Journal of Academic Libraries

关联数据：概念、技术及应用展望

□刘 炜

摘要 概述了关联数据概念的提出、基本内涵、技术实现和当前国内外的研究应用状况，对其在图书馆行业的应用作了简要介绍，点评了国内该领域的研究开发情况，重点阐述了对于图书馆在Web上发布书目数据和规范数据的重要意义，认为关联数据与网络时代的图书情报工作关系密切，是互联网发展到语义网时代，对网上资源和数字对象进行“编目”和“规范控制”的基础性技术，是数字图书馆进行信息资源发布和服务的核心技术之一。最后作者呼吁我国图书情报界重视这一技术，及早投入一定的资源和人力进行研究开发和应用推广，使图书馆大量的权威数据在互联网上占据一席之地。

关键词 关联数据 Linked Data 规范控制 语义网

引言：一个有序的知识世界

哲学家波普尔的心中存在一个超然世外、遗世独立的知识世界，负载却不依赖于具体的物质世界，依靠却不附属于个体的精神世界。这个世界总体上依赖于信息网络和各类载体而存在，具体上却不依附于任何个体的硬件设施；理解或解读这个世界需要人类大脑的参与，但它却有其自身的发展规律。遗憾的是在波普尔1994年去世前，这个世界还没有像现在这么具体、形象和几乎就要实现。这就是语义网的世界。

试想，如果每一本书都有一个独立的网址，每一个作者都有一条可以公开访问的记录，每个刊物、出版社，每个主题词、每个分类号……每个“知识点”，在网络中都有一个唯一标识，所有这些“资源”之间的关系都能从其标识所指引的地址里找到详尽的说明；甚至万事万物，不论是自然的、社会的或精神的，都有一个标识符，都建立起丰富的关联，计算机能够自动通过网络推理和挖掘知识，那将是一个多么有序的知识世界！

1 什么是关联数据？

“关联数据”所提出的技术架构，为实现这个有序的知识世界带来了曙光。

关联数据是国际互联网协会(W3C)推荐的一种规范，用来发布和联接各类数据、信息和知识，它希望在现有的万维网基础上，建立一个映射所有自然、社会和精神世界的数据网络，通过对大千世界万事万物及其相互之间关系进行机器可读的描述，使互联网进化为一个富含语义的、互联互通的知识海洋，从而使任何人都能够借助整个互联网的计算设施和运算能力，在更大范围内，准确、高效、可靠地查找、分享、利用这些相互关联的信息和知识。

从技术上看，关联数据是在万维网上发布任何“资源”的一种方式。语义万维网将资源定义为“任何有URI标识的东西”，分为信息资源和非信息资源两类，信息资源用以表达任何信息，通常以某种编码的文件形式而存在；非信息资源用以指代大千世界中的各类实体对象，可以是自然界、人类社会以及人类意识所创造的精神世界(概念、观念、抽象实体等)的所有对象。

关联数据通过HTTP URI方式表示和存取“资源”。如果这个资源是信息资源，则可以直接通过传统的Web方式获取；如果是非信息资源，则链接到一个以RDF/XML编码的、用以指代该“非信息资源”的数据文件，而不是其他任何格式的文档。这个RDF/XML编码的文件包含了关于这个“非信息资源”的元数据描述和与其他相关实体对象的关联关系描述。对象之间的关联关系通常可以用本体语言来编码，许多领域应用的知识体系都有规范的、可重用的本体，可用来建立实体对象之间的关联关系。

关联数据的 URI 除了能够在万维网范围内唯一标识资源对象之外，还能起到定位的作用，从而能够用以“关联”数据。具体的关联是依靠 RDF 文件中的大量资源链接来实现的，这些链接不仅决定了数据的语义，也通过“属性”而关联到其所能链接到的、大量的相关资源实体。这些“属性”本身也是资源，也应该有唯一标识符 URI 加以定义和描述，我们通常所称的“元数据方案”就是这类属性的集合，规定了所需进行描述的语义及其相互关系，其本身就可以看成是描述某些特定对象的本体。

