●王兰成 敖 毅 曾 琼

国外知识组织技术研究的现状、实践与热点*

摘 要 国外对知识组织的研究和实践已经进行得比较深入。知识组织技术的研究进展表现在:突出以语义网为代表的相关技术研究;以知识组织系统推动知识组织发展;以搜索引擎为代表的知识组织方式研究。国外的知识组织技术项目主要有 DSPace 项目、Haystack 项目、SIMILE 项目、Bricks 项目和 Corrib. org 项目。从文献组织到信息组织、从信息组织到知识组织,以图书馆、档案馆和情报中心为主的信息机构在这样的变迁过程中,具备了充分的理论基础并积累了大量的实践经验。参考文献 7。

关键词 知识组织 知识组织技术 信息集成 信息机构

分类号 G254

ABSTRACT In foreign countries, there have been deep researches in knowledge organization. In knowledge organization technologies, recent advances include semantic web and related technologies, pushing the development of knowledge organization with knowledge organization systems and researches on knowledge organization methods with search engines as representatives. Major knowledge organization projects in foreign countries include DSpace, Haystack, Simile, Bricks and Corrb. org. From document organization to information organization and the to knowledge organization, libraries, archives, museums, information centers and other information institutions are accumulating many practical experiences through such a transformation. 7 refs.

KEY WORDS Knowledge organization. Knowledge organization technology. Information intergration. Information institution.

CLASS NUMBER G254

国外对知识组织的认识比较早,在研究的领域、开展的项目以及实践等方面已经进行得比较深入。表现为有四股力量在推动整个知识组织研究的进行。一是以 W3C、ISKO 为代表的国际性组织,提出了一系列的技术标准和规范,并在全世界范围内,以会议、项目、研讨等形式进行推广;二是以美国国会图书馆、英国大英图书馆等大型图书馆、博物馆、档案馆为主的信息机构,他们积极地在自己的领域进行试验,并取得了一定的成果;三是以麻省理工学院、哥伦比亚大学、康奈尔大学、杜克大学、格但斯克技术大学等大学及其研究所为代表的学院派,借助大学本身雄厚的技术实力和大学图书馆对知识组织的迫切需求,开展了一系列的研究项目;四是以 Google、IBM、Microsoft、HP 等公司为代表的商业组织。一方面他们具有很强的技术研发实力,另一方面他们也掌握着主流的计算机软件产品,对知识组织研究的影响很大。最值得关注是这几股力量之间相互交叉,相互支持。往往是商家出钱、学校出智力、信息机构提供实践的场所而国际化组织进行推动,这样形成了一个良性的循环,对于知识组织相关技术、标准、规范、体系结构的研究起到了很好的推动作用。

1 知识组织技术的研究进展

国外在知识组织技术的研究进展主要体现在以下几个方面:

(1)突出以语义网为代表的相关技术研究。这方面的研究进展主要包括元数据理论,DC 元数据体系及其使用,使用 XML、RDF、OWL 构建 ontology 等方面。形成了以 W3C 为龙头,聚集一批世界一流的大学院校、实验室、公司形成的研究团队,并致力于推动 Internet 从第二代"互联网"向第三代"语义网"发展。

(2)以知识组织系统推动知识组织发展的研究。以美国国会图书馆为代表的信息机构提出,在数字化过程中要以知识组织系统来推动对数字信息的组织,实现从信息组织向知识组织发展,包括对词表、分类法、叙词表、概念图等的架构方式、组织方法、构建理论和应用。在这一方面,图书情报档案界大量借鉴了语义网研究领域的成果和技术,并积极地参与到相关项目的研究和实践中去。

(3)以搜索引擎为代表的知识组织方式的研究。随着 Internet 的发展,以搜索引擎为代表的知识组织

* 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。

方式也是一个重要的发展方向，通过对如自动分类、自动聚类、自动标引等自然语言处理技术的进一步研究和实践，期望对网络上浩如烟海的信息进行再加工，以方便人们利用。这方面贡献最大的莫如目前网上最大的搜索引擎公司 Google。

