图书馆关联数据:机会与挑战*

林海青　楼向英　夏翠娟

摘　要　关联数据在图书馆领域具有广泛的应用前景,通过采用关联数据技术,图书馆有机会在未来语义网建设中发挥主导性作用。图书馆应用关联数据具有潜在的四个基本模式:发布、消费、服务和平台。关联数据的应用会使机器成为图书馆的重要服务对象,图书馆不仅要为人服务,而且也要为机器服务,这将带来一系列的机遇和挑战。表2。图3。参考文献14。

关键词　关联数据　图书馆　机器用户

分类号　TP311.13　G250.7

Linked Data: Opportunities and Challenges for Library Services

Lin Haiqing, Lou Xiangying & Xia Cuijuan

ABSTRACT　Linked data principles can be widely applied to library services. Libraries have an opportunity to play a leading role in the development of semantic web by building linked data applications. There are four potential service models for library linked data applications: linked library data publication, linked data consumption, library data services provision and application platform development. The essay argues that library linked data practices will enable machines a significant part of the library services users. Libraries will not only provide services to people but also machines, which will possibly become a new remarkable challenge for libraries. 2 tabs. 3 figs. 14 refs.

KEY WORDS　Linked Data.　Libraries.　Machine Users.

关联数据是一种旨在提高网络数据机器可读性的技术框架,它通过构建网络环境下数据的引用和解引(reference/dereference)机制来建立数据之间的关联,从而实现数据在Web平台上的分享与重用。关联数据的核心是将数据和网络融合起来,一旦数据用关联数据的原理发布,数据就成为网络的一部分,实现网络即数据这个伟大的理想。

图书馆一方面作为数据的发布者,另一方面又作为数据的消费者,显然不能游离于关联数据运动之外,那么关联数据会给图书馆带来什么?图书馆如何应用关联数据?图书馆应用关联数据有哪些挑战?本文试图探讨这些基本问题。

1　图书馆关联数据应用现状

从宏观上说,关联数据给图书馆带来了机会,让图书馆有可能遵循一种泛在的技术规范提供服务,从而真正地将自己融入到整个信息世界中去。近年来图书馆关联数据应用有了长足发展,但总体上说目前还处于起步阶段。2010年9月在英国召开了国际知识组织协会大会,这次大会的主题是"关联数据:Web知识组织的未来"(Linked Data:The Future of Knowledge Organisation on the Web)。一位作者在报道这次大会

*　本文系国家社科基金项目"关联数据的理论和应用研究"(编号:11BTQ041)的研究成果之一。
通讯作者:夏翠娟,Email:cjxia@libnet.sh.cn

时,用的标题是“关联数据还处在早期阶段,但文化变化得很快”[1]。这个标题非常确切地概括了图书馆关联数据发展现状。

自 2006 年 Tim Berners-Lee 提出关联数据概念不久,图书馆界很快就对关联数据的应用作了有益探索。2008 年美国国会图书馆的 Ed. Summers 建立了 lcsh. Info 网站,将国会图书馆主题词表(LCSH)以关联数据的形式通过这个网站发布;同年瑞典国家图书馆也将瑞典全国联合目录 LIBRIS 采用了关联数据框架,成为首家关联编目数据提供者。更为重要的是,这两个项目不是孤立进行,而是相互连接起来的。LIBRIS 的瑞典语主题词通过 Summers 的 lcsh. info 提供的 URI,和美国国会图书馆主题词关联起来,完成了一个完整的关联图书馆数据的开拓性实验。随后有大量的图书馆关联数据项目涌现出来,从开放知识基金会网站登记的关联数据项目情况来看,图书馆关联数据项目已达到 51 个(截止到 2011 年 9 月)[2],这些关联数据项目总共提供了 4,576,472,613 个 RDF 三元组,平均每个项目包含 89,734,757. 12 个三元组。这 51 个项目中有 33 家提供了 SPARQL Endpoint 服务,占 64. 7%。

这些项目包括了图书馆书目数据、图书馆规范主题词数据、规范人名数据等,还包括 MARC、DDC 等图书馆标准、工具的关联数据形式。其中有 20 个书目数据,约占 39% 左右;13 个是规范控制数据,其中 6 个数据集是主题词服务,其余是人名控制和其他规范控制服务。还有 6 个关联数据集是专门的术语服务。由此可见,目前图书馆关联数据主要集中在书目数据、规范数据和术语服务三个主要领域。

关联数据的核心之一就是和其他数据集实现数据共享和相互关联。在图书馆关联数据的 51 个实例中,共关联了 56 个外部数据集,在数据集之间构建了 116 个关联关系,平均每一个图书馆关联数据项目和 2 个以上的外部数据集实现数据共享和重用。这 116 个关联关系中,总共包含了大约 381,238,848 个数据层面的连接。平均每个关联数据项目中有 7,475,272 条数据和外部数据相关联。其中 DBpedia,LCSH 等关联数据集是图书馆关联数据主要的关联对象,DBpedia 被关联了 12 次,独占鳌头。表 1 展示了 51 个图书馆关联数据项目和外部数据相关联的基本数据。