关联数据的发明人蒂姆·伯纳斯－李（Tim Berners-Lee）为关联数据总结了四个原则，很好地概括了上述关联数据的诸多特性：

（1）使用 URI 作为任何事物的标识名称，不仅是标识文档；

（2）使用 HTTP URI，使任何人都可以参引①（dereference）这一全局唯一的名称；

（3）当有人访问名称时，以 RDF 形式提供有用的信息；

（4）尽可能提供链接，指向其他的 URI，以使人们发现更多的相关信息。

其中第三和第四点要求 RDF 文件包含有用信息以及尽可能多的 URI，这就要求关联数据的 RDF 文件尽可能不使用“空白节点（blank nodes）”和少使用普通“文字（literal）”。在这里，“空白节点”是没有全局 ID 的本地资源（没有定义命名域的 URI，如 ISBN，DOI），“文字”指一个字串值（可以有类型以及语言属性），由于这两种描述方式都不能用来指代“资源”，因此过多地使用“空白节点”和“文字”不能起到数据（即资源）关联的作用，实现关联数据的目的。

总之，可以认为关联数据是一组最佳实践的集合，它采用 RDF 数据模型，利用 URI（统一资源标识符）命名数据实体，来发布和部署实例数据和类数据，从而可以通过 HTTP 协议揭示并获取这些数据，同时它强调数据的相互关联、相互联系和有益于人机理解的语境信息。

2 关联数据能做什么？

关联数据可以看成是语义万维网的一种简化实现，作为一种语义信息的编码、发布和利用方式，它的作用是基础性的和多方面的。从目前的研究开发项目来看，对关联数据的应用主要体现了两个方面的作用：一、提供“可信网络”的语义要素；二、作为跨网域数据整合的通用 API。它最终是为了用户更准确地、从更大范围、适时适地（just-in-time 和 just-in-case）地获取信息而服务的，但最终用户无需知道这些服务背后的技术细节，因此关联数据的“用户”，目前还主要是指图书馆、网站、信息提供商之类的机构组织，常被称为“信息中介”。

“可信网络”意为其信息资源的来源可追踪或可通过一定算法计算其“信度”的网络。关联数据的技术架构不仅提供了信息资源可以追踪来源（具有 URI）的 RDF 语义描述，而且为各类对象实体以及所涉及的大量概念术语提供了规范控制。例如对每个作品、表达、表现，或作者、机构、家庭等实体提供一个唯一的 URI 参引，或对每个主题、概念、术语、事件、分类词或属性词等，提供一个唯一的出处。这实际上就是传统图书馆学中“书目控制”（又称权威控制）的扩展：当人们提及某一实体，或某一概念术语时，系统能够给予自动的归并或参照。这种机制，就是规范控制。规范控制的结果，就是信息在一定程度上更加可信。

若要进行跨网域的数据整合，关联数据把 API（应用程序接口）统一为 HTTP 一种，只不过经过了简单的扩展而已（指 Hash 或 Slash 方式转发）。也就是说关联数据对数据访问方式进行了标准化，用户或代理无需知道某具体关联数据发布网站的体系架构、存储方式等任何技术细节，只要知道 Web 服务器地址，都可以直接用 SPARQL 进行访问。

据此，目前的关联数据应用系统的开发，基本上也可分为两类：“关联数据仓储系统”和“关联数据服务系统”。前者关心的是将数据发布为面向网络的关联数据仓储，后者关注不同仓储的整合应用和互操作。当然，这两者也不是截然分开的，某些应用兼而有之，是这两者的联合。

目前把各类数据发布为关联数据是一个热点，图书馆行业在这方面已成为先锋，不仅将本行业历久弥新的各类概念体系受控词表发布出来（即将各类知识组织体系发布成 SKOS），越来越多的元数据方案、本体，乃至图书馆传统的各类规范档（如书目

① 注：这里的“参引”（dereference），意指“为了获取引用资源的相关信息，在万维网上查找 URI 的过程”。下同。

记录、人名、地名、机构名等）都在探索以关联数据的形式发布，而且在发布过程中探索了领域本体（如FRBR）的应用。下一步我们可以期待，重点将会逐渐转移到跨应用的语义整合服务，例如各类术语体系或元数据的映射等。

当然，关联数据也不是万能的，它最大的敌人就是封闭，无法对封闭系统中的资源进行整合。目前图书馆购买或租用的大量资源库需要远程访问才能获得，如果这些资源库不提供一定的开放接口，关联数据就无计可施，最多利用本体和术语规范的关联数据，从服务整合的角度，提供一定的资源导航或术语规范的支持。

总之，关联数据相比于语义万维网技术来说，其实现更加简单，但背后同样有数学和逻辑学的支持，具有规范性和可靠性。作为一种数据发布技术，由于支持了语义描述，同时提供标准的服务接口，有效地提高了数据的可查找性和可重用性，其影响力正在日益显现，潜力十分巨大，已成为影响互联网基础结构的关键技术之一。

