关联数据驱动的 Web 应用研究

黄永文（中国科学院国家科学图书馆 北京 100190）

摘要：简要介绍了关联数据产生和发展的背景，分析了关联数据的定义、基本原则、特点以及与 Web2.0 API 的区别，提出了国内外基于关联数据的 7 种 Web 应用类型，最后从用户界面和交互方面、关联关系的有效性、数据融合和模式映射、关联开放数据的许可 4 个方面讨论了关联数据应用面临的挑战。

关键词：关联数据；Web 应用

Research on Linked Data-Driven Web Applications

Huang Yongwen
(Library of Chinese Academy of Sciences Beijing 100190)

Abstract The paper introduces the background of the linked data, analyzes the definition, characteristic and basic principle of linked data, and presents the seven types of linked data-driven web application. At last, it discusses the existing problem and challenges in the future.

Keywords linked data ; web applications

1 引言

关联数据（linked data）这个概念来自 W3C，Tim Berners-Lee 于 2006 年首次提出关联数据的思想及四个基本原则。关联数据通过网络把以前没有关联的相关数据连接起来，已经成为推动语义 Web 发展的重要力量之一，并得到了政府、企业、研究机构、图书馆等各方面的广泛关注。

从 2007 年起，关联数据发展很快。W3C的关联开放数据运动（Linking Open Data，LOD）正式启动，一些新的和期待已久的W3C标准也发布了，如SPARAL、GRDDL、RDFa等。目前，已经有 20 亿条传统网页上的数据被自动半自动地转换成了关联数据[1]，BBC、纽约时报、CNET 等大型媒体公司把他们的海量数据转换成了关联数据，Google、Yahoo!等搜索引擎已经开始利用公开的信息作为关联数据。英国和美国等也开始了政府信息语义网的相关工作，英国政府将在 2011 年 6 月把主要的政府信息发布成可以重用的关联数据，并建立起重用数据的通用协议[2]。根据LOD项目组的统计，截止 2009 年 11 月[3]，LOD云图中已有 100 多个开放数据集合（如DBpedia、FOAF、Flickr、DBLP等），提供大约 131 亿个RDF三元组，以及大约 1.42 亿个RDF链接。参加该项目的数据集还在不断增加。

近几年，围绕关联数据召开了一系列的国际会议。WWW LDOW（Linked Data on the Web）工作组于 2008 年和 2009 年先后召开了专门的会议，讨论关联数据的出版发布与浏览、关联数据的应用架构、关联算法和 Web 数据融合等方面的内容。DC-2009 会议也开始关注关联数据和语义 Web，将会议主题定为“关联数据的语义互操作”。2009 年的 ALA 年会开设了一个关联数据主题会场，讨论“从遗留数据到关联数据：图书馆为 Web3.0 做准备”。在华盛顿举办的 ISWC 2009 第八届国际语义网大会，也有许多对关联数据及应用方面的研究。即将在斯坦福大学举行的 AAAI2010 年春季论坛,主题定为“关联数据与人工智能”。

早期的研究多集中在关联数据的发布和浏览方面，以解决将不同格式的数据发布成关联数据的问题。随着网络上关联数据集合的不断增多，近两年围绕关联数据的应用研究逐渐增

多，成为关联数据的研究重点内容。本文试图通过国内外关联数据的应用研究，来梳理利用关联数据进行 Web 应用的现状。

2 关联数据的涵义

维基百科中的定义：关联数据是语义网的主题之一，描述了通过可链接的URI方式来发布、分享、连接Web中各类资源的方法。[4]

在TED大会上，Tim Berners-Lee认为关联数据就是一箱箱数据，当通过开放标准关联在一起时，从中可以萌发出很多新事物和新应用。他认为创建关联数据，应遵循如下四个原则：[5,6]

- 使用 URI（统一资源标识符）作为对象的名称；
- 通过使用 HTTP URI，人们可以定位到具体的对象；
- 通过查询对象的 URI，可以提供有意义的信息（采用 RDF、SPARQL 标准）；
- 提供相关的 URI 链接，以便可以发现更多的对象。

