

图书馆集群信息资源整合分析

范小松(乐山市图书馆 四川乐山 614000)

摘要 从元数据编码角度指出, 只有在不同信息系统中, 采用新一代的 XML 编码标准, 才能实现图书馆内跨资源类型和载体格式的多种元数据整合; 同时从元数据与馆藏实体分离的角度, 论述了在图书馆集群中, 构建书目和读者数据中心, 从而实现集群外资源的整合和共享。

关键词 图书馆建设 资源整合 资源共享

1 信息资源元数据分析

对于元数据, 最本质的理解是“关于数据的数据”(data about data)。所有的信息系统, 都需要对管理对象进行相应的描述, 以揭示这些对象的内部和外在特征以及其属性。这些描述, 其实就是关于对象的元数据。

图书馆界是最早采用和推广元数据规范的行业, 且在元数据规范和标准的研究和使用方面, 大大领先于其它行业。MARC (机读目录), 从 20 世纪 60 年代由美国国会图书馆确定至今, 已成为图书馆界应用最广、数据量最大的书刊资料对象描述的元数据规范。

然而, 随着信息化进程的发展, 图书馆已不再局限于传统纸质载体的书刊资源管理。越来越多的非书资源、数字资源被纳入了图书馆管理的范畴。很多新建的图书馆, 非书资源、数字资源所占的比重, 已超出了传统纸介质书刊。MARC 这种过于专业和复杂化的元数据虽经几次完善, 仍不能很好实现非书资源或数字资源的有效描述。因而当前图书馆界已在 MARC 体系外, 关注或研究更适应馆藏信息资源描述的元数据体系——其中, 都柏林核心 (Dublin Core, 简称 DC) 元数据以其更通用、适应面更广的元数据元素集以及更科学合理的限定词扩展原则, 成为当前最有代表性的元数据体系, 稍有影响的元数据体系多可看出具有 DC 元数据的影子。

DC 元数据为了更通用, 更适应各个信息系统, 提出了“语法无关原则”(syntax independence); 即 DC 只规定元数据元素的基本语义, 而没有规定某种具体的元数据编码格式, 允许在各类技术平台及应用中采用多种方式对 DC 元数据实现编码。正是因为“语法无关原则”的影响, 导致即使想采用某种规范编码的系统开发商无据可依, 更何况还有很多只图自己方便, 采用自己随意而定的私有编码的系统开发商了。所以, 当前图书馆界, 除了因为早已成为编码标

准的 ISO2709 的成熟应用, 能方便实现内部与外部书目元数据的整合外, 不同类型和不同载体格式的资源管理系统及其描述元数据体系, 由于缺乏一个大家认可的编码方案, 想要实现资源整合非常困难。

有鉴于此, 以美国国会图书馆为代表的, 致力于国际图书馆界信息规范研究和推广的单位及部门, 投入到 MARC 元数据新一代的编码方案, 以及其它元数据最佳编码方案的研究中。

随着 XML (Extensible Markup Language, 可扩展标记语言) 标准的确立, 一个更有利于资源信息表达、更具可读性、更有利于数据交换的元数据编码标准进入了图书馆界视线。美国国会图书馆在几年前确定了 MARCXML 规范, 并在其系统中实现了 MARCXML 格式数据的输出——美国国会图书馆的 Z39.50 服务器, 现已提供 MARCXML 数据的检索和下载。事实上, MARCXML 兼容于 ISO2709, 仍采用 MARC 的数字与字母标识体系, 只不过是更具结构化、更可读的 XML 格式而非复杂、机读 (人不可读) 的 ISO2709 格式进行编码。2006 年 7 月 24 日, 由丹麦国家图书馆发起, MARCXML 已被国际标准化组织批准注册为国际标准草案 (ISO/DIS 25577)。预计不久的将来, ISO25577 将取代 ISO2709 成为图书馆书目元数据更有效的交换标准。

迄今为止, XML 已经成为了 Web 上最理想的数据表达方式和数据交换的事实标准。所以信息界 (包括图书馆界) 新兴应用的信息系统及其元数据体系, 也多采用 XML 标准实现元数据编码。MARCXML 或 ISO/DIS 25577 编码方案的出现, 打破了 ISO2709 格式 MARC 元数据与 XML 格式其它元数据间的隔阂, 虽然 MARC 元数据是采用数字和字母标识符, DC 等元数据是采用表义词汇标识符, 但这些不同的标识体系对于计算机来讲实质都是一样的, 不需要将标识体系统一才能实现元数据整合。事实上, 由于采用 XML

编码标准,不同的元数据即使标识符不一致,都可以通过 XPath (XML Path Language, XML 路径语言) 这个统一通用的标准,来实现数据寻址,从而达到信息系统中索引抽取、内容显现的目的。

所以,只有将馆藏所有资源元数据体系采用 XML 格式编码,才能有效地实现书刊资料、声像、实物等非书资源,甚至于馆建读书论坛、专业学术论坛中的帖子资源的元数据整合,从而实现跨资源类型和载体格式的馆藏信息资源的整合。