3 关联数据是如何实现的？

关联数据是建立在Web技术之上的，Web技术主要涉及三个内容：HTTP、URL和HTML。

- HTTP是服务器操作的指令，规定了遇到各种请求（如GET/PUT /POST/DELETE）服务器如何响应，怎么处理；
- HTML是存储在服务器端的网页文件，将根据请求传送给浏览器，HTML的标准规定了文件的结构，允许包含丰富的超文本链接，并能嵌套各类其他文件格式，如果浏览器一端有相应的资源或程序就能够调用或运行。正是由于HTML，使整个万维网上布满了相互链接的文件，成为一个巨大的、不断膨胀的文件宇宙，这就是为什么说目前的万维网是文件的万维网（Web of Documents）的原因。
- URL本来是作为在这个文件宇宙中定位具体的文件而用的，后来演变成兼具名称作用，从而连同URN一起，统一作为URI的子类。

关联数据把上面三个技术作了进一步的限定和扩展，用URI同时解决命名和定位问题。在具体实现URI命名和定位时，由于该名称有永久性和易实现的要求，路径作为某个资源名称的一部分，不允许随意发生改变，并且在不同的软硬件平台和技术环境下都需要能够正确编码，这就需要作为关联数据标识的URI符合CoolURI规范。

同时对于同一个对象，必须允许有不同的描述与表达方式，例如对于“http://www.kevenlw.name/about/index.php”中关于kevenlw的FOAF[①]描述，既要有html文件（php可以认为是动态生成的html文件），通过浏览器显示给人看，又要有rdf文件描述kevenlw的各种性状属性以便机器获取相关元数据信息，如foaf文件：http://www.kevenlw.name/kevenfoaf.rdf。这两个文件其实描述的是同一个“东西”，因此不应该有不同的ID标识（注意：在这里是两个不同的URI，这是不规范的），必须在一个URI中区分这两类数据，同时让服务器有一种机制，能够自动地根据请求方的不同，传送不同格式的数据。

关联数据的具体实现方式解释如下：

一、对于来自客户端的对任何非信息资源的所有URI“参引”请求，均采用HTTP协议中的“内容协商”规则，返回其所请求的信息资源描述文件（对于非信息资源的请求是无法返回具体实物对象的，只能以描述该对象的代码文件代替）。一般信息资源描述文件有两类：即如果请求来自于普通浏览器（头信息中包含text/html请求，其他MIME文件类型，如图像文件、音视频文件等，可归入此类），则返回HTML文件的网页；如果请求为application/rdf+xml，则返回负责该对象语义描述的RDF文件。

二、具体的“内容协商”方式，通常有两种方案达成：

（1）采用HTTP协议的303指令重定向功能（如图1所示[②]）。客户端（浏览器）的URI请求由于不存在“东西”（非信息资源），服务器就会发送一个303 See Other给客户端，再由客户端根据重定向规则发送请求，具体根据客户端是HTML浏览器还是支持RDF的浏览器，决定HTTP文件头请求何种类型的文件（HTML或者RDF）。

该过程的具体流程如图2所示[③]：

① FOAF是个人信息描述的一种RDF格式，参见：http://www.foaf-project.org/。

② 示意图来自BBC关联数据项目报告，原图地址：http://www.bbc.co.uk/blogs/radiolabs/s5/linked-data/ui/images/slash303conneg.png。

③ 原图来自参考文献14，地址：http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/deref-ont-uri-rdf.png