简而言之，关联数据允许用户发现、关联、描述并再利用各种数据。关联数据以 URL 方式链接到一个对象，使得人们可以通过 HTTP/URI 机制，直接获得数字对象，对象可以是人、地点、概念等。关联数据有两个基本点：描述和连接，描述和连接都是为机器服务的，通过描述可以实现机器自动发现和确认，连接可以支持机器自动链接。

关联数据遵循了万维网的基本设计原则：简单、兼容、模块化设计，以及去中心化。关联数据通过采用RDF作为标准的数据格式，提供标准化的访问机制。与目前普遍采用的Web2.0 API相比，关联数据具有很多优点。利用Web2.0 API构建的应用，一般是程序层面来实现的，而利用关联数据构建的应用，则可以在数据层进行，关联数据为特定领域的应用开辟了新的可能性。下表列出了Web2.0 API与关联数据的比较：[7]

表 1 关联数据与 Web2.0 API 的比较

标准	Web2.0 API	关联数据
数据源总数	超过 1300 个	超过 100 个
连接机制	多种/变化的机制	HTTP，URI，RDF
全球 ID 的使用	很少	几乎所有的地方
发现方式	手工	手工、FYN、voiD
在应用中的使用方式	固定的	动态的可扩展的
数据交换格式	XML、JSON	RDF
API 类型	从 RESTful 到 SOAP	RESTful（只读）

3 关联数据驱动的 Web 应用的研究现状

随着关联数据的不断增加，不但减轻了整合分布式异构数据源的复杂性，同时也推动了基于关联数据的新应用。目前，国内外围绕关联数据进行了一系列的应用研究和开发，相关的研究项目主要有：OREChem 项目、瑞典联合目录（LIBRIS）、JISC 的 SemTech（Semantic Technologies for Learning and Teaching）项目、Thomson Reuters 的 Open Calais 项目、Bio2RD 项目、关联开放的医药数据项目（Linking Open Drug Data）等。

这些基于关联数据的 Web 应用主要集中在以下几个方面：

（1）利用关联数据实现数据网络和合作

OREChem项目[8]是微软资助的，由剑桥大学、康奈尔大学、印第安纳州大学、宾夕法尼

亚州立大学和南安普敦大学合作承担的项目。项目的主要内容包括：把现有的化学类数据源发布成LOD源，采用网格计算创造出新的LOD资源，开发和整合用于化学知识表示的标准本体，从发布的PDF中抽取语义化学数据。项目核心目的是设计和实施基于语义Web规则的互操作架构，允许化学研究人员对分布式的机构仓储、数据库和Web服务进行共享和重用。ORECheм项目是化学研究人员和信息科学人员共同开发和实施的，为化学领域的研究和学术发布提供了新的传播模式。虽然，项目的关注领域是化学，不过这项工作目前正在eScience网络基础设施下开展和实施，因此可以支持不同学科之间的连接。

瑞典联合目录（LIBRIS）[9]是全球第一个将书目数据发布成关联数据的联合目录，主要由瑞典皇家图书馆负责管理，开放其 200 多个成员馆大约 650 万条书目记录、20 万条规范文档记录（人名、地名、主题标目）。瑞典皇家图书馆已经开始创建从联合目录到Dbpedia的链接，为图书馆界开展关联数据的发布及应用提供了可贵的经验和思路。

（2）将语义 Web 嵌入到个人桌面环境中

Groza T等人[10]提出将文件工程为导向的自动元数据抽取与基于关联数据的信息扩展和可视化结合起来，并把结果文档无缝地整合到语义桌面。搭建关联数据与语义桌面的桥梁共分为 3 个步骤：1）抽取——采用面向文件工程的方法从出版物中自动抽取元数据；2）扩展——使用抽取出来的元数据，到关联数据云中进行检索，产生具有关联关系的元数据集合；3）整合——将丰富后的元数据嵌入到语义桌面环境中，再自动将元数据与已经存在的个人元数据关联起来。

Peter Sefton等人[11]提出采用ICE（集成的内容环境）、Fascinator（从网络到桌面的工具）、Lensfield（桌面数据处理环境）3 种系统，来拉近语义Web与eResearcher之间的距离。允许研究者通过不同的渠道来发布关联数据。