21 世纪以来，国内对知识组织技术的研究逐渐走向深入。以蒋永福、王知津、李秀云、王子舟等人为代表的一批学者，对知识组织的相关概念进行了深入的探讨，张晓林、盛小平、王军等学者的加入，拓宽和加深了知识组织的研究广度和深度，旅美华人学者曾蕾、张甲、秦健也不断地将国际上知识组织的相关信息介绍回来，这样，在国内图书情报学界掀起了一股知识组织研究的热潮。国内学者认识到了知识组织的重要性，在国家层面已经认识到了知识组织相关方面的重要性，先后实施了中国高等教育文献保障系统[1]（CALIS）和中国知识基础设施工程（CNKI）[2]等项目，并且这些项目已经取得了一批成果。

但是，国内对知识组织技术的研究呈现出两级分化的趋势：一方面是在图书情报学界，大量的论文局限于对知识组织的相关理论、概念进行反复的讨论和研究，却很难得出令人信服的结论；另一方面在技术层面，虽然也引进了如 ontology、概念图等代表国际先进水平的技术，但是大多是简单的模仿和重复，还缺乏有说服力的研究项目和研究成果。另外，从台湾学者发表的论文的引文看，两岸在知识组织方面的理论交流也有所展开，总体来看，国外的研究成果和研究项目比国内要在理论深度和可操作性上更胜一筹。

2 国外有关知识组织技术的项目实践

国外有关知识组织技术的项目建设主要有：

2.1 DSpace 项目

DSpace[3]项目是由惠普公司实验室和麻省理工学院图书馆联合开发的一个智能化动态数字对象存储系统，用于实现对多个分布式成员单位的数字资源进行仓储式管理。2002 年 11 月 4 日，DSpace 系统正式对外公布，2006 年 12 月 7 日发布最新的 1.4.1 版。DSpace 是一个以内容管理发布为设计目标，遵循 BSD（Berkeley System Dislribution）协议的开放源码数字化存储系统。该系统可以收集、存储、索引、保存和重新发布一个组织的研究资料。

该项目能够满足各类数字化信息机构的需要，包括构建机构库（Institutional Repositories）、学习对象库（Learning Object Repositories，LORs）、电子论文（eThesis），进行电子记录管理、数字保护、出版等等。DSpace 接受所有格式的数字化资料，包括文本、图片、视频以及音频文件等，内容也包括文章、技术报告、工作报告、会议报告、论文、数据集、图像等，并会将这些资料进行统一的标引以方便用户检索。在 DSpace 中，把不同的机构称为一个社区（Community），把其提交的数字化资料称为馆藏（Collection），把描述馆藏属性的条目称为项目（Item），这些项目由 DC 元数据（Dublin Core Result）来描述，项目再分为数据束（Bundle），数据束由数字流（BitStream）组成，数字流是不可以再划分的最小的描述单位。

DSpace 的整个技术框架分为三层（见图 1），分别是存储层、事务逻辑层和应用层。处于底层的是存储层，它实际上是一个文档管理系统，用 PostgreSQL 关系数据库管理系统进行管理，整个系统主要代码采用 JAVA 语言编写，选择 JDBC（JAVAAPI）驱动连接数据库；事务逻辑层是 DSpace 功能集成的地带，包括工作流、内容处理、系统管理及检索、浏览和授权等模块，每个模块提供 API 允许 DSpace 用户按自己的需要修改或扩展功能；应用层为用户的使用界面，主要功能包括 CNRI 资源调度系统、WEB 用户界面 OAI 协议、输入输出数据项等。