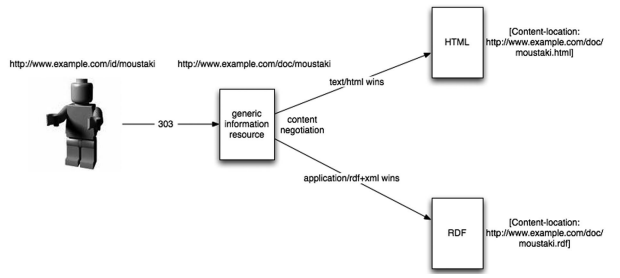

图 1　HTTP 协议 303 指令重定向示意图

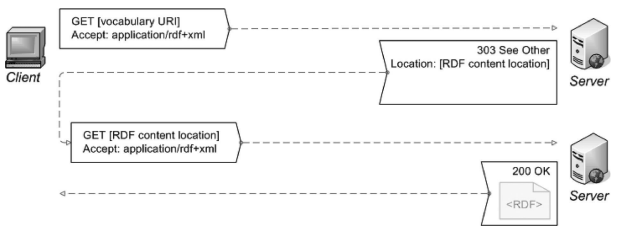

图 2　HTTP 协议 303 指令重定向流程示意图

URI 重定向通常采用以下惯例：

A

http://www.kevenlw.name/kevenliu（ID）

http://www.kevenlw.name/kevenliu.html

http://www.kevenlw.name/kevenliu.rdf

B

http://www.kevenlw.name/resource/kevenliu（ID）

http://www.kevenlw.name/page/kevenliu

http://www.kevenlw.name/data/kevenliu

C

http://id.kevenlw.name/kevenliu

http://page.kevenlw.name/kevenliu

http://data.kevenlw.name/kevenliu

（2）采用带“#”号（hash）的 URI 方式（如图 3 所示①）。“#”号前面的 URI 能够便于浏览器进行解析定位，而与后面带“#”号的片段标识符共同用来标识非信息资源，该片段标识符同时起到了类似于重定向的功能，允许支持 RDF 的浏览器参引到信息资源文件（在这里是静态的 RDF 文件）的所需位置。这种方式要求该片段标识符必须在 RDF 文件中是唯一的，且整个 RDF 文件不可过大，否则非常影响查询效率。

采用“#”号方式作为 URI 的例子如：

http://www.library.sh.cn/people.rdf #kevenliu

http://www.library.sh.cn/people.rdf #leonzhao

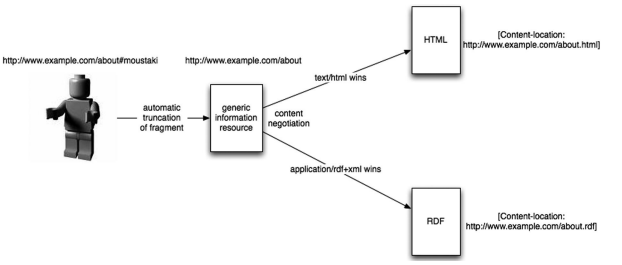

图 3　采用“#”进行“内容协商”定位资源描述的示意图

由于关联数据从技术上看只是一种简单的数据发布规范，规模较小的应用只需要对现有的 Web 服务器软件进行一定的设置，设定好资源对象的 URI 命名规范（以如上所述的各种方式），并将这些资源的 RDF 描述以静态文件的形式发布出来。对于海量数据仓储，则后台必须有支持关联数据规范发布方式的数据库管理平台，目前开源软件已经有著名的内容管理平台 Drupal② 全面支持关联数据，Ruby on Rails③ 据说也已开发了完整的支持模块。另一个做法是利用关系型数据库系统的管理功能，编制映射文件，实时地将数据表、行、列、值映射为 RDF 数据中的类、属性、资源、属性值（文本与连接）等。这种方式通常被称为 D2R 方式，即从数据库到 RDF 数据转换的方式。这样等于在原有的 Web 数据库三层应用架构基础上增加了语义构建层（即生成 RDF 数据以供 SPARQL 查询），大大简化了语义内容的构建难度，发布速度快，但也带来了语义标注一致性差、质量不高的问题。目前 LOD④（即开放关联数据 LOD：Linked Open Data）中有很多大型数据集都采用了这种方式发布。关于关联数据发布的详细解释，可以参考 Chris Bizer、Richard Cyganiak 和 Tom Heath 合著的 How to Publish Linked Data on the Web 一文⑤。

① 示意图来自 BBC 关联数据项目报告，原图地址：http://www.bbc.co.uk/blogs/radiolabs/s5/linked-data/ui/images/hashconneg.png。

② 参见：http://drupal.org/

③ 参见：http://www.rubyonrails.org/

④ 参见：http://linkeddata.org/

⑤ 参见：Chris Bizer，Richard Cyganiak，Tom Heath. How to Publish Linked Data on the Web.［2011－01－18］. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

4 关联数据在国外的研发应用现状

2006 年 7 月蒂姆·伯纳斯—李提出关联数据，由于其主要是一套应用规范，而不是难度很高的技术开发，很快成为互联网研究和应用的一个热点领域。在 2007 年开放关联数据运动的推动下，不久便出现一大批实验性的应用，表示关联数据应用范围的云图不断增大①，关联的开放数据呈几何级数飞速增长，截止 2010 年 11 月，LOD 中的数据集合已有 100 多个，其中 RDF 三元组数据已达 131 亿。其内容也逐步扩展，从早期的地理信息、生命科学数据、百科词条等，发展到目前涉及媒体、出版、政府信息、图形图像等，几乎无所不包。