表 1　图书馆关联数据外联情况

关联对象数据集	相关联的图书馆数据集	关联三元组数
Dbpedia	21	3, 476, 139
Lcsh	12	4, 781, 964
geonames-semantic-web	7	1, 013, 067
Viaf	6	2, 592, 161
dnb-gemeinsame-normdatei	5	10, 523, 740
stitch-rameau	5	194, 100
dewey_decimal_classification	4	2, 700, 782
marc-codes	3	30, 370, 972
Openlibrary	3	13, 043, 082
Lexvo	2	9, 600, 000
Lingvoj	2	1, 421
rkb-explorer	2	1, 616, 565
semantic-web-dog-food	2	371
stw-thesaurus-for-economics	2	6, 063

注:据 the Data Hub (http://ckan. net/) 提供的信息统计。

分析图书馆关联数据外联资源分布情况，发现外联资源主要是规范数据服务，尤其是规范主题词服务，特别是国会图书馆主题词表以关联数据形式开放后，成为图书馆关联数据的一个链接中心。OCLC 的 VIAF 也是图书馆关联数据外联的重要数据源。

和以往不同的是，图书馆关联数据的发展得到了外界的重视和推动。由于图书馆保存了大量的书目数据，这些数据构建了一幅完整的人类知识地图。当人们试图将 Web 构建成一个大规模的数据空间时，没有书目数据将是不完整的。2010 年 5 月 W3C 成立了一个图书馆关联数据孵化小组，专门探讨图书馆如何应用关联数据技术来增进现有的各种图书馆技术，如元数据、元数据标准和其他技术协议在网络环境下的互操作性，鼓励和促进图书馆将他们的数据在网络环境下实现互操作并向其他领域开放。

2 关联数据，图书馆的机遇

W3C 的图书馆关联数据孵化小组于 2011 年 8 月活动结束时，发表了一份研究报告。这份报告建议，图书馆应该从两个方面来拥抱信息网络：利用关联数据将图书馆的数据变成可利用的；将数据网络融入到图书馆服务中去。报告认为，图书馆数据应该完全整合到其他网络资源中去，为信息搜寻者创建显著的图书馆数据可见度，并把图书馆服务直接带给他们。报告提出了一个让图书馆人振奋的观点，那就是图书馆能够在数据网络运动中担当领导者的角色[3]。类似的观点还见诸 Eric Miller 的一次演讲，Eric 提出，图书馆面临独一无二的机遇，那就是他们不仅能为关联数据世界贡献数据，而且能够引领多种关联数据建设的实践[4]。

关联数据不仅为图书馆奠定了一个新的活动舞台，同时也需要图书馆在这个舞台上扮演主要角色。如果没有图书馆的参与，关联数据的发展是不完备的，甚至是不可能实现的，这也是为什么 W3C 报告会认为图书馆将在关联数据运动中担当领导者的角色。

关联数据不是一个封闭孤立的数据空间，而是一个社会性的数据环境，Heath 提出了一个关联数据的基本架构[5]（见图 1）。

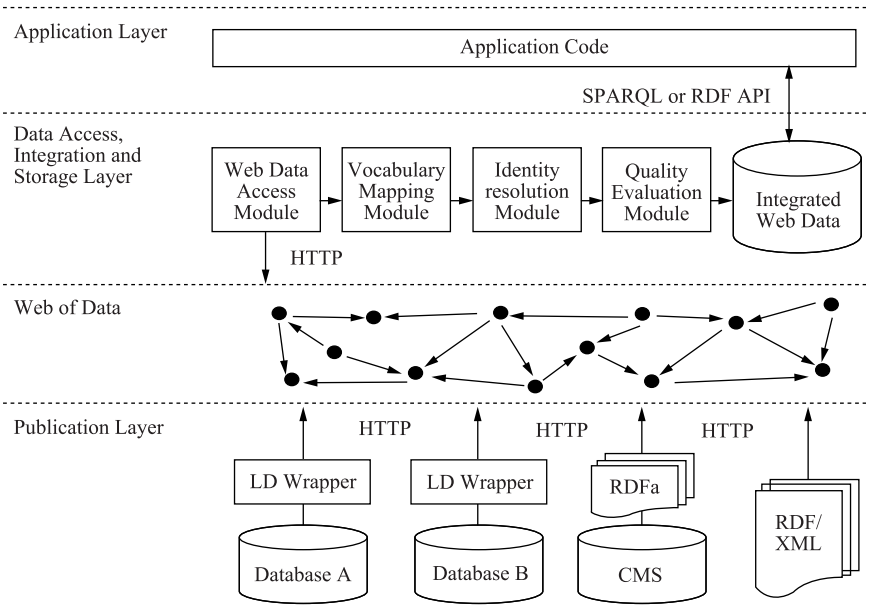

图 1 关联数据框架[5]