2 常用信息系统接口分析

在单一的系统内部甚至相同系统之间比较容易实现单一类型资源整合,采用 XML 编码标准,可进一步拓展为跨资源类型、跨载体格式、跨元数据体系的资源整合。但是,对于异构系统间,由于无法保证获取对方系统内部体系和数据结构,想要实现单一类型资源的整合比较困难,更何况想实现跨资源类型、跨载体格式的资源整合。

Z39.50 协议就是为了打破异构系统这种数据互操作障碍而提出的系统接口标准。图书馆界应用 Z39.50 协议已经多年,因为 Z39.50 协议的下载部份规定比较统一,所以服务器端利用 Z39.50 协议发布书目信息、前端下载服务器端数据的应用比较成熟。但 Z39.50 协议上载部份是通过 Extended Service 中的 Update 功能实现的。顾名思义,Z39.50 Extended Service 就是让大家自行扩充的部分。当前,国际上较为成熟的 profile 是 UCP (Union Catalogue Profile, 联合编目规程),如果开发商不利用成熟的 profile 加以约束,而随意扩充,就会产生较大的差异,这也是当前国内宣称支持 Z39.50 协议数据上载的系统多只是局限在一个小范围甚至变相的同构系统用户群体内(统一采用某个开发商的上载协议模块)应用的原因。

Z39.50 协议上载规定不统一导致系统兼容性差,加之 Z39.50 也是一个非常复杂、软件开发成本很高的协议,所以在图书馆界内部,Z39.50 接口尚不能很好承担起异构系统间的数据互操作,更何况那些不知 Z39.50 协议为何物的其它信息系统。事实上,当今最开放、最标准的接口非 Web Service (网络服务) 莫属,Web Service 接口是一套与平台和编程语言无关的标准集,它定义了应用程序如何在 Web 上实现互操作性。Web Service 接口和 Z39.50 接口相比,具有很多优势:

一是 Web Service 通过 SOAP 这种可在任何传输协议(诸如 TCP、HTTP、SMTP 等)上使用的轻量级协议,实现远程调用。因而采用常用的 HTTP 协议 80 端口,而避免了采用 Z39.50 协议不得不另开特殊端口(例如 210)的防火墙设置限制,这样,Web Service 接口就具有很大的便利性。

二是由于 XML 高度结构化和可读化,特别是与平台和厂商无关的优点,XML 成为不同 Web service

表示数据的基本格式。采用 Web Service 实现 XML 数据互操作,技术上更容易,开发成本更低。

三是 Web Service 可以通过 WSDL (网络服务描述语言) 这样一个基于 XML,既便于机器阅读又便于人理解的语言,描述 Web service 及其函数、参数和返回值。一些最新的开发工具既能根据 Web service 自动生成 WSDL 文档,又能导入 WSDL 文档,生成调用相应 Web service 的代码。

Web Service 接口如今已经成为一个通用的接口,如果图书馆界新一代的信息系统提供 Web Service 接口,除了比 Z39.50 协议更容易实现异构系统间的信息资源整合外,还由于当前 Google、Amazon 等网上信息巨头都已公开其数据库的 Web Service 接口,只需要双方授权,图书馆馆藏数据资源即可实现与这些海量数据库的整合和数据互操作。

3 集群外信息资源共享成本分析

对于集群外数字资源,比如 Google、Amazon 等数据库,以及其它数字信息系统,只要公开 Web Service,不需要对方干预(仅需要授权),即可独立地、低成本地将其信息整合进馆藏信息系统中。但是,对于集群外传统纸介质书刊资源或实物资源,想要整合进集群信息系统中的成本就比较高。

当前,小型图书馆(室)仍是采用传统方式管理书刊,无法形成数字化的书目信息甚至全文、声像等数字文档。想要整合这些集群外信息资源,首先必须实现这些资源的数字化。然而,经费不足、专业人员不足的小型图书馆(室)根本无法完成自有资源的数字化工作。

飞速发展的互联网,为图书馆信息系统突破本地局域网,将工作前端从本地扩展到全球互联网分布提供了快捷便利的条件。当前解决小型图书馆自动化甚至于信息化发展迟缓的最佳方案就是:让这些小型图书馆通过 ADSL 等宽带接入方式,成为有条件、有资源、有专业人员的图书馆互联网络工作前端,实现更大范围内的信息提交、数据加工和资源管理。事实上,由于 Web Service 这种更开放的系统接口保障,并不是只有没启动自动化管理的小图书馆(室)才需要作为网络客户端。即使已实现自动化管理的图书馆,也可互为网络客户端,共享共建信息资源,根据授权实现双方数据库的互操作。这种以成熟应用图书馆信息系统的图书馆为中心,在保障本馆局域网内工作前端的基础上吸收互联网上工作前端的图书馆集群建设方式,可以更方便共享资源、共享设备、共享人力,从而更容易实现集群外信息资源的整合。

图书馆集群建设,首要一点就是要将资源元数据与馆藏实体信息分离,即统一一个书目数据库(或其它元数据数据库),同一种书目的多个相互独立的馆藏信息(如图书册条形码、馆藏地点等)独立存贮并与书目挂接。只有这样,才能保障多人、多方共同建