对于数字材料的存储，DSpace 提供了“数字流保存”和“功能性保存”两种模式。前者是把一个数字文档按照原始形状保存起来，可能数年后不再有软件能读出来，只有使用特殊的软件或者由特殊的专家对其进行编码才能辨认出来；后者则要求在技术格式和媒体介质变化的情况下始终保持其可用性。显而易见，“功能性保存”是比较理性的存储模式，但需要更多的经费支持。DSpace 把文件格式划分为三个层次：①支持的格式，DSpace 利用格式迁移技术对这种文件进行功能性保存；②知道的格式，此格式的文件意味着无法完成功能性保存，但是作为一种流行的格式，可以尝试通过第三方提供的转换工具完成格式迁移从而实现功能性保存；③不支持的格式，此格式意味着 DSpace 按照目前的技术不能完成功能性保存。DSpace 采用分级权限的管理体系，每种用户都有自己特定的使用范围，每种资料也都有各自的使用权限。匿名用户可以检索、浏览和下载资料；授权用户还可以向 DSpace 提交自己的材料，如课件、论文等，经过专业的图书馆员审核和元数据编辑，就可以方便地归入馆藏。

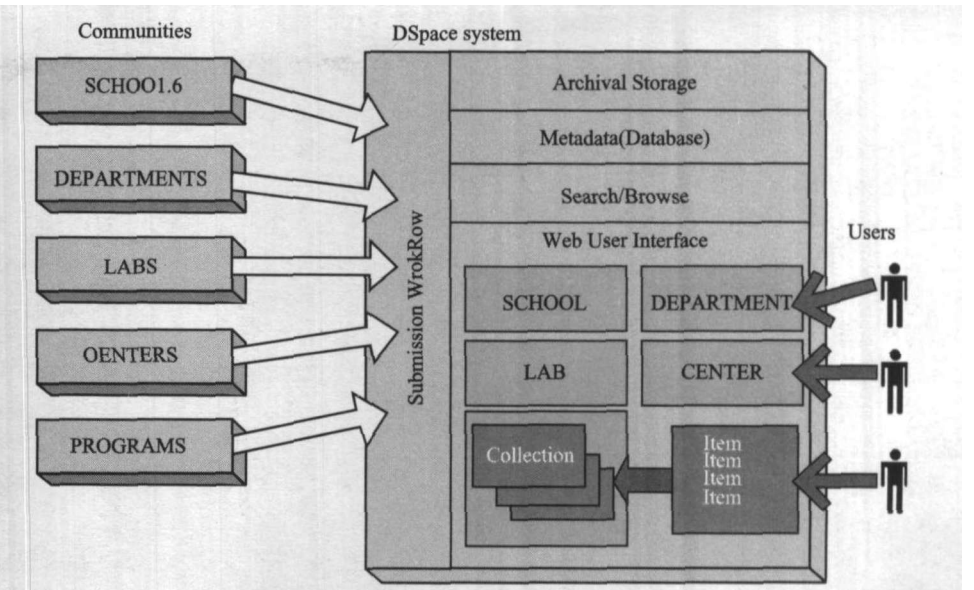

图1 DSpace 技术框架

DSpace 提供了一个比较完善的数字图书馆系统,包含了从资源著录到内容发布的主要功能。图书馆进行二次开发的门槛较低,对技术支持的力度要求较小,因此受到了普遍的欢迎。

2.2 Haystack 项目

Haystack[4]是麻省理工学院计算机科学与人工智能实验室(Computer Science and Artificial Intelligence Laboratory,CSAIL)的一个项目,目的是研究如何使人们能够以最合理的方式来管理自己的信息。

通常的应用程序都是处理由开发者定义的信息类型和关系,Haystack 项目抛开了这由应用程序设置的障碍,其目标是让用户在信息的视图之间定义他们最有效的安排和连接。这种个性化的信息管理将极大地改善每个人查询的能力,能够在需要时找到其需要的东西。目前的研究集中在以下几个方面:

(1)通用信息客户端:Haystack 客户端提供管理每天的信息的能力,例如安排约会、阅读和创建 email、组织相册等等。它探寻通过将信息在一个单一的地方进行处理,使得用户将注意力放在信息而不是使用程序上。