从图1看出,关联数据应用有三个功能层:数据发布层、数据存取整合和保存层、数据应用层。数据发布层主要由数据发布者构成,它们是关联数据网的数据提供者,数据应用层由关联数据消费者构成,它们主要是应用关联数据来满足自身的数据需求。中间一层,即数据存取、整合和保存层是由关联数据的第三方参与者构成,它提供了一系列基础服务,如本体词汇的维护、不同本体词汇之间的相互映射、数据标识的规范控制等。这一层其实是关联数据网的基础设施,它构建了关联数据发布者和消费者之间的桥梁。

图书馆显然可以存在于这三个功能层中,它可以作为数据的发布者,而成为发布层的主要组成部分;它又可以成为关联数据的消费者;更重要的是,图书馆以其得天独厚的优势可以担当第三方的角色,成为数据存取整合和保存层的主力军。尤其在构建关联数据网的信任机制方面,图书馆可以作出重要贡献。

语义网层次结构的顶层是信任层,它确保语义网数据是可靠的。关联数据作为语义网的具体实现,是一个开放的数据环境,如何确保这个开放环境的有序性,建立有效的机制来确保数据的可信度,辨别"好"的关联数据和"坏"的关联数据,成为数据网络发展的关键。图书馆数据是高质量的,由训练有素的专业人员搜集、修订和维护,正因为如此,图书馆数据有可能成为关联数据信任机制中迫切需要的支柱[6]。

关联数据信任机制的重要组成部分是确定数据的来源(data provenance),追溯数据的来源是确定数据可靠性的主要手段,数据的内容由谁提供? 这些数据通过哪些过程处理过? 谁保存这些数据? 这一切都是判断数据可靠性的依据。Hartig [7]2009年曾经提出了一个关联数据来源的概念模型,这个模型定义了数据来源的基本组件:行为者(actors)、实施(executions)和制成品(artifacts)。一个数据来源信息可以表述为一个判断:行为者实施一个过程形成或施用一个制成品。行为者主要是数据创建者,包括数据创建实体,数据创建服务和数据创建设备。行为者还包括数据发布者、数据服务提供者等。如何表述和保存这些信息,是数据来源机制的核心,同时也是挑战。

图书馆拥有独特的信息资源,尤其是维护了海量的人名和机构、规范、数据,这些数据经过训练有素的图书馆员搜集规范查验整理,记录了大量人名与机构名称变迁的历史,成为具有很高可靠性的数据源,完全可以成为追踪数据来源的基础资源。同时,机器虽然可以完成一部分数据来源的追踪工作,但是很多情况下还需要人工干预。尤其是数据来源表述的一个重要手段是注释方法(Annotation method)[8],这种方法和图书馆界的信息组织方法非常相似。图书馆界拥有一大批经过专门训练,信息整理组织经验丰富的专业人员,他们完全可以承担关联数据的整理组织和来源追踪确认的工作。在长期的全球性书目控制工作中,图书馆界还形成了一个全球合作分享数据的机制与模式,这种机制同样可以运用到数据网络的规范组织中去。

3　关联数据在图书馆领域的应用模式

自2008年以来已有很多图书馆关联数据应用项目出现,这些应用项目覆盖了图书馆服务领域的许多方面,书目数据和规范数据是图书馆关联数据应用的主要对象。纵观现有的关联数据应用,我们可以将图书馆关联数据应用归纳为四种模式:发布、消费、服务和平台。

3.1　发布

将图书馆数据以关联数据的形式发布出来,以便其他网络可以利用这些数据,这是图书馆关联数据应用的主要模式。图书馆书目数据关联数据化已经成为一个热点,根据开放知识基金会网站图书馆关联数据组的数据统计,共有九个国家级图书馆发布了18个关联数据集,这些数据包括书目数据、主体规范数据和名称规范数据(见表2)。