除了关联数据专题会议之外，2007 年以来几乎每个互联网国际会议都以关联数据作为主题或最重要的分主题，如全球互联网大会（WWW）、语义万维网年会（ISWC）、AAAI 年会、DCMI 国际元数据年会等。自从 W3C 的 2007 年年会（即 WWW2007）之后，关联数据就开始作为一个专门的分会场——LDOW：Linked Data On the Web，于每年召开。该会议已成为关联数据领域最重要的会议，会上所探讨的主题代表了最新的研究和开发动向，目前已从最初的关联数据的发布和浏览，到关联数据的应用架构、关联算法、Web 数据融合、关联数据的消费和关联服务等诸多方面。

关联数据领域的一个鲜明特点是边研究边应用，在实践中不断得到检验和完善。目前涌现出一批非常知名的应用，如美国和英国政府的政府信息、英国广播公司（BBC）、纽约时报、路透社、百思买等。

以下以 BBC 为例，简要介绍关联数据对于组织机构内部数字资产管理和利用所带来的变化。

BBC 是世界上最大的广播电视公司之一，创立于 1920 年，目前有 32 种语言的国际服务，8 个全国电视频道，1 个高清频道，大量的地方频道，10 个国家电台，40 多个地方电台等，积累了难以想象的资料和素材，管理、发现和重用这些资源都是巨大的挑战，更别说开放出来给公众使用。

BBC 矢志成为业界翘楚。它的网站 bbc.co.uk 开设于 1994 年，是同行业中最早的网站，语义网技术使它燃起了新的希望，它希望建立先进的语义媒体库，不仅利用网站进行节目推广，而且可以发布、推送、组织和存档节目，支持知识搜索，使其积累的大量内容成为储存人类记忆的脑库。于是它利用关联数据技术，给每个节目（每一集）都建立了自己专属的网页和静态地址（CoolURL），每个知识单元都有自己的结构化描述和永久地址，而且每个网页都可以由所有这些知识单元根据模版自动生成，同时以同样的方法建立了 455465 位艺术家的信息，682473 个播出节目，7851093 个音轨，以及 31112 个 Labels 的完整资料。BBC 还采用了鼓励用户贡献信息和纠错的机制，用户的参与使信息库的完整性和准确性不断得到提高。BBC 认为关联数据技术使其网站和数据的可用性得到大大增强，用户的体验得到巨大提升，搜索引擎的查询效果得到优化，资源的可查找性、可点击性和可传播性都得到很大提高。现在 BBC 的整个网站同时又是一个 API 平台，它采用了 RESTful 发布，与 Web 无缝集成，保证了链接的永久性和数据的开放性，并且其系统的各组成部分松散耦合，互有联系却互不干扰，整个系统进入可持续发展的良性轨道。

5 图书馆行业的关联数据应用

自从 2008 年瑞典国家图书馆首家以关联数据的形式发布了 LIBRIS 国家书目，并将其中的数据与 DBPedia 相关联之后，到 2010 年，已有逾 20 个图书馆的关联数据集②。

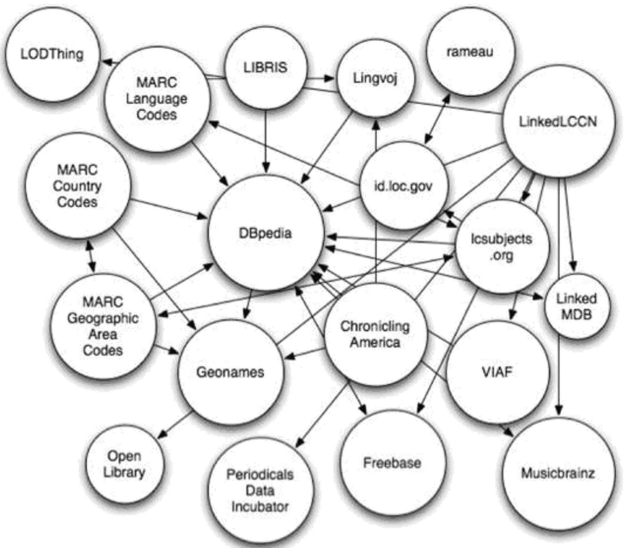

图 4 2010 年已有的图书馆领域关联数据集

① 参见：http://richard.cyganiak.de/2007/10/lod/

② 来自 Ross Singer2010 年 Code4Lib 报告 http://code4lib.org/conference/2010/singer 中的图书馆关联数据云图。

