



关联开放政府数据的研究与应用进展

袁远明^{1,2}, 吴产乐^{1,2,3}, 艾浩军^{1,2}

(1. 武汉大学计算机学院 武汉 430072;

2. 国家多媒体软件工程技术研究中心 武汉 430072; 3. 武汉东湖学院 武汉 430212)

摘要

开放政府数据持续稳定增长,已成为政府与市民间至关重要的沟通桥梁。利用关联数据技术发布政府数据,不仅能提升政府数据的透明度,更能提高政府数据的重用率。本文对利用关联数据技术进行数据发布的相关概念、发布原则及优势进行了描述,在此基础上深入分析涵括内容表达层、创建层、互联层、浏览/查询层4层结构的关联政府数据技术体系,总结英美发达国家在关联政府数据上的应用实践,最后展望政府关联数据在我国智慧城市建设中的应用以及挑战。

关键词 关联数据;开放政府数据;语义网

1 引言

影响公众及公共事务等多方面的政府数据不断增加,为满足民众对政府职能与义务的知情需求、促进政府公共数据潜在价值的应用开发、增强政府不同部门间的数据互操作,开展开放政府数据的研究与应用对构建服务型政府有十分重大的意义。

然而,传统政府数据中缺乏语义互操作的格式化数据(xls、csv)以及不具备语义关联的非格式化数据(txt、doc、ppt、pdf、html)大量存在,异构的政府数据间的互操作难以实现。

改变当前政府数据的发布标准将是解决问题的重要手段。互联网和语义网等新技术的发展,已使各国政府意识到基于开放标准进行政府数据发布的重要性,应更多地考虑数据之间的关联和复用,同时使得机器能够理解并发现更多的相关数据。

2 相关概念

2.1 关联开放政府数据

关联数据采用资源描述框架(resource description

framework, RDF)数据模型,利用统一资源标识符(uniform resource identifier, URI)命名数据实体,通过 HTTP 浏览并获取这些数据,通过“链接”揭示数据的关联关系以及能被计算机理解的语境信息。关联数据也称 RDF 数据, RDF 是用于描述 Web 资源的标记语言,是由“主体(subject)—谓词(predicate)—客体(object)”3部分(即三元组)组成的资源描述通用模型。关联政府数据的发布不要求使用者将数据拷入原数据集,只需创建不同数据源间的数据链接即可发现相关信息,以满足政府数据发布的需求^[1]。

开放数据尚无统一定义,来自维基百科的理解为:开放数据将资料开放给任何人使用,不论是出版还是做其他的运用,不受著作权、专利权以及其他管理机制的限制^[2]。所有的定义最终均以实现信息的开放与获取、共享与重用为目标。

开放政府数据作为以建立更加开放、透明、高效政府为目标的开放政府运动的重要部分,是政府与市民间的重要交流通道,其最直接的优势是让政府更加透明,市民能访问不透明的原始政府数据。

关联开放政府数据(linked open government data,



LOGD)是一项很有前景、能更加高效访问开放政府数据的技术,也是关联数据技术在开放政府数据方面的重要实践应用。

2.2 LOGD 的原则及优势

关联数据社区^[3]提出了一系列在网络上合作发布与互联结构化数据的最佳原则^[4]:使用 URI 作为事物的名称,允许用户使用 HTTP URI 查找这些名称;当查找一个 URI 时,以 RDF 提供有用的信息,包含与其他 URI 的 RDF 链接,以便发现更多的相关信息。数据提供者遵循上述原则发布数据,将原始数据加入关联政府数据集合,并被各种程序加以利用。

关联政府数据的优势体现在如下 3 个方面^[5]。

- 开放:能以开放的形式被各种应用获取并使用。
- 组合化:关联数据之间或与其他关联数据间可以混搭,如城市医疗健康方面的政府数据与该市的人口、环境等方面的数据联合起来,用于评估政府医疗投入的效率。
- 可拓展性强:互联方式的改变不会影响数据本身。