表 2　图书馆关联数据项目统计

	书目数据	主题规范数据	名称规范数据	其他
联合国粮农组织		AGROVOC		
OCLC		Dewey Decimal Classification (DDC)		
OCLC		VIAF: The Virtual International Authority File		
匈牙利国家图书馆	Hungarian National Library (NSZL) catalog			
德国国家图书馆	20th Century Press Archives	GemeinsameNormdatei (GND) STW Thesaurus for Economics		
捷克国家技术图书馆		Polythematic Structured Subject Heading System		
日本国会图书馆		Web NDL Authorities-National Diet Library of Japan National Diet Library of Japan subject headings	Web NDL Authorities-National Diet Library of Japan	
法国国家图书馆	data. bnf. fr-Bibliothèque nationale de France			
瑞典国家图书馆	LIBRIS			
美国国会图书馆		Library of Congress Subject Headings LCSubjects. org Library of Congress Subject Headings	Library of Congress Name Authority File (NAF)	Chronicling America
芬兰国家图书馆		Yleinensuomalainenasiasanasto		
英国不列颠图书馆	British National Bibliography			
德国北莱茵－威斯特法伦图书馆服务中心	lobid. Bibliographic Resources			lobid. Index of libraries and related organisations
剑桥大学	Cambridge University Library dataset			
挪威科学技术大学		TEKORD		
法国高校与研究机构图书馆联合目录	Sudoc bibliographic data	theses. fr		

注:据 the Data Hub (http://ckan. net/) 提供的信息统计。

书目数据和规范数据是图书馆界原生的数据。首先,关联书目/规范数据实现了真正意义上的数据开放,图书馆数据因此成为一种通过网络向其他应用提供的数据服务。其次,关联数据技术也可以将书目数据和其他数据融合起来,使书目信息更加丰富和完整,在书目数据多元化的环境下,图书馆面临的挑战不在于是否需要发布关联书目数据,也不在于如何发布关联书目数据,核心问题在于如何界定关联数据的主要功能需求。用户是否能够从图书馆发布的关联书目数据和规范数据中获得更多、更可靠、更准确的信息,并在此基础上提供增值服务。第三,从图书馆界内部业务过程看,关联数据的应用确保了数据的重用和分享,使得图书馆数据流程更加清晰。

图书馆的原创性数据不只是书目数据和规范数据,图书馆,特别是学术图书馆还能提供更多的数据增值服务,比如资源导航数据、研究信息等。这些数据增值服务也可以通过关联数据的形式开放出来,其他系统通过消费图书馆提供的增值性数据,在系统中嵌入图书馆的服务。

3.2　消费

从消费关联数据的角度看,关联数据具有很强的数据整合和重用功能。图书馆系统可以通过消费关联数据的方式来整合外部数据,通过关联数据将各种数据源无缝地关联起来,将图书馆资源建成一个广域分布的数据库。尤其重要的是,关联数据不仅是裸数据,也描述了数据之间的相关关系,关联数据对关系形式化描述,形成一张关系地图,使得机器可以通过理解和处理数据之间的各种关系,发现新的数据。通过关联数据图书馆系统可以按图索骥,集成更多的信息与功能。

