●王兰成

图书馆、情报和档案学数字信息群的知识集成研究进展与关键问题*

摘　要　图书馆、情报和档案学大学科数字信息群的知识共享是学科各领域协调发展和跨领域知识集成的重要基础。实现学科数字信息群的知识共享集成和挖掘,需要创新研究相关关键技术。目前国际上的研究处于起步阶段。提出我国对数字信息群知识集成研究的方法和基本路向。参考文献 13。

关键词　数字信息群　知识集成　信息构建　信息技术

分类号　G250

ABSTRACT　The knowledge sharing of integrated digital information in library, information and archival sciences is the important basis for the coordinative development of various disciplines and the integration of cross-disciplinary knowledge. The author introduces the latest developments in the field and proposes some methods and directions for Chinese researchers. 13 refs.

KEY WORDS　Integrated digital information.　Knowledge integration.　Information architecture.　Information technology.

CLASS NUMBER　G250

图书馆、情报和档案学数字信息群(以下简称数字信息群)的知识共享是学科各领域协调发展和跨领域知识集成的重要基础,实现学科数字信息群的知识共享、集成和挖掘,需要创新研究相关关键技术。目前国际上已具备较成熟的学科知识集成条件,处于研究起步阶段,但在我国,缺少统一的学科理论和技术体系支持,该研究基本属于空白。

1　数字信息群知识集成的研究现状

国际上,匹兹堡大学信息科学学院院长 Ronald Larson 教授和卡耐基梅隆大学计算机科学学院副教务长 Howard Wactlar 认为,数字图书馆、数据网格和永久档案的整合工作正在活跃地进行,目的是为收集、出版、共享、储存和保存各种形式的信息提供一个可行的基础架构。当各个领域都聚焦于各自研究中至关重要的知识和功能时,知识集合共享原则就势在必行了。英国联合信息系统委员会成员和电子图书馆计划主任 Chris Rusbridge、伯明翰大学 Stephen Pinfields 教授、中国科学院孟广均教授都在研究复合型图书馆。学术界的基本观点认为,复合型图书馆是传统图书馆走向现代化的必然形式,复合型图书馆走向数字图书馆是一个长期过程,应加强相关的研究、探索和创新。

华人学者秦健和曾蕾博士认为,开放资源(open source)是数字图书馆的下一步工作重点之一,知识集成中元数据方面存在的问题是元数据标准太多,按简单的都柏林核心元素集来整合将失去很多细节和影响上下文的衔接;一些理论上很吸引人的模型,难以在实践中应用推广(如 METS, RDF);元数据库还存在各种质量问题,缺乏各种控制词汇和标准。斯坦福大学图书馆张甲认为,图书馆业务重点正从数据收集转向资源利用,从信息技术重点转向一体化服务,其要素包括跨系统连接、内容相关服务、读者资格认定和信息资源的组织及相关标准等。

1977 年,美国档案工作者协会(SAA)咨询委员会任命组成的"国家信息系统特别工作组"(NISTF),专门研究档案目录数据库的著录与档案信息的电子化共享等问题,建立了传统档案著录元素与 MARC 格式之间的成功映射,最终制定了适合美国档案目录数据库的机读目录交换格式。1982 年 10 月,该格式经美国档案工作者协会、美国国会图书馆及美国机读目录格式标准化办公室开发部正式批准,成为档案数据库著录的标准,即"档案机读目录格式标准"(MARC AMC),使得档案馆为 MARC 系统提供机读目录成为可能,为美国档案乃至整体信息资源的共享奠定了基础,而且能够被用作对其他非档案类资料的管理[1-8]。

在技术开发层面,采取开发主体上跨机构、开发对象上跨库、跨地区和跨资源的合作开发,能够最大限度地满足各方面利用者的多元化需求,图书馆、博物馆和档案馆等不同文献保管机构的馆藏类型互有交叉,突破机构界限可以更大程度地集中相同主题文献[9]。一个成功的案例是加拿大新的法律合并了国家图书馆、国家档案馆和肖像馆[10],其联合搜索系统能够使利用者一次性检索出所有的文献(档案)著录、数字实体和网络资源,从而建立起大学科信息群的最佳工作方式。

* 本文系国家社会科学基金项目(项目编号 05BTQ011)的研究成果之一。

研究表明,国内外尚没有发现通过DC(Dublin Core)或EAD(Encoded Archival Description)的兼容来集成学科领域知识的研究,也没有发现能够替代MARC描述学科丰富的馆藏资源的元数据,国内档案界缺少对档案MARC方面的研究。国内目前没有开展基于XML和相关元数据对跨领域的多种信息资源的知识集成进行研究和开发。