其中至少有以下5个国际、国家级的书目数据/规范数据开放了关联数据服务：

- 美国国会图书馆及其主题标目（LCSH）（id.loc.gov）
- 德国国家图书馆的联合权威档（Gemeinsame Normdatei）（d-nb.info/gnd/）
- 法国国家图书馆（BnF）的RAMEAU主题标目（stitch.cs.vu.nl/rameau/）
- OCLC的杜威分类法及国际虚拟权威档（VIAF）（dewey.info/和viaf.org/）
- 匈牙利国家图书馆的目录和叙词表（oszk-dk.oszk.hu/resource/DRJ/404）

另外DC元数据、应用了FRBR的RDA词表、BIBO书目本体（http://bibliontology.com/）、SKOS知识组织编码模式和OAI-ORE对象重用和交换模型都可作为数据关联的语义工具。

目前这类词表和KOS已经如雨后春笋一般涌现出来。较著名的有：

- STW经济学叙词表（zbw.eu/stw）
- 社会科学叙词表（lod.gesis.org）
- GEMET环境叙词表（eionet.europa.eu/gemet）
- Agrovoc（联合国粮农组织叙词表）（aims.fao.org/）
- 纽约时报主题标目（data.nytimes.com/）
- 科学出版物词表（dblp.rkbexplorer.com）

因为有了如此进展，Antoine把2010年称为图书馆关联数据元年①。

图书馆行业所具有的经年累积的高质量数据，包含了大量的、值得揭示和参照复用的内容实体，只是这些东西都隐藏在书目记录内部，没有独立标识，也缺乏结构化描述，特别是其相互之间的隐含关系尤其值得揭示，但工作量浩大，必须开发一定的规则算法，由机器进行批处理。

IFLA也注意到了关联数据与图书馆的密切联系，于2010年6月发布了《关联数据与图书馆》的专题报告[1]，由德国国家图书馆的Jan Hannemann和Jürgen Kett执笔。文章介绍了德国国家图书馆在应用关联数据技术方面的进展，包括三个具体的实例：德国作家Bertolt Brecht的规范数据、国际图联（IFLA）海牙总部的机构规范数据和主题“Führungskraft”（英语：“Executive”）的标目，探讨了关联数据对于图书馆的意义和应用前景，对于全球图书馆如何互通互联数据、并在此基础上探索新的服务内容和方式，进行了全面深入的思考。

由于图书馆行业有着独特的“规范控制”经验和长期积累的数据优势，万维网协会W3C专门成立了“图书馆关联数据孵化小组（Library Linked Data Incubator Group）”②，由DCMI的元老Thomas Baker领衔，汇集语义网、特别是关联数据方面的高手，集思广益，充分挖掘现有图书馆领域的相关专业知识，如元数据模型、元数据模式、标准和协议等，重新定义需求、编制指南、开发新的标准，鼓励图书馆界将它们的各类数据和规范档以关联数据的形式发布到互联网上，提高图书馆数据在万维网上的互操作性，使图书馆行业成为万维网上最重要的语义数据提供者，并探索和寻求与其他相关领域的数据和应用进行协同的可能性。

孵化小组目前已完成了约50多个用例（Use Cases）的收集和编写③，内容涉及书目数据、规范控制、词表发布、档案和异构数据、参考引文、数字对象、资源集合、社会性应用等各个方面，还在不断增加，涵盖非常广泛，几乎包括了图书馆行业数据可能想到的所有方面。

尽管已经取得了不小的进展，图书馆界应用关联数据的困难也十分明显，主要表现在以下4个方面：一是缺乏可资利用的、公认的术语词表，各类KOS、本体尚未经过严格的编码应用检验，而且目前也不够用、不统一；二是缺乏成熟的方法和可以立即上手的工具；三是数据的版权属性不明朗，有时可能有法律风险；四是做这个事情还是缺乏经验，需求掌握也不是很充分，为什么做？有什么用？能不能达到预期目的？还都是未知数。

6 国内的研究与应用

国内最早引介关联数据，应该是2008年12月在上海召开的“数字环境下图书馆前沿问题研讨班”

① 参见：http://talis-linkeddata-libraries.s3.amazonaws.com/I-saac-LLD10.pdf slide 6；“2010, Year 1 of Library Linked Data”。