(2) 基于关系的探索:为 Eclipse 设计的 Relo 插件用于帮助用户探索和理解大的信息空间的一部分。项目目前的关注点在软件领域。

(3) 重新找到前面看过的信息:为 Firefox 浏览器设计的 Re:Search 扩展使得用户可以将他们过去的检索结果无缝地整合到当前的检索结果中。

(4) 轻量级结构化资料发布:Exhibit 是一个轻量级结构化资料发布框架,使用户只需通过编写 HTML 和一些 CSS 以及 Javascript 代码,就能够创建支持排序、过滤以及丰富视觉表现的网页。

(5) Steroids 上的标签:Piggy-Bank 扩展被设计成用于让 Firefox 浏览器的用户能够收集和浏览从原始网页上链接来的"语义数据"(以 RDF 格式)。

2.3 SIMILE 项目

SIMILE[5],即不同环境下的元数据语义互操作(Semantic Interoperability of Metadata and Information in unLike Environments,SIMILE),该项目最初的灵感来源于 DSpace,有四个团体对这个项目进行了资助:惠普研究实验室,W3C,麻省理工学院图书馆和麻省理工学院计算机科学与人工智能实验室。

由于现在每个数字图书馆的使用方式都有所区别,这给用户使用数字图书馆带来了极大的麻烦。另外,像 DSpace 这样的数字图书馆系统,当扩展到一定的程度时,必然会面临着需要进行元数据互操作的问题。在语义网时代,RDF、RDFS 和 OWL 为人们使用本体来描述他们的元数据,并使这些元数据能够得到更普遍地复用创造了条件。但是因为绝大多数的人都不是经过培训的语义网开发者,他们需要一些能够帮助他们完成这些工作并且进行评估的工具。SIMILE 项目的目标就是试图解决以上这些问题。其主要目的是扩展 DSpace,强化其支持任意的模式(Schema)和元数据,并提供一个分发数字化馆藏资产的架构。虽然 SIMILE 项目最初是面向图书馆界的,但是其开发的工具也可以很容易地在其他具有相似问题的领域进行使

用。由于本体定义专家的匮乏，要实现创建RDF，并且转换现行的基于XML的元数据为RDF格式就需要相应的工具，SIMILE项目就试图为元数据专家（例如图书馆员）制造高质量的RDF提供工具。

为了让图书馆用户能够实际地浏览RDF元数据，SIMILE开发了一整套网络应用来实现使用标准网络浏览器浏览RDF的功能。这套工具由Longwell和Knowle组成。Longwell是一个浏览器，向用户屏蔽了底层的RDF模型；而Knowle是一个以节点为中心的图导航浏览器，它的目标用户是那些想看看或者调试底层RDF模型的人。Longwell工具需要对被测数据结构有全盘的了解，但一般情况下是很难得到一个RDF模型的全局视图的，而且没有多少工具可以对RDF模型进行概括并快速的给出一个将被操作数据的模型。Welkin项目就是为此而创建的。Welkin是一个交互式图形化RDF浏览器，不需要预先配置就能够将任何RDF模型进行可视化（有点像Knowle，但不像Longwell），并且将RDF显示为一群节点和弧线的集合。Welkin（由来自意大利帕维亚大学的志愿者Paolo Ciccarese编写）对理解和挖掘不熟悉的数据集的布局特别有用。不像其他作图方法关注使用复杂的布局算法来得到合适的图像，welkin试图使用一种交互的方法增强用户的能力，允许用户挖掘、缩放、拖动、选择、聚集、过滤以及突出显示图中的节点和弧线。