2 实现数字信息群知识集成的方法研究

我国迫切需要实现包括图书馆、情报学、档案学多种信息资源的知识集成,以促进国际上和国内各领域之间的信息资源共享和交流。

2.1 方法研究的主要观点

由于国内外文献界长期使用的MARC格式标准在网络应用方面存在明显缺陷,以XML(Extensible Markup Language)形式存储和管理元数据已形成共识。但国内外文献界主流没有淘汰MARC,而使它进入了一个新的发展阶段,如:国外已有实现图书、档案MARC数据的XML文件形式加以利用,国内图书馆界已有用XML研究对MARC书目数据的定义,国内外仍在使用丰富的MARC馆藏资源和建立书目MARC与DC元数据的转换等等。RDF(Resource Description Framework)仍需要领域提供更丰富的语义支持,其基础框架Warwick的容器结构可以包含DC、MARC等不同类型的元数据。

阶段性成果所形成的主要观点包括:

(1)MARC AMC是MARC在档案领域的具体应用。MARC与DC、MARC AMC与EAD格式的数据可以互相交换,互为利用和补充,共同形成了网络乃至整体利用图书馆、档案信息资源的系统的梯度检索体系。MARC AMC所提供的对档案目录信息的概略性著录对网络检索可以起到引导作用,而通过利用EAD所置标的检索工具的详细全文信息则可起到深入查检的作用[11]。MARC AMC格式使档案目录与图书目录信息发生了集成,构成了整体的目录信息系统,并构成了元数据复杂等级体系中的一个重要组成部分,它们可以向利用者提供信息资源的内容与本质,及与其他资源间的关系,即真正发挥了学科集成、共享的目录信息的文化与知识属性。

(2)定义XMARC文档指用XML语法来阐述CNMARC,它符合所有XML和RDF语法规范,其特点是能表达所有CNMARC数据中的信息,并且不受CNMARC中对字段大小和层次多少的限制,可以进一步扩展;其语义标记为检索提供了很好支持,针对标记的索引可以构造复杂的检索条件;其数据与其他网络系统的交换十分方便,为数据的馆际共享和与国际数据交换奠定了基础;可以充分利用现有资源,升级方便;很容易与DC、EAD等其他元数据格式实现互操作,使学科多信息资源在元数据级的集成成为可能。

(3)内容DTD/SCHEMA是关于XMARC内容描述的文档类型定义[12],它把MARC字段的字段说明作为其一个文档元素的名字,同时把该字段所包含的子字段的字段说明作为该元素所包含的子元素的名字。框架DTD/SCHEMA是关于XMARC资源描述框架的文档类型定义,它把MARC字段的字段说明作为文档元素字段的一个属性。该属性名为"字段说明",把子字段的字段说明作为文档元素子字段的一个属性,该属性名为"子字段说明",从而能够将各种文献对象的XMARC统一起来。

2.2 知识集成方法研究的基本路向

研究的思路是先基础理论研究后技术开发,先研发领域知识后进行学科知识集成实践,充分借鉴国外已有成果,吸收总结已取得的相关成果。基本方法是:

(1)对图书馆、情报学、档案学领域中馆藏、电子和网络数字信息群进行知识集成理论的探索与实践。

(2)系统地比较和研究MARC、CNMARC、MARC AMC、EAD、DC、METS、RDF等相关元数据。

(3)提出跨领域的多种信息资源的知识集成的元数据理论并进行具体研究。

(4)研究基于XML的领域内容和元数据框架的DTD方案和XML SCHEMA方案,设计元素可扩充定义的软件重用模板,以定义跨领域资源文档,根据元数据之间的对应关系(mapping)进行相互转换。

3 实现信息群知识集成的技术研究

我国迫切需要创新研究学科数字信息群的知识动态集成方法与技术,以最大限度地支持信息服务和实现知识的价值。

3.1 技术研究成果的主要观点

B. Inmon定义数据仓库DW(Data Warehouse)是面向主题的、集成的、稳定的、随时间变化的数据集合,用以支持管理决策的过程。实现学科多种信息资源共享的关键是知识集成,Web Services和数据仓库则是一个综合的解决方案。Web Services技术为各领域的知识集成提供了完整的体系结构;采用数据仓库的数据处理技术能够对各领域的异构异种海量数据进行事前处理,以进一步提高数据存储和查询效率;基于数据挖掘和知识本体技术,能够从海量数据中有效提取隐含的、潜在有用的信息和知识,能够跨领域对知识进行编码。如:主题块数据的知识自动标引在减少分词歧义性、缩短标引时间方面有许多潜力可挖,以知识本体构造、维护基于XML的知识词典,实现跨领域的知识检索。目前国外数字图书馆系统中见有相关成果介绍,国内没有系统的研究成果。

阶段性成果所形成的主要观点包括:

(1)通过知识本体技术,将学科信息群的知识按照不同主题进行组织,形成等级类目,同时规定类目的特性及其之间的关系。利用主题概念语义关系,在改进的整词二分分词词典上采用区间最大词长,研究设计预处理特义中文禁用字词的切分方法和长词匹配短词推进的中文抽词标引方法,运用于学科数字信息群的知识集成能有效地减少领域的分词歧义性和缩短标引时间,使研究更有实用价值[13]。