设和完善某种资源的元数据描述,并且不影响馆藏实体的独立管理(如流通借还和实体信息的增删等)和联合目录发布。

4 实际应用介绍

地处人杰地灵的历史文化名城乐山的乐山市图书馆,除了需要利用MARC元数据实现馆藏书刊纸介质资源的管理外,还承担着地方文献、嘉州画派艺术、地方文化名人、旅游资源等特色数据库的建设,也承担着扶持和引导周边区县图书馆(室)从手工管理转型为计算机自动化管理并深化挖掘整理馆藏信息资源的职责。

为此,乐山市图书馆、乐山师范学院图书馆、西南交大峨眉分校图书馆为首的乐山市图书馆学会达成共识:以上述三个图书馆局域网为基础和中心馆,通过互联网接入方式实现互联互通,形成一个图书馆集群。其它图书馆(室)通过ADSL宽带接入方式,成为这个集群的远程工作前端,实现本地馆藏实物的流通以及资源信息的数字化整理。目的就是最大限度共享信息资源、人力资源和硬件资源,充分整合馆内外信息资源,形成规模更大的信息资源整合效应。

然而,当前常见的图书馆管理系统多是沿用前几年的理念和开发思路,局限于书刊资源的管理,仅凭MARC元数据描述体系并不能很好适应书刊类型外的其它资源的描述,无法满足乐山市图书馆界的需要。并且,三个图书馆现有的图书馆管理系统虽然相同,在同构系统条件下,很容易实现互联互通,但由于采用ISO2709等专业性太强的编码标准,无法保障以后与其它行业信息系统间的资源共享,不利于充分发掘出辛苦建设的数据价值,更担心因积累的专业局限性数据越多,以后可能会更加剧数据转换成本压力。

乐山市图书馆学会根据行业发展趋势,并未简单选择常见系统作为图书馆集群升级软件,而是与业内具有创新思维和开发实力的系统开发商合作,基于XML格式(含MARCXML)和Web Service标准,结合图书馆应用需求,量身定做了一个通用信息管理平台,在满足现有图书馆自动化管理功能并兼容传统图书馆自动化系统的基础上,采用DC元数据实现了地方特色资源等非书资源的描述和管理。也因为书目元数据与非书资源元数据都是采用XML标准(含MARCXML),所以这些资源信息已实现整合。

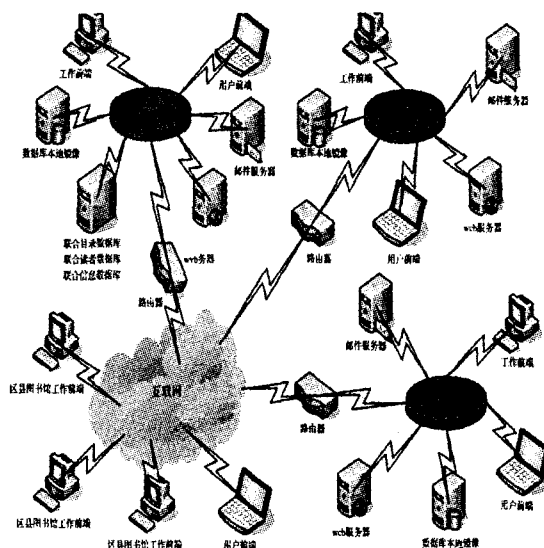
夹江县、五通桥区、沙湾区图书馆等尚未实现图书馆自动化管理的区县图书馆,采用ADSL宽带接入方式,已成为乐山市图书馆集群成员馆。这些成员馆通过ADSL宽带接入,普通工作人员即可利用专用前端软件,将本馆图书册条形码、馆藏地点等馆藏信息与乐山市图书馆书目数据库挂接,并利用乐山市图书馆读者库存贮本馆读者数据,从而实现了本地馆藏的借还流通操作。乐山市图书馆利用自己的WWW服务器,对外(包括成员馆)提供乐山市联合目录查询和

读者借阅信息查询,并利用本馆专业人员,定期为成员馆提供业务统计分析报表。

现在采用互联网工作前端实现图书馆自动化管理或数字化信息建设的成员馆,待条件成熟,也可添加本地服务器,形成独立的局域网系统体系,从而与图书馆集群中现有三个图书馆中心一道,共同承担为互联网工作前端服务或发展自己的互联网工作前端的工作。

乐山市图书馆集群,下一步将形成一个联合编目中心和读者数据中心,各成员馆共享共建这个书目中心,并挂接各成员馆馆藏信息,实现通借通还和地方资源联合馆藏。还将提供专业论坛、读书论坛等学术交流平台,通过系统与读者的交互,实现信息资源的自采集和自管理,为乐山市图书馆集群整合的信息资源提供重要的补充。

附乐山市图书馆集群布局图:



参考文献

- 1 MarcXchange. <http://www.bs.dk/MarcXchange/index.htm>.
- 2 MARCXML. <http://www.loc.gov/marc/marcxml.html>.
- 3 Web Services Activity. <http://www.w3.org/2002/ws/>.

(编辑:卢骥)

(校对:王涛)

(收稿日期:2009-08-27)