（3）关联数据及语义技术在教育和教学方面的应用

JISC-SemTech项目[12]是 2008 年由JISC资助的研究教育和教学方面的语义技术，项目在对采用了语义技术的教育工具、在英国高等教育中实际使用的语义工具和服务进行调查的基础上，提出了 3 个阶段的发展路线，主要是分步骤实现跨高等教育机构创建关联数据，以便机构联盟内部之间共享教育、教学资料、课程资料等资源，并构建教育类本体，实现基于本体的数据分析和教育感知推理应用。

资源列表管理工具（RLM）主要是对由老师指定的图书、期刊文章、网页/音视频内容资料等教学参考资料进行管理和发布的工具。由于现有的RLM工具与资源发布平台、图书馆目录、虚拟学习环境（VLE）之间缺乏数据之间的互操作性，学生不能对教学参考资源进行标注、评价，不支持学生根据特定学习任务需要增加新资源，缺乏反馈机制和互动性服务，因此Talis提出采用语义Web技术来解决RLM工具存在的问题，具体包括：使用现有的本体对资源统一描述、采用关联数据原则改善数据互操作性、使用现有的模式和本体（如FOAF和SIOC）来描述关系、鼓励学生和教师丰富数据的语义以支持环境感知推荐功能等。Talis系统[13]不仅实现了统一描述学习资源，还丰富了资源的语义描述，已于 2008 年 9 月在普利茅斯大学实施，Talis打算到 2010 年实现在使用它的RLM工具的三分之一英国大学客户群中推广和使用。

（4）基于关联数据构建 Mashup 服务

关联数据很好地描述了数据的结构信息，在数据层就建立了链接机制，可以通过 URI 来确保机器能够自动链接各种数据，为信息聚合的智能化自动化提供了基础。利用开放数据构建 Mashup 服务的项目主要有：Bio2RDF 系统、Paggr 系统、SensorMasher 平台、Revyu 等。

Bio2RDF系统[14]是实现生物知识整合的Mashup系统，它基于 3 个步骤的方法构建了生物数据的Mashup系统。Bio2RDF系统成功地将语义Web技术应用到公开发布的数据库中，与共享

的通用本体关联起来，创建了生物方面的知识空间。

Benjamin Nowack[15]提出了构建在基于HTML的widgets通用模式之上实现Web信息集成和显示的Paggr系统。目前，大多数Web widgets都是基于专有的API进行构建的，Paggr系统则是采用SPARQL的操作方法来实现的。可以组织成类似WIKI的仪表板，提供基于浏览器的开发工具，允许开发者协作式地创建widgets和写SPARQL脚本。在2009年的欧洲语义ESWC会议中，第一次公开发布了Paggr系统。面向特定的会议任务创建了个性化的paggr Web应用，可以浏览即将召开的会议和会议论文，从选定的公司或有关特定主题的论文中发现演讲者，或者跟踪Twitter和Identi.ca服务中的与ESWC相关的讨论。

（5）基于关联数据实现本体的再利用

LOD 云图中有些是由本体和实例相结合的资源，如 DBPedia、YAGO、UMBEL、GeoNames 等，这些数据集提供了大量的概念、概念之间关系以及具体的事实。DBPedia 是 Wikipedia 关联数据的结构化的表现，提供了大量跨领域的知识库。YAGO是涵盖了Wikipedia和WordNet的大部分内容，YAGO 经过人工修正，其知识组织的正确性达 95%以上。UNBEL 将 OpenCYC Ontology 中的 2 万主题概念类目及其相关关系抽出来。

一些关联数据Web应用直接利用LOD云图中的本体资源，作为知识库的基础。目前围绕DBPedia实现的应用比较多，如Falcons语义搜索引擎、UAd 分析工具等。大多数知识库只是面向特定领域，由相关专家进行人工维护，很难随着领域的变化做出及时的调整。而DBpedia知识库是跨领域的，由发布者自己维护，具有较好的动态性。Falcons语义搜索引擎[16]是中国东南大学开发的，提供对象、概念及文档的语义检索，Falcons系统收集了DBpedia 知识库的内容。UAd 分析工具[17]是利用关联数据帮助市场研究人员跟踪Web上讨论的工具，它采用了DBPedia中的分类框架，提供基于Web的论坛（通过SIOC和FOAF）中讨论的内部关联。

