信息技术

文章编号:1003-7136(2008)02-0009-04

由XML和Web Service引发的信息资源整合革命

席毅强[1],范小松[1],江汇泉[2]

(1.乐山市图书馆,四川 乐山 610400;2.数字平台(北京)软件有限责任公司,北京 100081)

摘 要:当前图书馆馆藏书刊目录数据库多采用MARC元数据,并以ISO2709标准进行编码,而馆藏其它数据库多采用不同的元数据和元数据编码格式。并非是元数据不统一才导致图书馆难以整合不同数据库信息资源,根本原因在于这些元数据的编码格式不统一,接口不规范。文章从元数据编码角度指出,在不同信息系统中,只有采用新一代的XML和Web Service信息标准,才能实现跨资源类型、跨载体格式和跨系统的多种元数据整合,从而充分实现信息资源的共享共建和整合效应。

关键词:图书馆建设;资源整合;资源共享

中图分类号:G250.7 **文献标识码**:A

1 馆藏资源元数据分析

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介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。甚至很多新建立的图书馆,非书资源、数字资源所占的比重,已超出了传统纸介质书刊。仅靠MARC这种过于复杂、侧重于书目的元数据体系已不能完成图书馆现有馆藏资源的描述。因而MARC体系外的其它元数据体系,就成为图书馆界关注或应用的对象——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,以及更科学更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。

由于DC元数据为了更通用,更适应各个信息系统,它提出了"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,除了利用早已成熟的ISO2709标准,能方便实现内部与外部书目元数据的交换和信息整合外,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,想要实现元数据的交换和信息资源整合非常困难。

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、更有利于数据交换和格式转换的元数据编码标准进入了图书馆界视线。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标准。所以信息界(包括图书馆界)新兴的信息系统及其元数据体系,也多采用XML实现元数据编码。DCMI(Dublin Core Metadata Initiative,都柏林

收稿日期:2007-11-05

核心元数据提案)项目组织虽然提出了DC的"语法无关原则",但推荐XML为DC元数据的最佳编码格式。

图书馆的MARC数据库系统,由于是采用20世纪60年代制订的ISO2709编码标准,MARC数据想与其它信息系统数据进行交换或整合非常困难。为了继承图书馆界几十年来积累的海量书目信息资源,且与当前其它数字信息系统,比如搜索引擎数据库、网上书店系统、出版发行系统、电子图书等进行数据共享和互操作,图书馆界迫切需要一种比ISO2709更容易交换和识别的数据编码方案。在这种需求下,美国国会图书馆近年来确定了基于XML标准的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,只不过是以比ISO2709更具结构化、更可读的XML格式对MARC元数据进行编码。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。

新兴的DC元数据采用表义词汇标识元素,而MARC元数据采用数字和字母标识元素,这些不同的标识体系看起来大相径庭,但不同的标识体系对于计算机来讲实质都是一样的。事实上,由于采用XML编码标准,不同的元数据即使标识元素不一致,都可以通过XPath(XML Path Language,XML路径语言)这个统一通用的标准,忽略标识符的差别,实现数据寻址方式的统一,从而实现同时从多个元数据体系中索引抽取、内容显现的目的。MARCXML或ISO/DIS 25577编码方案的出现,就打破了ISO2709格式这种行业色彩过浓且复杂难处理的数据局限,为馆藏书目数据与其它类型资源元数据提供了一个整合的基础,为实现跨资源类型、跨载体格式的馆藏信息资源整合扫清了障碍。

2 馆藏信息系统接口分析

在单一的系统内部甚至相同系统之间都比较容易实现单一类型资源整合,采用XML编码标准,更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都比较困难,更何况想实现跨资源类型、跨载体格式的资源整合。

Z39.50(ISO23950)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协议的下载规定比较统一,所以服务器端利用Z39.50协议发布数据、前端下载服务器端数据的应用比较成熟。但Z39.50协议的数据上载功能是通过Z39.50 Extended Service中的Update实现的。顾名思义,Extended Service就是让大家自行扩充的部分。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它是一个过于完美的,复杂的重量级的协议。标准正式文本厚达156页,2002草案也有147页。标准的实施需要软件开发者了解数据结构,网络通讯,编码解码,数据库等诸多方面的知识。标准的复杂性使Z39.50的实施工作面临技术风险和较高的开发成本。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。

2.1 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。

2.2 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提供者表示数据的基本格式,具有越来越多的通用性。

2.3 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。采用Web Service实现XML数据互操作,技术上更容易,开发成本更低。

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 Web Service 是实现联合编目的最佳方案

当前，国内外常见的资源整合方式是通过联合编目系统，实现书目数据的共建和共享。这些联合编目系统，分同构系统联合编目与异构系统联合编目两种类型。同构系统因为规则一致，很容易实现书目数据的上传和下载，但异构系统间，由于开发商的不同，想要形成一致的规则很不容易。即使是 Z39.50 协议，也多为数据下载解决方案，而由于自行扩充的数据上载功能，也因差异较大导致兼容性较差，以至于不得不采用统一开发的上载模块作为系统补丁——这只是打着异构系统互操作旗号的同构系统，有违 Z39.50 初衷。

虽然 ZIG(Z39.50 Implementers Group，Z39.50 实施小组)提出了 ZING(Z39.50 International: Next Generation)这个新的协议集，欲利用 Web Service、XML、SOAP 等网络新技术简化 Z39.50 协议，突破已有的应用限制，帮助开发人员低门槛实现 Z39.50 协议。但是，ZING 仍只是对 Z39.50 协议的另一种实现模式或包装模式，它多作为 Web 用户与 Z39.50 服务器间的网关，仍无法回避网关之后的 Z39.50 复杂性。所以，虽然越来越多的系统与机构宣布支持 ZING，但 ZING 尚未成为 Web 上的主流检索协议。