3 LOGD 的技术体系

依据以上发布原则及要求,归纳分析关联数据的技术体系,该体系可分为关联数据的表达、创建与发布、互联、浏览与检索 4 个层次,如图 1 所示。

3.1 表达

LOGD 表达的是具体的政府数据内容,数据表达采用 RDF 数据模型,用 URI 予以标识,能通过 HTTP 调用,如 <http://www.w3.org/people/EM/contact#me>, 这个 URI 可用来指代 Eric Miller 的人名信息,对应的关联数据的 RDF 三元组表达为{'<http://www.w3.org/People/EM/contact#me>', 'Full Name', 'Eric Miller'}。HTTP URI 是抽象的概念,需要采用内容协商^[6]实现 URI 到具体数据记录的映射,从而获取如 RDF/XML 文本数据、HTML 网页等具体的内容。

3.2 创建与发布

LOGD 创建是实现 LOGD 的关键环节,采取 3.1 节的表达方式描述对象的内部结构及其内含语义,其关联深度取决于内容对象本身和所属元数据格式的丰富程度,创建过程是基于元数据格式将元数据转换成 RDF 数据。

政府数据大多以关系型数据库 (relational databases, RDB) 形式进行存储,需要考虑如何将 RDB 数据转换成

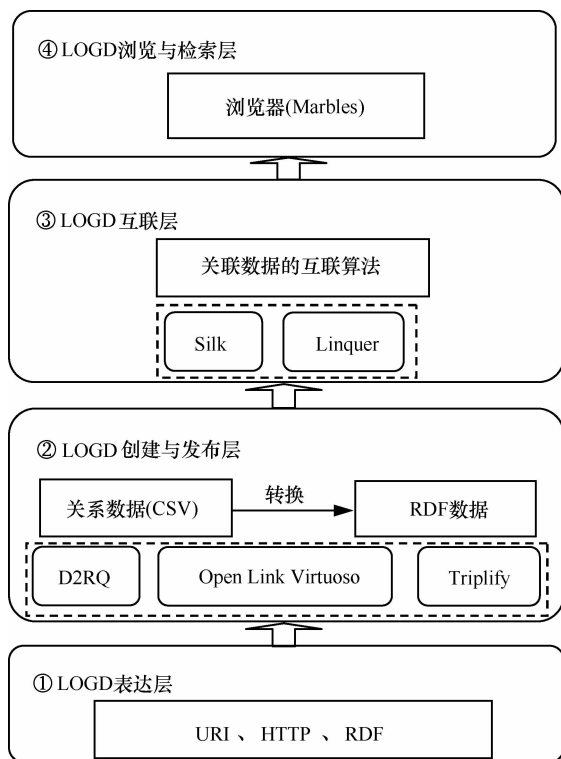


图 1 关联数据的整体技术体系

RDF 数据。RDB 的模式是二维表,而 RDF 三元组是用二元关系表达主客体间的关系,两者虽有不同但都是基于现实世界而构建,具备建立映射的条件。RDB 和 RDF 数据的具体映射方式见表 1。

表 1 RDB 数据到 RDF 数据的映射

RDB 数据	RDF 数据
column header	class
table cell	entities
cell values	subject, object
column	predicate

使用转换工具将 RDB 数据转换成 RDF 数据,当前流行的转换工具有 D2RQ Platform/D2R Server^[7,8]、Open Link Virtuoso's RDF Views^[9]、Triplify^[10]。Linked MDB、DrugBank 等很多数据集都采用 D2RQ Platform 进行数据发布。

3.3 互联

LOGD 的互联层建立了同一数据集中内容对象间的关联关系,与其他数据集的互联是其主要工作,也是整个 LOGD 体系的关键环节。

关联数据的互联技术是近年来的研究热点,在消费关

联数据专题研讨会(COLD)发布的研究热点中,关联数据的互联算法居于首位。基本互联方法主要有共同键匹配、字符串匹配、子图匹配3类,关联架构主要有 Silk^[11]、LinQuer^[12]。其中,Silk 是基于规则的互联框架,是根据两个数据集中数据的属性相似度计算它们之间的互联关系;LinQuer 是一种从语义上发现关系数据集间的关联,从而生成 SQL 查询的架构,包含 LinQuer 语言、Web 接口、一个能将 LinQL 查询转换成 SQL 查询的 API、一个能更容易用 LinQL 查询编写的接口。由于 LinQuer 采用模块化和通用化进行构建,使得 LinQuer 能容易地使用用户定义关联发现算法进行充实和丰富。