（6）关联数据的语义 Web 搜索引擎

关联数据的语义 Web 搜索引擎主要包括两种应用类型：一是抓取和存储 Web 上的关联数据，形成语义 Web 搜索引擎，如 Arnetminer、Falcons、SWSE、Swoogle、Sindice、Watson 等；二是在已有的搜索引擎之上，增加与关联数据的连接，如 Yovisto 等。

Arnetminer[18]是清华KEG实验室研制的关于学术研究网络的搜索和挖掘引擎，它使用了语义Web本体，扩展了FOAF。主要提供研究人员及其出版物的搜索，其中的数据主要来自DBLP，具体包括如下功能：检索学术研究人员或特定领域的专家，获得人员的详细情况和出版物；检索会议或出版物，获得更为详细的信息；检索两个研究人员之间的联系，如两个教授之间的可能联系路线，根据两个学术机构之间的路径长短进行联系加权。

Yovisto[19]是学术报告和会议视频的搜索引擎。它提供基于内容的演讲录音搜索，可以有效访问超过 6200 个来自于世界各地的大学和科研机构的演讲录音。Yovisto通过关联数据丰富了搜索引擎的检索结果，改善用户的使用体验。将Yovisto的内容与关联数据网连接起来，将外部的其他信息纳入到Yovisto中，同时还使用外部信息交叉连接再回到Yovisto自己的内容中。

（7）利用关联数据实现自动语义问答

利用关联数据实现自动语义问答的应用主要包括：关联开放的医药数据项目（Linking Open Drug Data）、DBpedia 移动服务（DBpedia mobile）、语义 CrunchBase Twitter 机器人 (Semantic CrunchBase Twitter Bot)等。

关联开放的医药数据项目[20]把不同来源的关于医药方面的数据进行关联，并在此基础上回答相关的科学和商业问题；DBpedia移动服务[21]是移动环境下的关联数据应用，是一个以定位为中心的用于移动设备的客户端应用，地图与位置信息来自DBpedia数据集；语义 CrunchBase Twitter 机器人[22]支持人们就有关硅谷公司方面的问题进行提问，如董事会、首席执行官、业务问题等。语义CrunchBase是关联数据的封装器。

4 关联数据应用面对的挑战

随着关联数据规则被越来越多的数据提供者所采纳，目前已经产生了许多关联数据的应用。然而，在实际的应用中还存在一些问题，如关联数据的质量和可信度、关联 URI 的有效性等。为了更好地利用和共享关联数据，还有许多问题需要研究和解决，还面临着许多挑战。

（1）用户界面和交互方面的挑战

从用户的角度来看，关联数据的最大好处是可以提供多个分布式异构数据源的整合的关联的访问。关联数据浏览器允许用户在不同数据源之间进行浏览，不过在关联数据的导航和检索结果显示方面还不尽人意，需要进一步完善。例如，关联数据浏览器的导航控制应该为用户提供实体之间的前进和后退功能。关联数据浏览器需要从传统以文件为中心的浏览，转向以实体为中心的浏览视图，改变为用户提供应用服务的角度和焦点。另外，用户不仅需要浏览实体之间的链接，还需要方便地分析大量数据。

（2）关联关系的有效性问题

LOD 项目推动了许多机构将数据发布到 Web 上，并与其他数据源相互关联。不过一旦 LOD 数据源有所变动，数据源之间的关联可能会产生断链，会引起基于 LOD 的应用随之也发生错误。目前的实践是忽略这些问题，把它留给应用层来解决，当具体的应用发现断链时再解决。不过，为了降低应用层对断链的处理，LOD 数据源应该提供关联集成的高可用性。同样，关联数据源也应提供监测和修正机制，以维护数据参照的完整性。

（3）数据融合和模式映射问题

数据融合是将代表真实世界里的同一对象的多个数据项整合成统一的数据的处理过程。从分布式的数据源检索到数据后，应该在显示给用户之前，以有效的方式进行数据融合，或者做进一步整合处理。目前，大多数关联数据应用都是分别显示不同数据源的数据，并没有做更多的融合处理。为了进一步处理和集成数据，需要将来自不同数据源词表的术语映射到目标应用模式中。关联数据源有的使用他们自己的词表，或者在现有的多个词表的基础上，根据需要增加自定义的术语。目前，W3C 推荐采用 RDF 模式和 OWL 来定义基本的术语，如 owl:equivalentClass、rdfs:subClassOf 等。为了支持不同数据源模式之间的转换，需要数据源发布它们本地的术语体系和与 Web 中相关数据源的术语体系的对照，还需要解决数据冲突问题，对来自不同数据源的相同实体进行融合。