② 参见：http://www.w3.org/2005/Incubator/lld/

③ 参加：http://www.w3.org/2005/Incubator/lld/wiki/UseCases

上，刘炜所作的“语义互操作与关联数据”介绍①，当时是为了宣传次年在韩国召开的DC－2009国际元数据会议主题，希望国内同行关注这一新的技术动向。美国著名图书情报学家曾蕾教授在同一个会上所作的题为“术语注册和网络服务系统当前技术和应用”②的报告，更为详细地介绍了关联数据技术及其应用现状。随后曾蕾教授前往中国国家图书馆和中国人民大学图书馆，又作了两场同题报告，传播了正在国外兴起的“关联数据”研究和应用。

关联数据与元数据具有天然的联系，从某种程度上可以说关联数据是元数据语义表达和实现其功能需求的最佳方式，就像业界普遍认为RDF是当然的“元数据格式”一样，RDF作为一种数据表达方式（三元组），其在Web上开放发布的最简单便捷的形式，就是“关联数据”的一整套被称为“最佳实践”的规范。尽管这些说法可能不是非常严格准确，但还是从某种程度上揭示了这些概念之间的关系。

DCMI的国际元数据年会从2008年柏林会议就有大量的关联数据讨论，这时已经经历了国外2007年关联数据的持续升温。在美国雪城大学秦健教授的推荐下，刘炜为《现代图书情报技术》组织了一个DC-2008年会会议录中有关语义网应用的翻译文章专辑，其中有两篇涉及关联数据，分别介绍了瑞典国家图书馆以关联数据形式发布书目数据[2]，以及美国国会图书馆主题标目的关联数据应用[3]。这两个应用可以说是图书馆行业在这一领域应用的先驱和样板。

由武汉华中科技大学主办的2009年“数字环境下图书馆前沿问题研讨班”③又一次涉及了关联数据主题。这次会议上由于有曾蕾教授的强烈推荐，引起了大家对关联数据的高度重视和强烈兴趣，开始认识到这是代表发展方向的一个技术领域，将对未来的网络信息资源组织和应用产生重大影响。这次会议上曾蕾和刘炜分别作了“关联的图书馆数据”④和“关联数据：意义及其实现”⑤的报告。

2010年8月上海市图书馆学会在普陀区图书馆召开了一年一度的“图书馆前沿技术论坛”，主题定为“关联数据与书目数据的未来”⑥，参加会议交流的除了上海市在该领域从事研究开发的一些专业人员之外，远在大洋彼岸的曾蕾教授也通过远程会议系统为会议作了第一个报告，会议特别邀请了新西兰奥克兰大学图书馆的资深技术专家林海青先生、中国科技信息研究所的白海燕女士和嘉兴学院的黄田青先生，一共进行了8场专题报告⑦，最后还进行了讨论和互动，全国各地约有近20位对关联数据感兴趣或正在从事研究的同行也参与了网络直播和交流。上海图书馆学会学术委员会主任范并思教授在开幕致辞和闭幕总结中对这次会议给予了高度评价。

从国内见诸专业刊物的文章来看，关联数据的研究尚不普及。除了上面提到的两篇翻译文章之外，总共只有不超过10篇论文，其中有两篇是综述文章，黄永文的综述[4]主要侧重图书馆应用的角度，沈志宏、张晓林的综述[5]则从技术发展所提供的可能性角度，介绍得更为全面系统。

其他文章也都较为详尽地介绍了关联数据技术的内容和发展[6][7]以及国外有关项目的应用开发情况[8]，白海燕[9][10]和范炜、邹庆的论文[11]涉及了项目开发和技术实现。这些论文的作者单位也反映出国内对关联数据感兴趣的机构集中在中国科技信息研究所、中科院文献情报中心等少数几家。另外已经有两篇学位论文涉及了这一主题[12][13]。

中国科技信息研究所是国内较早跟踪关联数据技术，并积极探索其应用可能性的单位，曾经有多个项目与此有关，最早的项目可以追溯到2008年在国家科技图书文献中心立项的“NSTL联合目录的分层组织与关联构建”，该项目主要探讨了FRBR在NSTL应用的可能性，提出了NSTL书目本体，并在DC-2009上发表了一篇短文（挂图Poster）。后来该所又立项了“基于关联数据的信息组织深度序化”，并成功申请2010年度国家社科基金项目“图书馆资源组织语义化研究”，全面研究了关联数据的实现技术，并进行了基本开发试验。目前基于上述成果又开展了资源整合和服务整合的研究开发，分别立项了“基于关联数据的服务融合与资源扩展”和“基于DOI的科研资源整合研究”等项目，该所在十二五规划中也打算基于关联数据技术，全面调研关联数据