3.4 浏览与检索

建立互联后的 LOGD 需要浏览器浏览数据,并依据 RDF 链接实现数据间的导航,当前较常见的关联数据浏览器包括 Tabulator RDF Browser、Disco Hyperdata Browser、OpenLink RDF Browser、Object Viewer、Marbles RDF Browser 等,均运行在服务器端。

关联数据浏览器有别于传统浏览器,它运行在服务器端,为用户提供基于 Web 浏览器的界面;还有些 RDF 浏览器(如 LongwellMSpaceFacet)因无法在不同数据源之间浏览,不能称为关联数据浏览器。搜索引擎技术用于检索所需的数据,常见的关联数据搜索引擎包括 Falcons、Sindice、Watson、Semantic Web Search Engine、Swoogle 等。

两者的有效结合,能以更智慧的方式提供浏览服务,Marbles 就是较好地融合两种技术的关联数据浏览器。

4 LOGD 的应用实践

开放数据在信息组织与信息发现中的应用相当广泛。在政府公共信息服务领域中,美国最先开展关联数据的实践应用,基于关联数据标准建立了 data.gov 政府数据门户网站。英国(data.gov.uk)、加拿大(data.gc.ca)、澳大利亚(data.gov.au)、新西兰(www.data.govt.nz)等国家相继建立起本国的政府数据网站。与此同时,相关研究机构也纷纷成立,世界首个开放数据研究所于2012年5月在英国成立,目标旨在帮助公共部门更有效地使用政府数据,更好地发掘开放数据的商业价值。

4.1 美国政府数据

2009年5月,全球第一个国家政府层面的数据门户网站(data.gov)正式上线,数十万项政府专用的数据被白

宫开放到互联网,是政府各部门数据得以互操作的重大实践成果。

data.gov 是一个数据可自由获取、公众与政府机构互动、API 开放的网络信息共享平台。数据采集方式为分工协作、多点聚合,数据集来源有两类:一类以联邦政府所辖的能源安全、财政、司法、教育等部门发布的数据集为纵向数据链;另一类以美国各市州为横向数据链。data.gov 以都柏林核心元数据集为标准,采用目录分类方式组织政府数据,提供类目、机构、关键词等搜索途径。其以一套完整的元数据分类体系集中管理数据资源,按照“Raw Data Catalog”、“Tools Data Catalog”、“Geo Data Catalog”3类将资源编制为一级目录,并在每个一级分类检索模块中,为所有数据资源建立了两个独立的二级分类体系,一个是以数据所属的社会领域角度为出发点,另一个是以数据的来源机构为出发点,以提高检索效率。

政府各部门专门设有信息主管,负责数据的审核、提交与发布、与 data.gov 的沟通等工作,定期在 data.gov 发布可供公众自由获取的高价值数据集。最早有11个机构提供了76项数据集,之后数据集不断攀升,见表2。

表2 data.gov 数据集的发展情况

时间	Raw Data	Tools Data	Geo Data	Total Catalog
2009年11月18日	728	353	110 077	111 158
2010年11月16日	2 951	642	303 352	306 945
2012年07月08日	4 613	1 275	416 314	422 202

4.2 英国政府数据

继美国政府数据网站 data.gov 后,英国政府开放数据门户网站 data.gov.uk 于2010年1月正式投入使用,旨在建立政府数据间的关联,便于公众获取政府数据。

data.gov.uk 已从上线之初的2500多个政府数据集发展到超过8607个,涉及人口与健康、交通与环保、教育与商务等领域^[13]。该网站由“万维网之父”Tim Berners-Lee 等人创立,在数据组织方面使用 RDF、URI、SPARQL 查询语言、关联数据 API 等高效的国际标准和新技术,保证与关联数据原则及协议的一致性,较好地实现数据的有效性 & 互联。

英国各届政府重视更多地发布政府数据,新一届政府提出了“数据权”的新概念,指明数据权是信息社会一项基本的公民权利,承诺深入推进以使之制度化。



data.gov.uk 的建立,使得政府数据更加透明化、政府信息更加关联化,对政府与社会关系的转变有重大意义。

4.3 我国 LOGD 的设想

当今,智慧城市已成为全球城市发展的新热点,在全球智慧风潮和国家政策的鼓励下,北京、上海、武汉等纷纷加入“智慧城市”建设的行列。智慧政务作为智慧城市的首要建设项目,是进一步加快推进服务型政府建设的重大举措。智慧政务的目的就是要增强政府信息的透明化水平,提升政府数据的重用效率,实现政府数据的透明、开放、共享以及政府智慧化决策与管理。

