索。

由上述 MARC 字段构建的简要式 MARC 数据结构体系,具有如下几个重要特点:

(1) 由于简要式 MARC 数据结构体系所确定的 MARC 字段比较简洁,因此,记录和描述电子信息资源的复杂程度大大降低。采用简要式 MARC 数据结构体系对电子期刊、光盘数据库、电子论坛、电子报告等进行著录和标引,可操作性强,省时省力。

(2) 对于电子信息资源具有重要参考价值的信息,简要式 MARC 数据结构,能够通过若干子字段较为全面完整地记录和描述。如 856 字段包含了定位电子信息资源所需的各种信息,这些信息均可通过子字段加以标识,如子字段 \$u(电子资源地址)、\$d(路径)、\$f(电子文件名称)、\$s(文件大小)、\$z(注释)等。而“都柏林核心”的电子资源定位的资源识别字段难以将这些子字段信息详尽记录和描

述。

(3) 简要式 MARC 数据结构体系不仅覆盖了“都柏林核心”15 个字段的全部内容,而且对电子期刊和光盘数据库的出版周期、更新周期等一些特殊的外部特征在 3 号相关字段作了附注;对有丛书、丛刊的电子信息资源,在 4 号相关字段作了连接说明。

3.2 简要式 MARC 数据库的软件设计

电子信息资源简要式 MARC 数据库的建设,是根据电子信息数据源,按简要式 MARC 数据结构填写记录输入单,通过相应软件的支持,进行记录输入单的计算机录入、修改,并且形成由头标区、目次区和数据区组成的电子信息资源简要式 MARC 数据通讯格式(即电子信息资源简要式 MARC 数据库)。传统关系型数据库管理系统软件设计方法难以实现上述功能。在开发研究过程中,我们采用了先进的、技术含量高的 C++ 开发工具。其软件设计步骤可概括为:

(1) 将电子信息资源按 MARC 格式中字段定义的格式进行归类,确定电子信息资源中每部分所对应的字段名及子字段名,并将各个字段按字段名从小到大依次排序。

(2) 统计所有字段的数目和各字段数据的长度(包括字段结束标记),按 MARC 通讯格式规范计算出 MARC 记录的总长度和数据内容同记录头之间的偏移量,然后加上记录的属性就生成了固定长度的头标区。

(3) 根据 MARC 格式目次区的定义,将各个字段的字段名、字段长度、字段数据的偏移量分别填入目次区。其中偏移量为前一字段的字段长度加上其偏移量(其中第一个字段的偏移量为 0),然后打上目次区的结束标记。

(4) 将各个字段的内容依次填入数据区,并在每个字段和记录的结束处打上相应标记。

(5) 逐条处理电子信息资源,即可生成国际标准的 MARC 通讯格式。

在软件设计过程中,我们较好地解决了两个难点问题:一是,如何把格式各不相同的电子信息资源统一化?我们是将用户自由定义的电子信息资源的分类同 MARC 格式中的字段名及子字段名的关系保存在配置文件中,在主程序中读取配置对应文件。二是,由于写入 MARC 数据库文件时是先写入头标区,再写入目次区,最后写入数据区,但实际上是由数据区中的内容统计并生成目次区和头标区的。对于这个问题的解决,我们是先统计计算出头标区和目次区后再依次写入文件。

采用先进的 C++ 开发工具设计完成的简要式 MARC 数据库,在建库时,我们能够按屏幕提示根据实际需要进行增加字段、删除字段、删除记录、MARC 生成、重新检索、修改、查看详细信息等操作,使得电子信息资源的计算机录入比较方便、灵活, MARC 数据修改与维护较为容易、可靠,

MARC 数据通讯格式的生成更是快捷、高效。

3.3 数据库接口技术的开发

电子信息资源简要式 MARC 数据库,能够实现馆际间的网络传递与交流。要满足用户完全基于 Internet 的检索需求,不仅需要把电子信息资源简要式 MARC 数据库转换成 sybase 或 SQL Server 等大型电子信息资源数据库管理系统,更重要的是要进行数据库接口技术的开发,使电子信息资源数据库与 web 服务器互联,实现 web 电子信息资源数据库的动态查询。传统的方法是利用通用网关接口 (GGI) 开发方式。这种方式不仅开发难度大,而且服务器资源消耗也大,大大降低了网络的整体效率。使用新的组件 ActiveX 与 XML (Extensible Markup Language,可扩展标注语言) 新型的综合开发方式,能够获得更高的网络效率,更短的响应时间。采用这种新型的开发方式,其软件设计要点是:

(1) 将电子信息资源数据库中的数据以 DTD 方式简化为 XML 文档结构。DTD (Document Type Definition, 文件类型定义) 是 XML 3 个主要要素中的一个,它定义了电子信息资源数据库的元素、元素的属性及元素和元素属性之间的关系。

(2) 利用 ASP (Active Server pages, 活动服务器主页) 转化电子信息资源数据库为 XML 数据源。ASP 也是一种数据库接口技术,利用它建立电子信息资源数据库的连接和转化。这种方法有两点优势。第一,XML 数据源可以使用脚本来生成。这意味着可以便捷地转换大批量的数据。第二,XML 数据源是从电子信息资源数据库中的数据动态生成的,当电子信息资源数据库被更新时,XML 数据源也随着更新,而不需要额外的开销。

(3) 使用 XML —QL 查询语言编写查询程序,实现电子信息资源数据库 XML 数据源的查询。XML —QL 查询语言是在查询语言 (UnQL 和 StruQL) 基础上设计的。它有类似 SQL 查询语言 Select From where 的结构 (where In Construct)。按照这种结构方式编写检索程序,用户可从题名、关键词、主题词、出版者、发行者、分类号、ISSN、ISBN 以及具有检索意义的所有子字段单项及组配检索。更为重要的是,用户可通过提供的电子资源定位信息进行超链接,以较高的检索效率,极短的响应时间获取所需的电子信息资源。