还有一个问题是如何将已有的XML数据集转换为RDF格式。现在缺乏能够让人眼前一亮的对XML数据集进行浏览的工具。Gadget通过提供数量、唯一值以及XML属性中唯一值的百分比等概要信息来帮助数据管理员理解XML数据集的结构。当一个数据集没有模式，或者要了解数据集的模式哪一部分可能被使用时，Gadge-t也很有用（对简化转换步骤非常有用，因为能够避免转换那些根本不可能用到的部分）。RDF最强大的地方在于其提供的模型定义和其模型的高分布性特性。但是，在RDF/XML序列化方面一直被XML社群和挖掘模型能力感兴趣的潜在用户认为不够友好。基于以上原因，SIMILE开始了一个新的项目DEFizers，用于创建和分类软件工具和脚本，它能够将数据从现有语法转换为RDF。语义网的发展现在面临着“没有好的应用程序就不可能创建更多的RDF数据；但没有更多的RDF数据就不可能创建好的应用程序”的困境。通过让专家（例如图书馆员或其他元数据专家）使用这些工具更容易地将普遍和广泛有效的元数据源转化为RDF，为解决这类“鸡与蛋”的问题提供了一种现实方法。

除了开发自己项目的工具和软件以外，SIMILE也在支持MIT的Haystack项目的工作。在为SIMILE和Haystack开发RDF浏览时，发现如果有一个通用的本体操纵怎样显示RDF，将对开发RDF的浏览工具更为有利。Fresnel就是这样一种本体，它是一个描述如何用人类友好的风格提交RDF的通用本体。通过针对RDF的样式表（s-tyle sheet），可以对如何向用户表示某些抽象数据进行限制。

2.4 BRICKS项目

整合文化知识服务资源建设（Building Resources for Integrated Cultural Knowledge Services，BRICKS）是欧盟资助的一个项目[6]，目的是为共享与开发数字文化资源研究和实现先进的开源软件解决方案。BRICKS基础软件基于LGPL协议，是一个开源的软件框架，目的是构建分布式数字图书馆管理系统。项目从2004年1月1日开始，到2007年6月30日结束，共持续42个月投资1220万欧元。目前，已经发布了第二个原型系统。

BRICKS整合项目是欧洲第六届框架程序会议选择过程中的绝对胜出者，被选为欧洲文化领域最重要和最具创新性的ICT项目。BRICKS的目标是整合已有数字资源到一个公共和共享的数字图书馆中，一个涵盖了“数字博物馆”、“数字档案馆”以及其他数字记忆系统的概念，项目的成果将成为一个工厂的主要财产，在项目期间，该工厂得到了欧盟及其成员的资助，以后将自己维持。

BRICKS行动被构架为一个将共享的数字遗产的影响最大化的过程，不过其尊重欧洲文化存在多样性。项目采用由下到上的方法，根基于本地系统的动态社团的交互能力，最大化使用已有资源BRICKS将在以下三个方面作出贡献：①调整数字领域记忆机构的任务；②为开发数字文化内容研发一个共享的视图；③为建设一个可以互操作的合作的文化资产而鼓励文化合作。

BRICKS委员会是一个世界性的文化遗产机构，是研究组织，技术提供者以及数字图书馆服务领域的其它角色的联邦。委员会面向和验证项目成果，并且合作推进创建BRICKS文化遗产网络，将提供获取和维护欧洲数字记忆。

2.5 Corrib.org项目

Corrib.org[7]是一个聚集项目（cluster project），最初是从国际数字化企业研究协会（Internatnal Digital Enterprise Research Institute，DERI International）和波兰的格但斯克技术大学（Gdansk University of Technology）的合作开始的，致力于语义网相关技术的研究。

Corrib.org 中目前共有 6 个项目,主要是两家合作进行研发的各类开源项目,研究领域集中在以下 5 个方面(括号中为相应的项目名称):①数字图书馆(JeromeDL,MarcOnt);②信息检索(S3B,JeromeDL);③用户管理和 DRM(FOAFRealm);④分布式系统(点对点)(HyperCuP);⑤eLeaning(Didaskon,S3B)。