ZING 仅是为了继承已有 Z39.50 资源，保护已有 Z39.50 系统投资的折衷方案。事实上，在 Web 用户与 Z39.50 系统之间增加一个独立的网关环节，看似回避了对现有系统增加 Web Service 接口的改造投入，维持了现有 Z39.50 资源的稳定性。但这个新增加的网关环节仍需不菲的开发投入，更何况，多出的这个环节是否会带来新问题更是未知的。

显然，Web 用户通过以 Web Service 表达的 SRW/SRU - Z39.50 网关，来访问数据库的 Z39.50 接口，远不如直接访问数据库提供的 Web Service 接口更简捷。所以，新一代的联合编目系统建设，采用 Web Service 这个支持面更广、开发成本更低的解决方案，必将大大优于采用 Z39.50 或 ZING 的解决方案。各个前端，通过服务器端数据库系统公开的 Web Service，即可轻易获得对数据的授权操作，根本不需要再为复杂的 Z39.50 体系而感到痛苦了。

特别是，由于 Web Service 与 XML 天然的联系，联合编目中心通过 Web Service 可以非常方便地实现以 XML(包括 MARC)编码的 MARC、DC 等多种类型资源的元数据联合编目，且实现这些元数据信息的整合管理，极大地扩展了当前联合编目的应用范围。

多年的 Z39.50 接口建设，在为网络数据查询和联合编目带来贡献的同时，也成为联合编目系统转型的阻力：

* 是否该为 Web 用户增加一个 SRW/SRU - Z39.50 网关中间环节，从而保证现有 Z39.50 资源不变？

* 是否该改造已有联合编目系统，新增一个 Web Service 接口，从而保证 Web 用户或前端以低成本方式直接实现数据授权访问？

* 是否该在新建联合编目系统时，即把 Web Service 接口作为最终解决方案，从而为可能有的其它接口需求扩展打下坚实的基础？

面对上述问题，我们该作何种选择呢？

4 实际应用介绍

地处人杰地灵的历史文化名城的乐山市图书馆界，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，还承担着大量的地方文献、嘉州画派书画艺术、地方文化名人、旅游资源、学位论文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和挖掘整理馆藏信息资源的职责。

为此，以乐山市图书馆、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，并形成一个图书馆集群。其中心馆下面的图书馆(室)通过 ADSL 宽带接入方式，成为这个集群的远程工作前端，以实现本地馆藏实物的流通以及资源信息的数字化整理。其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整合效应。

然而，当前市面上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。虽然三个图书馆现有的图书馆管理系统相同，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。

尚未实现图书馆自动化管理的区县图书馆，通过中心馆代建代管馆藏书目数据库方式和采用ADSL宽带接入方式，也将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)提供乐山市图书馆系统的联合目录查询和读者借阅信息查询，中心馆专业人员也将定期为成员馆提供业务统计分析报表。目前，已有3个区县馆和1个高校馆成为乐山市图书馆集群成员馆并运行正常。

目前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的三个图书馆中心一道，共同承担为互联网络工作前端服务或发展自己的互联网络工作前端的工作。

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，各成员馆共建共享这个书目中心，并挂接各成员馆馆藏信息，从而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，从而为乐山市图书馆集群整合的信息资源提供重要的补充。

参考文献：

[1]Expressing Qualified Dublin Core in RDF/XML. http://www.dublincore.org/documents/2002/05/15/dcq-rdf-xml/.

[2]SRU: Search and Retrieve via URL (Standards, Library of Congress). http://www.loc.gov/standards/sru/.

[3]SRW: Search/Retrieve Web Service. http://www.loc.gov/standards/sru/srw/.

[4]数字图书馆标准与规范建设：Z39.50协议应用指南.http://cdls.nstl.gov.cn/mt/blogs/2nd/archives/docs/Z39.50协议应用指南.pdf.

[5]MarcXchange. http://www.bs.dk/MarcXchange/index.htm.

[6]MARCXML. http://www.loc.gov/marc/marcxml.html.

[7]Web Services Activity. http://www.w3.org/2002/ws/.

The Revolution of Integrating Information Recourse Result from XML and Web Service

XI Yi-qiang[1], FAN Xiao-song[1], JIANG Hui-quan[2]

(1. Leshan Public Library, Leshan 610400, China;
2. Digital Platform Co Ltd, Beijing 100081, China)

Summary: Currently most of the databases about catalogue in library conform to MARC metadata with ISO2709. But for other resources in library often conform to different metadata and encoding. In library, it is hard to integrate those different databases, this is because of the different encoding of metadata and the unstandardized interface but not the different element of metadata. Based on metadata encoding, the article points out that in different information systems, adopting the standard of the new generation of XML and Web service, the inter-type, inter-format or inter-systems resources in library can be integrated. And the library can fully realize the effect of co-constructing, sharing and integrate the information resources.

Key words: library construct; resource integration; resource sharing

作者简介：

席毅强(1957-)，男，武汉大学图书馆学系，大专，经济师，乐山市图书馆馆长。

范小松(1971-)，男，西南师范学院图书情报专业，学士，馆员，乐山市图书馆技术部主任。

江汇泉(1971-)，男，北京大学图书情报专业，学士，馆员，数字平台(北京)软件有限责任公司经理，中国社科院联合编目项目课题组成员，主要研究元数据编码、特色资源数据库建设等。