我们可以通过一个具体实例,来展示图书馆系统如何通过应用关联数据来整合不同来源的数据与功能,以丰富图书馆数据服务的内容,向读者提供完整的信息资源。

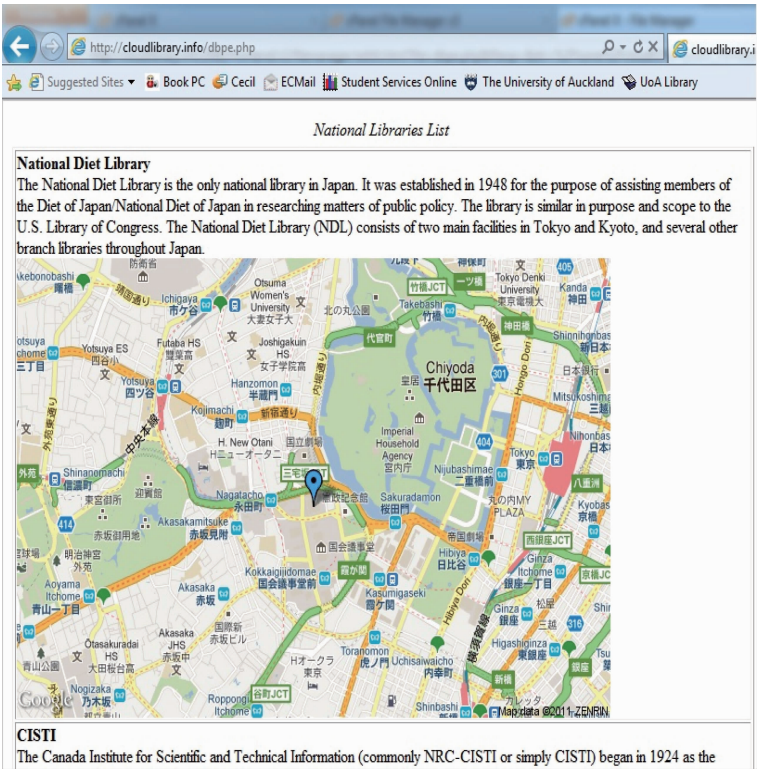

图 2　关联数据消费样例

图2是关联数据消费的一个试验网站的网页截图,这个网站列出了将近50个国家图书馆的名称和简介,同时将图书馆的地理位置标注在地图上。所有国家图书馆的简介来自DBpedia,Wikipedia的关联数据版本。通过DBpedia提供的图书馆位置的经纬度,再调用Google的地图API,将图书馆的位置标注在地图上。为实现这个功能,系统先运行一个SPARQL查询:

```
$query = "
    PREFIX dbp: <http://dbpedia.org/resource/>
    SELECT ? library ? long ? lat ? ab ? b
    WHERE {
        ? library dcterms:subject dbp:Category:National_libraries>.
        ? library geo:long ? long.
        ? library geo:lat ? lat.
        ? library dbpedia-owl:abstract ? ab FILTER (LANG(? ab) ='en').
        ? library rdfs:label ? b FILTER (LANG(? b) ='en')
    }";
$searchUrl = 'http://dbpedia.org/sparql? query='.urlencode( $query).'&format=json';
$ch = curl_init();
curl_setopt( $ch,CURLOPT_URL, $searchUrl);
curl_setopt( $ch,CURLOPT_RETURNTRANSFER,true);
$response = curl_exec( $ch);
```

通过这段程序,从DBpedia获取四组信息:各国家图书馆URI、名称、经度、纬度和简介。当获取经度纬度后,用以下Google MAP API将图书馆位置标注在地图上:src = "http://maps.google.com/maps/api/staticmap? zoom=14&size=800x412&maptype=roadmap&markers=color:blue|". $lib[lat][value]."," . $lib[long][value]."&sensor=false。其中,$lib[lat][value]和$lib[long][value]均来自上述SPARQL语句的运行结果。

从这个实例我们看出:

①关联数据的消费是通过机器完成的,换句话说,关联数据是为程序设计准备的,它提出了一系列共同遵守的规范,确保系统开发者能够准确理解数据的含义,并使用这些数据。这些含义就是语义网所说的语义。在本例中,语义体现在dcterms:subject,geo:long,geo:lat,dbpedia-owl:abstract和rdfs:label。这些语义描述保证了应用系统能够准确地使用从DBpedia中获取的数据。

②SPARQL,包括SPARQL语句和SPARQL Endpoint,是消费关联数据的重要组件。SPARQL语句具有非常强大的表达能力,几乎可以表达所有的对RDF数据的查询需求。消费关联数据的核心是构建SPARQL语句,并通过SPARQL EndPoint运行这个语句,从而获得所需要的数据。

3.3 服务

虽然关联数据本身是一种数据服务,但是图书馆数据仅仅以关联数据的形式发布出来是不够的,用户来消费这些关联数据不仅仅是要连接几个数据点,而是需要更多的功能和服务。所以图书馆关联数据需要提供各种增值服务,将数据和功能绑定在一起,这样才能充分发挥图书馆的数据优势。比如图书馆在提供关联规范数据的同时,是否能够提供抽词服务、规范控制词映射服务等。

目前,图书馆基于关联数据的服务还是一个空白,不仅没有开展这样的先导性服务,即便研究课题也鲜有涉及。然而,放眼图书馆界以外的关联数据世界,还是可以看到很多有宝贵

借鉴价值的应用,汤森路透公司的 OpenCalais 就是一个很好的服务模式[9]。OpenCalais 是一个集合自然语言处理、机器学习等技术来实现自动生成基于文本内容的语义元数据的网络服务,它可以分析文本的内容,在文本中发现各种实体,如人物、机构、事件等,并将这些实体提取出来并以关联数据形式发布,便于搜索引擎发现和索引(见图 3)。

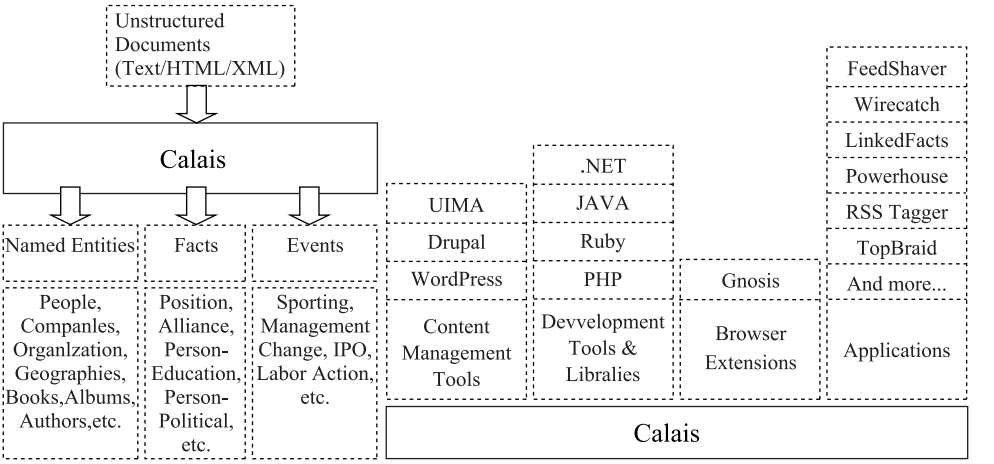

图 3　Calais 功能结构[9]