① 参见：http://www.lib.sjtu.edu.cn/adls/download/12－18/1218AM－C2.pdf

② 参见：http://www.lib.sjtu.edu.cn/adls/download/12－17/1217PM-A7.pdf

③ 参见：http://202.114.9.60/dl6/

④ 参见：http://202.114.9.60/dl6/pdf/26.pdf

⑤ 参见：http://202.114.9.60/dl6/pdf/24.pdf

⑥ 参见：http://www.libnet.sh.cn/tsgxh/list/list.aspx?id=6604

⑦ 参见：http://www.kevenlw.name/archives/2199

在NSTL服务系统中的应用场景，探讨利用该技术进行知识组织系统的构建、知识关系抽取、海量文献自动标引、检索结果的扩展、异类资源整合检索、多维分面信息资源的组织与检索、数据融合与混搭等前沿领域应用的可能性。

7 问题与展望

关联数据是一项与图书情报工作密切相关的技术，是互联网发展到语义网时代、提供对任何网上资源和数字对象进行“编目”和“规范控制”的基础性技术，是数字图书馆进行信息资源发布和服务的核心技术之一。可能囿于技术障碍，我国图书情报界还没有充分认识到这一点，甚至还没有引起一些大型的、肩负指引行业发展方向的机构的充分重视，未能投入足够的人力和资源进行跟踪研究和开发试验。目前仅有的一些研究由于缺乏必要的交流而很难达成一致理解，甚至无法避免谬误和弯路。关联数据从技术上看是非常简单的，但要应用得好，必须要有领域专家、内容管理专家和网络应用开发人员共同参与，仔细调研需求，同时需要对于标准规范有深刻的理解，在模型和架构方面达成一致，即使可以边摸索实践边服务推广，也需要有一个基本的研究团队和交流环境，这些是制约目前国内关联数据研发和应用的主要问题。希望通过本文的回顾、总结和呼吁，能够使大家认识到关联数据的价值、内涵和意义，并引起一些相关机构和专家的重视。

参考文献

1 Jan Hannemann, Jürgen Kett. Linked Data and Libraries.[2011－01－18]. http://www.ifla.org/files/hq/papers/ifla76/149－hannemann-en.pdf
2 Martin Malmsten. 将图书馆目录纳入语义万维网. 李静雯译. 现代图书情报技术, 2009, 3(3): 2－8
3 Ed Summers, Antoine Isaac, Clay Redding, Dan Krech. LCSH, SKOS 和关联数据. 姚小乐、刘炜译. 现代图书情报技术, 2009(3): 8－14
4 黄永文. 关联数据在图书馆中的应用研究综述. 现代图书情报技术, 2010(5): 1－7
5 沈志宏, 张晓林. 关联数据及其应用现状综述. 现代图书情报技术, 2010(11): 1－9
6 黄永文. 关联数据驱动的Web应用研究. 图书馆杂志, 2010(7): 55－59
7 李亚婷, 曹洁, 彭洋, 鲍莹. Web环境下关联数据的应用. 情报理论与实践, 2010(11): 122－125
8 白海燕. 关联数据及DBpedia实例分析. 现代图书情报技术, 2010(3): 33－39
9 白海燕, 朱礼军. 关联数据的自动关联构建研究. 现代图书情报技术, 2010, 26(2): 44－49
10 白海燕, 乔晓东. 基于本体和关联数据的书目组织语义化研究. 现代图书情报技术, 2010. 9. 18－27
11 范炜, 邹庆. 词表资源关联化. 情报理论与实践. 2010(5): 21－25
12 宁小敏. 语义关联数据模型及其检索机制的研究[博士学位论文]. 武汉: 华中科技大学, 2008
13 娄秀明. 用关联数据技术实现网络知识组织系统的研究[硕士论文]. 上海: 华东师范大学, 2010

作者单位：上海图书馆，上海，200031
收稿日期：2011年2月8日

Overview on Linked Data: Concept, Technology and Implementation

Liu Wei

Abstract: The paper outlined the initiation of Linked Data, introduced its concept, implementation and current status of applications at home and abroad, and put emphasis on its deployment in library and information area. It also foresaw the impact on the library information services through the Web, and reviewed the related research and development in China. It concluded that, with the help of Linked data, it will be brought back the authority control to the Web at a certain level as bibliographical data and authority files in legacy library system transformed and uploaded onto the Web. Chinese librarianship has the responsibilities to catch up with the new achievement of the development of linked data technology.

Keywords: Linked Data; Authority Control; Semantic Web; Bibliographic Record