3 国外有关知识组织技术研究的热点

从国外知识组织的研究项目及发表论文看,当前国外对知识组织研究热点主要集中在以下几个方面。

(1)对 DC 元数据的深入开发使用

DC 元数据自问世以来,由于其简单易用,并在揭示不同学科领域的信息内容上都能够发挥作用,因此很快得到了广泛的应用。从最初的 15 个元数据到后来的堪培拉修饰词,DC 的内涵越来越丰富,能够表达的语义含义也越来越多,应用的范围也越来越广,并且被寄予了作为各类元数据方案进行互操作的中介的厚望。从 DC2006 会议的相关文献和发言中可以看出,DC 逐渐在向精致化、复杂化的方向发展,其体系已经远远超出了最初 15 个元素时的范围,并且在越来越多的项目中,都以 DC 作为元数据互操作的中介,通过实现各类元数据与 DC 之间的语义映射来达到互操作的目的。这一方面对在更广泛的领域进一步发挥 DC 的作用起到了很好的推动;另一方面,越来越精致复杂的 DC 会不会背离其简单易用的初衷,从而向 MARC 一样变成专用工具,也引起了人们的注意和担心。

(2)对语义网技术的研究

Robert Lee 提出了语义网的概念,并且设计了语义网的多层架构后,语义网的研究就成为了一个热点。特别是现在 Internet 得到普遍应用后,人们越来越发现,需要不是信息。因为现在信息已经太泛滥了,他们更需要的是能够直接满足要求的知识。对语义网的研究现在有几个热点:一个是 ontology 的构建。随着 W3C 的 OWL 语言的推出,已经出现了像 Jena、sesame 这样开源的开发包对其进行支持;还有关于 ontology 的存储、RDF(S)的存储、检索的研究也很集中。这主要是因为目前使用文件形式进行存储的效率太低,也不利于检索利用。随着各类关系数据库管理系统纷纷在最新的版本中推出了对 XML 的直接支持,这一点上应该有较大的改善。按照 Lee 对语义网的规划,在 ontology 以后,应该是 Logic 和 Proof,这方面已经有人开始研究,但是还处在比较初级的阶段。

(3)对元数据互操作技术的研究

元数据在网络时代的重要性越来越显现出来了。但是各种元数据方案的层出不穷在带来了方便的同时也带来了烦恼,就是关于元数据间的互操作问题。从目前掌握的文献看,已经在很多项目中使用 DC 作为元数据操作的中介。这一方面是 DC 本身固有的这种特性所决定的,另一方面,如何避免在向 DC 映射过程中造成的语义损失,以进行"无损"的互操作还有待进一步的研究。另外,现在也出现了另外一种研究方向,就是定义元数据的 ontology,通过 ontology 来达成语义的一致性,从而实现不同元数据方案之间的互操作,这可能也是一条值得探索的路子。

(4)对数字信息资源整合的研究

对数字信息资源整合的研究起始于企业信息化领域,随着各种信息系统的使用,在企业经营过程中积累了不少各种类型的历史数据。将历史数据集成起来,从中发现企业运行中的知识,是企业信息化的迫切需求。相同的需求也出现在信息机构中。对数字信息资源整合的研究集中在实现异构数据集成,通过 Web Services 实现应用集成,以及使用 XML 对数据进行标准化描述。目前在企业应用方面已经有了一些成熟的解决方案,为信息机构实施数字信息资源整合提供了参考。

参考文献

1 CALIS 简介. [2007-04-15]. http://www.calis.edu.cn/calisnew/calis_index.asp?fid=1&class=1
2 CNKI 网格资源共享平台. [2007-04-16]. http://tpi.cnki.net/produces_doc/grid.htm
3 [2007-04-20]. http://www.dspace.org/
4 [2007-04-22]. http://havstack.csail.mit.edu/
5 [2007-04-26]. http://simile.mit.edu/
6 [2007-04-26]. http://www.brickscommunity.org/
7 [2007-04-28]. http://www.corrib.org/

王兰成 南京政治学院上海分院信息管理系教授,博士生导师。通讯地址:上海市江湾五角场。邮编 200433。

敖 毅 曾 琼 南京政治学院上海分院军事信息管理系博士研究生。通讯地址同上。

(来稿时间:2007-07-10)