(2)知识集成是指对已经存在的多个异构数据库，在尽可能少地影响其本地自治性的基础上，构造具有用户所需要的某种透明性的分布式数据库，以支持对物理上分布的多个数据库的全局访问和数据库之间的互操作性。应立足于对书目数据和档案数据的知识集成，在集成架构中加入数据仓库元素，利用数据仓库对集成的知识数据进行统一视图的组织和管理，利用知识集成的元数据对跨领域数据进行描述，并在该元数据基础上建立各类视图。

(3)文档与知识是有区别的。目前许多搜索引擎的数据搜索都是位于网络表层的静态信息，无法挖掘到位于数据库里的深层数据，从而面临着3个难题：一是如何从数据库得到请求响应，二是如何将搜到的数据进行组织，三是如何整合这些信息并呈现出来。第二代相关性技术根本无法做到这一点，智能化技术根据关键词和内容之间的关系来确定则有可能实现。

3.2 知识集成技术研究的基本路向

本课题开发构建跨领域信息整合的数据仓库平台，研制跨领域中文信息处理通用计算机系统，示范性地开发学科信息群共享与集成的实验系统。这无论是对于促进和推动领域数字信息资源发展还是提高学科信息利用的质量，都有十分迫切、重要的理论意义和应用价值。根据重大信息系统项目的需求在实际应用研究方面取得突破。基本方法是：

(1) 以知识本体构造基于XML的知识词典，主要用等级结构直接显示主体概念之间的关系，按学科体系进行系统排列，为知识检索提供支持。

(2)研究数据仓库在知识集成中的应用模式，利用Web Services建立学科数字信息群的知识集成架构。各数据库系统的知识集成，可以利用Web Service按照统一的数据格式提供出来，集成时只需调用各馆(部门)暴露数据，不必关心所采用的数据库系统甚至操作系统方面存在的异构。

(3) 基于数据挖掘对跨学科领域知识进行提取和整合技术研究。传统的基于关键字匹配或基于学科分类的检索工具不能令人满意的主要原因之一，是它们无法挖掘概念之间的内在联系，搜索出更深层的信息联系，而采用本体论可以达到这一目的。

(4)综合以上成果，研究与开发学科数字信息群的知识集成实验系统。

本文试图研究新的基于多种学科数字信息资源的描述机制，这不仅是我国图书馆学、情报学、档案学数字信息群的知识共享研究的需要，而且能够奠定大学科知识集成的理论基础，将大大拓展学科各领域的研究内容，推进学科知识的利用水平，促进我国相关标准的尽早制定与完善。

本文还试图研究运用国际上领先的信息技术，研制跨领域中文信息处理通用计算机系统，示范性地开发图书馆、情报学、档案学学科信息群共享与知识集成的实验系统，这无论是对于促进和推动领域数字信息资源发展还是提高学科信息利用的质量，都具有重要的应用价值。这方面重要实践成果目前在国内学科研究领域还没有发现。

参考文献

1 张甲，秦健，曾蕾. 数字图书馆的下一步. 数字图书馆前沿问题高级研讨班，深圳，2004
2 Elements. Qualifiers and Schemes(RDF)(2000, July) Retrieved August 30, 2000 from the WWW: http://www.oclc.org/oclc/corc/documentation/index.htm
3 Ronald Larsen, Howard Wactlar. Knowledge lost in information, Report of the NSF Workshop on research directions for Digital Libraries. Junel5-17, 2003, Chatham, MA
4 Caplan, P. (1996). Metadata for Internet Resources: The Dublin Core Metadata Elements Set and Its Mapping to USMARC. Cataloging & classification quarterly, 22(3-4), 43 (16pages)
5 Margaret Procer and Michael Cook. Manual of Archival Description. Third Edition, 2000
6 Wayne Eckerson, Colin White. Evaluating ETL and Data Integration Platforms. Retrieved from the WWW: http://www.dw-institute.com
7 Zachary G. Ives, Efficient Query Processing for Data Integration, University of Washington, 2002
8 Busse S., Kutsche R., Leser U. Federated information systems: Concepts, terminology and architectures Brelin: Techniques University, 1999.9
9 冯惠玲. 档案信息资源的有效开发. 国家档案局：中国档案信息化发展战略论坛，2005-06
10 http://www.collectionscanada.ca, 2005-06
11 王兰成. 基于SCHEMA模式的XAMC信息描述及其检索实现研究. 情报学报，2004(4)
12 王兰成，冯文杰. 两种MARC的XML DTD信息描述机制研究与比较，中国图书馆学报，2004(1)
13 王兰成. 数字图书馆XMARC文档的关系模式与数据获取研究. 计算机科学，2004(10)

王兰成 解放军南京政治学院上海分院军事信息管理系教授，博士生导师。通信地址：上海市。邮编200433。

(来稿时间：2005-07-28)