为了实现这个功能,OpenCalais 提供了四种服务模式:内容管理系统工具、开发工具和库、浏览器扩展和应用系统。这四种模式基本上体现了基于功能的网络服务的主要模式。尤其是内容管理工具将 OpenCalais 的功能嵌入内容管理系统中去。

图书馆界关联数据服务可以借鉴 OpenCalais 的服务模式,不仅开放数据,同时开放功能,这样才能将图书馆数据真正嵌入到社会信息基础结构中去,使图书馆数据不仅存在于网络,而且成为网络的一部分。

3.4　平台

平台其实是一种开放环境,用户可以利用平台提供的基础资源和功能来实现自己的应用。提供应用平台,是图书馆尤其是大型图书馆应用关联数据的有效途径。关联数据应用平台化的一个有效尝试是美国国会图书馆的 Recollection 项目。

Recollection 是国会图书馆全国数字化信息基础设施和长期保存项目(National Digital Information Infrastructure and Preservation Program)的一部分。2009 年,国会图书馆和 Zepheira 公司合作,开发一个用于收集和发现数字化资源的平台环境,Recollection 就是这个项目的成果。Recollection 允许学者、图书馆或其他机构上传各种数字化资源,生成各种显示界面,包括交互式地图、时间列表等,并且可以让这些数字资源嵌入各种应用中。Recollection 采用了关联数据框架,其关联数据特征体现在三个方面[10]:

① 用 URI 来揭示资源,这是关联数据四原则中的第一个原则,即采用 URI 作为资源的唯一标识符,这样任何网络资源都有一个唯一的名称。各种资源从幕后走向前台,成为网络结构的一部分,易于被发现和指证。

② 利用 HTTP 协议来存取资源,这满足了关联数据的第二个原则,HTTP 协议是万维网的基础协议,被广泛支持。几乎所有系统和开发工具都支持 HTTP 协议,从而使得资源的存取方法具有广泛的适用性和便利性。

③ 利用通用的数据格式最大限度地促进数据的重用和分享。Recollection 可以提供 RDF/

XML,HTML,Semantic wikitext,JSON 等多种机器可读的数据格式,这些数据携带各种语义信息,以便各种应用系统理解数据的含义,准确地使用这些数据。这是关联数据第三原则所要求的。

虽然根据 Tim Berners-Lee[11] 关联数据应用的五星级标准,Recollection 目前只是四星级的关联数据应用,但是这个尝试却是图书馆提供关联数据应用平台的先导,值得关注和研究。

4 关联数据与图书馆服务的转型

图书馆显然能够从关联数据应用中获益,有学者总结了图书馆会在以下方面从关联数据应用中获得好处[12]:①关联数据能够让图书馆以通用的格式(RDF)来发布各种事实性调查数据,这些数据能够容易被其他系统汇聚和利用,从而使图书馆能够有效地支持"基于证据的决策"(evidence-based decision-making);②通过采用关联数据技术,图书馆能够成为一个关联枢纽,这个枢纽可以连接各种图书馆相关者,将其整合在一起,形成一个真正的集成图书馆系统。③关联数据有助于图书馆实现"智能联合检索"(smart federated search)。如果各种数据都以标准的数据格式发布出来,那么很多智能联合检索的问题都可以迎刃而解。④关联数据还有助于实现基于语义的搜索引擎。

图书馆利用关联数据最为重要的价值不在于具体的技术改善,而是关联数据将从根基上改变图书馆在整个社会信息基础结构中的地位。关联数据本质上是一种 Web 数据服务,是面向机器的。关联数据在图书馆界的应用必然带来图书馆用户概念的变化,图书馆不仅要为活生生的人服务,同时也要为机器服务。当机器成为图书馆的主要服务对象后,图书馆的性质就会发生质的转变,图书馆可能会从前台服务转为后台服务,成为整个社会信息系统的一个基础设施。图书馆员的角色也会发生变化,他们通过控制资源的源头来确保数据整合的可靠性,其他事情就可以交给机器依据关联数据原则去整合,这样既体现了图书馆员的智力贡献,又借助了机器的高效率,使得图书馆能够应付瞬息万变的信息世界。

国会图书馆的 Daniel Chudnov[13] 在一篇文章中描绘了图书馆为机器服务的一个例子,他在讨论关联数据对图书馆 OPAC 会产生什么影响时,提出在移动终端日益普及的今天,图书馆可以将现有的指代性元数据(Surrogate metadata)和对象结合起来,使图书馆的 OPAC 不只是图书馆馆藏的描述,而是通过关联数据提供更多的信息,这样用户可以通过移动终端访问 OPAC 来获取和具体馆藏相关联的整个知识对象。