（4）关联开放数据的许可方面

为了鼓励数据拥有者发布 Web 数据，同时为了保证数据的使用者以不侵权的方式使用数据，提供关联数据发布规范的适当框架是非常重要的。目前，已经有了基于版权概念的创作开放许可，然而版权法并不适用于数据，可以考虑采纳诸如 ODC PDDL 框架（Open Data Commons Public Domain Dedication and License，开放数据共享公共领域奉献和许可）。为了方便地使用 Web 中发布的关联数据，数据拥有者除了遵循发布规范外，还需要提供获得这些规范的方式（如在用户界面上等），方便使用者获得利用关联数据源的明确术语规范。

参考文献：

1 刘伟.关联数据：意义及其实现.[2009-07-17].http://www.kevenlw.name/archives/1435

2 HM Government.Putting the frontline first: smarter government.[2009-12].
http://www.hmg.gov.uk/media/52788/smarter-government-final.pdf

3 W3C SWEO Linking Open Data community.Linking Open Data Project Description.[2009-11].
http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

4 wikipedia.linked data.[2009-11].http://en.wikipedia.org/wiki/Linked_data

5 Tim Berners-Lee.Linked Data. [2009-06-18].http://www.w3.org/DesignIssues/LinkedData.html

6 曾蕾.关联的图书馆数据.数字环境下图书馆前沿问题研讨班,2009,7,武汉
http://cnlib20.ning.com/profiles/blogs/guan-lian-de-shu-ju-linked-1
7 Michael Hausenblas. Linked Data Applications. [2009-07].
http://linkeddata.deri.ie/sites/linkeddata.deri.ie/files/lod-app-tr-2009-07-26_0.pdf
8 Lagoze,Carl.The oreChem Project: Integrating Chemistry Scholarship with the Semantic Web. In: Proceedings of the WebSci'09: Society On-Line, 18-20 March 2009, Athens, Greece.
9 Anders Soderback,Martin Malmsten.LIBRIS－Linked Library Data. [2009-01-16].http://blogs.talis.com/nodalities/2009/01/libris-linked-library-data.php
10 Tudor Groza etc.Bridging the Gap between Linked Data and the Semantic Desktop.The Semantic Web-ISWC 2009. [2009-10].http://data.semanticweb.org/papers/iswc/2009/in-use/paper206.pdf
11 Peter Sefton etc.Boundaryless eResearch: use the Web, use Linked Open Data. Nov 2009. 3rd eResearch Australasia Conference,Sydney,Australia
12 Thanassis Tiropanis etc.JISC - SemTech Project Report.[2009-07].
http://www.jisc.ac.uk/media/documents/projects/semtech-report.pdf
13 Chris Clarke, Fiona Greig.Case Study: A Linked Open Data Resource List Management Tool for Undergraduate Students. [2009-01].http://www.w3.org/2001/sw/sweo/public/UseCases/Talis/
14 Francois Belleau etc.Bio2RDF: Towards a mashup to build bioinformatics knowledge systems.Journal of Biomedical Informatics, 2008(41)
15 Benjamin Nowack.Paggr: Linked Data widgets and dashboards.Web Semantics: Science, Services and Agents on the World Wide Web, 2009.7(4)
16 falcons.http://iws.seu.edu.cn/services/falcons/objectsearch/index.jsp
17 同7
18 LI Juanzi etc.Arnetminer: expertise oriented search using social networks.Frontiers of Computer Science in China.2008 2(1)
19 J Waitelonis,H Sack.Augmenting Video Search with Linked Open Data.[2009-09].
http://www.hpi.uni-potsdam.de/fileadmin/hpi/FG_ITS/papers/Harald/ISemantics09-Final.pdf
20 A Jentzsch etc.Linking Open Drug Data. [2009-09].http://triplify.org/files/challenge_2009/LODD.pdf
21 Christian Becker,Christian Bizer.DBpedia Mobile: A LocationEnabled Linked Data Browser.LDOW2008, April 22, 2008, Beijing, China
22 同7