(4) 将查询结果以 XML 文档的形式在浏览器端输出。这种显示方式的优点,一是将电子信息资源数据库查询结果的显示格式和电子信息资源数据库的数据相分离,实现对同一电子信息资源数据的多种处理,以满足不同用户的需求。二是节省服务器资源消耗,从根本上克服以 HTML 文档形式显示查询结果的局限性。

(下转第 72 页)

应该做到理论和实践相结合,只有这样,文献学才能具有强大的生命力。

《文献学导论》将文献学基础理论知识与实用知识有机结合,重视对文献利用的指导,这是其第二个特色。该书对文献的分类、社会功能、文献交流及其规律、文献工作标准化与现代化等文献学的基本理论问题做了深入的探讨,使读者能够更加深入地认识学习文献学的作用与意义。同时,为了体现文献学的实践性及《文献学导论》一书的写作初衷,即对读者进行文献利用的指导,使读者通过阅读全书后,能够顺利利用文献,作者非常注重实用性较强的文献学知识的介绍,重视对文献利用的指导。例如,在“文献服务”一章中,作者专门设立了“文献利用指南”一节,对图书馆目录的使用方法及使用技巧做了充分的介绍。在这一节中,为了指导读者更好地查找和使用因特网上的文献资源,作者还对因特网上文献的主要类型、因特网上文献的查找工具等内容进行了详细的介绍。再比如,在“文献生产”一章中,作者根据自己的研究成果,对非法出版活动与非法出版物进行了探讨,向读者详细介绍了非法出版活动和非法出版物的定义、非法出版活动的特征、非法出版物的主要类型、非法出版物的认定和取缔等内容,特别是对有关如何识别非法出版物知识的介绍,对于读者识别非法出版物,具有很高的实用价值。

3 文献学研究的继承与发展相结合,丰富和发展了文献学研究内容

我国的文献学研究,已经从侧重于以对古代文献的目录、版本和校勘为主要内容的古典文献学,发展到以对文献的收集、分类、交流和利用等为主要研究内容的现代文献学。这说明,文献学的研究内容是在不断变化和发展的。文献学只有不断丰富和发展其研究内容,才能长盛不衰。

《文献学导论》一书对文献学研究内容进行了较大的丰富和发展,这是该书的第三个特点,也是其创新性的体现。作者以文献及文献工作作为文献学的

研究对象,根据文献工作发展的实际,对文献学的研究内容进行了丰富和发展,并且取得了显著的成果。

该书对文献学研究内容的丰富和发展,首先体现在对文献学原有研究内容的不断深化上。随着计算机技术、光盘技术等高科技成果在文献学工作中的应用,新型文献载体不断出现。因此,对于文献载体的研究,就不能仅仅局限于对传统的简牍、缣帛和纸张的研究,还要对磁、光、电等载体进行研究,不断深化研究内容。《文献学导论》不仅对传统文献载体进行了介绍,而且对常见的新型文献载体,如感光膜载体、磁膜载体、光盘和电子纸等进行了介绍,使有关文献载体研究的内容更加丰富。

该书对文献学研究内容的丰富和发展,还体现在作者根据文献工作发展的实际,扩展文献学的研究内容、创造性地将非法出版活动及非法出版物、文献评价等内容纳入文献学研究的领域。非法出版活动是文献生产活动中的一股逆流。随着出版技术的不断提高,非法出版物的数量越来越多,对文献收藏和利用等工作的危害也越来越大。如果不对非法出版活动加以研究,将严重危害文献工作的发展。因此,将对非法出版活动和非法出版物纳入文献学研究领域是有其道理的。与此相类似,将文献评价工作纳入文献学的研究领域,也是与文献工作发展的实际相适应的。《文献学导论》一书对这些内容做了深入的研究,这是其它文献学著作所没有涉及的,对丰富和发展文献学研究内容做出了有益的尝试。

当然,作为一本颇具新意的文献学著作,不足也是难免的。由于该书在写作时比较注重实用性,因而全书对某些文献学基本理论问题的阐述还稍显不足。另外,虽然主编在全书的加工、润色和修改等方面做了大量的工作,但由于该书是多人分工写作,因而全书的体例仍不是十分统一,有些章节的论述仍不充分。

王贵彬 南京大学信息管理系研究生,发表论文多篇。
通讯地址:南京大学信息管理系。邮编 210093。

(来稿时间:2001-06-12)

(上接第 66 页)

参考文献

- 1 付立宏. 信息资源网络化与图书馆服务创新. 图书情报知识, 2000(3)
- 2 肖燕. 关于强化图书馆研究开发功能的思考. 中国图书馆学报, 2000(3)
- 3 林海青. 数字图书馆的信息组织. 中国图书馆学报, 2000(1)
- 4 贺亚峰. 都柏林核心集与机读目录. 江苏图书馆学报, 2000(6)
- 5 (美) Steven Holzner 著; 潇湘工作室译. XML 使用详解. 北京:

机械工业出版社, 1999

谈春梅 南京邮电学院图书馆副研究馆员. 通讯地址: 南京市. 邮编 210003.

叶继元 南京大学图书馆研究馆员, 博士生导师. 通讯地址: 南京市. 邮编 210009.

汪令全 安徽省马鞍山市电信局(南邮毕业生, 原课题组成员)工作. 通讯地址: 安徽省马鞍山市电信局. 邮编 243000.

(来稿时间: 2001-05-21)