关联数据的应用还将加强图书馆目录服务的功能,文献[14]报道了挪威奥斯陆公共图书馆采用关联数据提升图书馆编目系统的实践,这个称作 Pode 的项目旨在通过各种 Mashup 技术,如 Z39.50、SRU 和 FRBR、关联数据技术等为用户提供更强的检索功能来获取各种公共数据。这个项目连接了 OCLC 提供的杜威十进制分类系统关联数据服务,同时由于采用了 RDF 格式,提供了更强大的智能化的查询功能,因为 RDF 提供了对关系的描述,这就使用户可以基于关系查询。

5 关联数据应用的技术性挑战

关联数据在图书馆应用的挑战首先在于总体架构的设计方面。关联数据不是一种具体的技术,而是一种模式,一种数据组织和共享的框架结构。其中数据的发现和检索机制是决定关联数据应用成功与否的关键。关联数据的检索与提取可以有多种模式,问题的关键在于如何构造动态的关联数据查询和提取模式。目前比较定型的关联数据检索机制应该是 SPARQL,SPARQL 可以和关系数据库的 SQL 相提并论,它提供了一套完备的检索描述机制。从目前现有的图书馆关联数据的应用看,越来越多的图书馆在发布关联数据的同时也提供 SPARQL Endpoint 服务,但总体上,发现机制和检索机制方面的探索还非常薄弱。

其次,关联数据是一种语义数据,从道理上

说,数据的语义都被很好地描述出来。由于语义描述系统的多样性,如何将不同语义系统的数据无缝地整合在一起,换句话说,如何实现不同语义描述系统,即本体之间的互操作,包括映射和匹配,也是一个根本性的问题。消费关联数据的关键之处在于整合和转换不同的语义系统,将其合并到本地语义系统中去,确保本地语义系统的一致性。另一方面,由于语义描述体系的多样性,在本体整合过程中不可避免地会带来本体冗余,即同样的信息被不同的本体系统描述着,给本地系统消费关联数据带来不便,甚至有时每个本体描述系统的一致性和完整性都没有很好地得到保障,造成很多麻烦。在消费 DBpedia 数据时就遇到这样的问题。

第三,关联数据和其他 Web 服务的整合问题,由于关联数据很少,尤其是可靠的关联数据更少,在整合数据时不可避免地要整合用其他方式发布的数据,这些不同来源的数据处理方式是不一样的,同样一种类型的数据会有不同的处理方法,造成系统设计的复杂性。

第四,在本地系统中消费关联数据,需要考虑本地系统的功能实现,消费关联数据有两种基本模式,一种是本地化消费,即将关联数据收割到本地系统。本地化消费比较可靠,由于数据已经在本地系统,处理起来更加方便,安全性和可靠性都可以得到很好的保障。本地化消费关联数据的重点是如何设计同步机制并设定同步时差,当外部数据源更新后,本地数据也可以实时更新。另一种是动态消费。动态消费很好地解决了数据的同步问题,但随之而来的是本地处理,由于数据不存放在本地系统,导致本地数据的分布式状态,给建立本地索引机制、查找机制带来一定的麻烦。

最后需要指出的是,关联数据的应用是社会性的,一个基于关联数据的应用会广泛集成外部数据和功能,所以图书馆服务会越来越融合并依赖于外部社会信息环境。关联数据的消费不是普通的链接,而是一种深度的数据嵌入,甚至是功能嵌入,如果外部数据链断裂,将直接影响本地系统的稳定和功能实现。从这个角度看,消费关联数据的一个代价就是将自己系统的稳定性、可靠性和数据的完整性都建筑在不可掌控的外部系统,这就给图书馆服务的可靠性带来挑战。我们曾经试验过一个基于关联数据应用的资源导航系统,该系统在运行过程中有几次数据丢失,其原因是 DBpedia 连接不上,这样就直接影响了本地系统的运行。当然这个问题不是关联数据固有的,但关联数据对外部数据有很强的依赖性,连接的可靠性问题就显得尤其突出。再者,第三方服务的稳定性也会直接影响关联数据应用的可靠性。很多关联数据需要集成第三方的服务,如一些系统的 SPARQL Endpoint 服务是由第三方提供的,那么第三方服务的稳定性也将决定本地系统的可靠性。在上述资源导航试验系统中,由于 DBpedia 的 SPARQL endpoint 提供方连接不上,数次导致系统瘫痪。

参考文献:

[1] Hyams E. Early days in linked data but culture changing fast[J]. Library & Information Update, 2010,9(10):23-25.

[2] Open Knowledge Foundation. Library linked data: The data hub—The easy way to get,use and share data[EB/OL]. [2011-09-08]. http://ckan.net/group/lld.

[3] Thomas Baker E B. LLD XG final report (Draft of the general part)[EB/OL]. [2011-09-09]. http://www.w3.org/2005/Incub-ator/lld/wiki/DraftReportWithTransclusion.

[4] Miller E. Linked data and libraries[J]. The Serials Librarian,2011(60):17-22.

[5] Heath T A. Linked Data:Evolving the Web into a global data space[R]. Synthesis Lectures on the Semantic Web:Theory and Technology,2011.

[6] Hannemann J,Kett J. Linked Data for Libraries[EB/OL]. [2011-09-13]. http://www.ifla.org/files/hq/papers/ifla76/149-hannemann-en.pdf.

[7] Hartig O. Provenance information in the web of data[R/OL]. Proceedings of the Linked Data on the Web Workshop,LDOW'09. Madrid,Spain:2010. [2011-09-28]. http://events.linkeddata.org/ldow2009/papers/ldow2009_paper18.pdf.

(下转第112页)

literature management to information resources management[M]//Ji Baocheng, Liu Dachun. A Report of Humanities and Social Science Development. Beijing: Press of Remin University of China, 2009: 360 – 370.)

[4] 中国共产党第十七届中央委员会. 中共中央关于制定国民经济和社会发展第十二个五年规划的建议[OL]. 北京: 中国共产党第十七届中央委员会第五次全体会议, 2010. [2010 – 10 – 27]. http://news.qq.com/a/20101027/001797.htm. (The Seventeenth Central Committee of the Communist Party of China. CPC Central Committee suggestion on twelfth five years plan of national economic and social development[OL]. Beijing: Chinese Communist Party's Seventeenth Central Committee Fifth Plenary Session, 2010. [2010 – 10 – 27]. http://news.qq.com/a/20101027/001797.htm.)

[5] 中共中央, 国务院. 国家中长期教育改革和发展规划纲要(2010 – 2020年)[OL]. [2010 – 07 – 29]. http://www.gov.cn/jrzg/2010 – 07/29/content_1667143.htm. (The Central Committee of the Communist Party, the State Council. Outline of national medium and long-term educational reform and development plan (2010 – 2020)[OL]. [2010 – 07 – 29]. http://www.gov.cn/jrzg/2010 – 07/29/content_1667143.htm.)

[6] 叶继元. 图书情报学(LIS)核心内容及其人才培养[J]. 中国图书馆学报, 2010(6): 13 – 19. (Ye Jiyuan. Core contents of the library and information science and professional training[J]. Journal of Library Science in China, 2010(6): 13 – 19.)

[7] 刘延淮. 企业技术创新与知识产权信息服务平台建设[C]. 北京: 2010中国企业知识产权大会, 2010. (Liu Yanhuai. Technical innovation of enterprises and intellectual property information service platform construction[C]. Beijing: 2010 Chinese Enterprise Intellectual Property Conference, 2010.)

叶继元 南京大学信息管理系教授, 博士生导师。通讯地址: 南京市汉口路22号南京大学。邮编: 210093。

(收稿日期: 2011-09-02; 修回日期: 2011-10-14)

(上接第67页)

[8] Omitola T G. Provenance in Linked Data Integration[R]. Future Internet Assembly. Ghent, Belgium, 2010.

[9] Thomson Reuters. How does calais work?[EB/OL]. [2011 – 09 – 28]. http://www.opencalais.com/about.

[10] Johnston L. MacDougall, K. Use case recollection [EB/OL]. [2011 – 09 – 07]. http://www.w3.org/2005/Incubator/lld/wiki/Use_Case_Recollection.

[11] Berners-Lee T. Linked data: Design issues[EB/OL]. [2011 – 09 – 18]. http://www.w3.org/DesignIssues/LinkedData.html

[12] Byrne G, Goddard L. The strongest link: libraries and linked dafta[J/OL]. D-Lib Magazine, 2010, 16 (11/12).

[13] Chudnov D. Connecting linked data, OPACs, and inline exhibits[J]. Libraries in Computers, 2009(5).

[14] Westrum, Anne-Lena. The key to the future of the library catalog is openness[J]. Computers in Libraries, 2011(4).

林海青 奥克兰大学图书馆亚语部亚洲学学科馆员和中文资源馆员。通讯地址: Asian Languages Collection, The University of Auckland Library, Private Bag 92019, Auckland Mail Centre, Auckland 1142, NEW ZEALAND.

楼向英 浙江理工大学图书馆馆员。通讯地址: 浙江杭州下沙高教园区浙江理工大学图书馆713室。邮编: 310018。

夏翠娟 上海图书馆系统网络中心研究开发部工程师。通讯地址: 上海市淮海中路1555号。邮编: 200031。

(收稿日期: 2011-10-19)