4.3.1 部门数据集之间的互联提升政府数据的利用效率

提升政府数据的利用效率,需要在开放政府数据的同时,实现政府各个部门数据集之间的互联。通过数据集间的互联挖掘潜在价值信息,实现数据交互重用,并最终实现城市智慧化决策与管理。

(1)城市可持续发展水平智慧评估

如经济部门、环保部门、卫生部门,都在各自网站上公布了自己的业务数据集,若 3 个网站的数据集内部及外部都实现了关联,分析人员可将这 3 个网站背后的 GDP 数据、污染数据和人们的健康指数联接糅合起来形成新的应用,通过应用挖掘潜在信息,评估城市的可持续发展水平。

(2)政府医疗投入效率评估

如城市卫生部门的医疗健康方面的政府数据,可以和城市人口统计部门的人口数据、环保部门的环境数据等联合起来,从而对政府医疗投入的效率进行评估。

(3)部门间数据的高效共享

城市行政服务中心作为提供城市电子政务服务的重要载体,是连通各部门数据信息的重要枢纽。而连通各部门数据实现数据间的共享重用,需要建立以各部门数据集为基础的统一数据交互模型,该模型可采用 RDF 架构、URI 资源标识方式、关联数据等高效的国际标准和新技术来构建。以市民办理房产交易为例:行政服务中心的房地产管理处窗口工作人员可以通过共享民政部门的个人婚姻信息、公安部门的户籍信息审查相关资格信息,市民不需要在多个部门收集资格材料就可以快速进行房产交易。通过数据关联实现部门数据的高效共享,明显提升行政服务能力。

4.3.2 智慧城市 LOGD 组织和应用

在建设智慧城市及服务型政府的指引下,基于关联数

据技术建立国家及城市的 LOGD 网站十分必要。当前政府体制内不同部门之间的利益难以协调,难以实现数据的整合以及公民对政府的监督和问责。可以参考国外开放数据的建设经验,建立国家信息化推进办公室,统一指导国家开放政府数据的研究与应用工作,加大开放数据关联技术研究,按照关联数据的内容组织框架及原则,建立国家开放政府数据网站。并对开放政府数据进行立法,保证开放政府数据的有效使用。更重要的是,在运用关联数据技术处理国家开放政府数据的同时,应考虑到具体的国情和中文信息处理方面的问题,只有与现实情况相结合,才能做好政府关联开放数据的工作。

5 结束语

基于关联数据标准发布政府公共数据,让机器更好地理解 and 处理这些数据,充分重用和挖掘政府公共数据潜在的价值,以简化行政服务流程,提升行政服务水平。开放政府数据作为开放数据的重要应用领域,关联开放政府数据将更好地推动开放数据在其他行业的发展。我国在借鉴英美政府实践经验的基础上,建立了我国的开放数据许可环境,并健全了相关法规。

笔者在描述关联开放政府数据相关概念的基础上,分析了关联开放政府数据 4 层结构的技术体系,总结了国外典型开放政府数据的应用实践,重点讨论了关联政府数据在我国智慧城市建设中的应用与挑战。面对我国政府尚无开放数据、关联数据技术尚未得到有效应用的现状,建立我国关联开放政府数据的网站,实现政府数据的关联化将是一个长期的过程,有许多工作要去做。

参考文献

- 1 Linking open data, 2012
- 2 Open data. http://en.wikipedia.org/wiki/Open_data, 2012
- 3 Christian Bizer, Tom Heath, Tim Berners-Lee. Linked data—the story so far. International Journals of Semantic Web Information System, 2009, 5 (3):1~22
- 4 Tim Berners-Lee. Linked data. W3C Design Issues, 2006
- 5 Qian Guofu. Government data online release based on linked data. Library and Information Service, 2012(5)
- 6 Best practice recipes for publishing RDF vocabularies. <http://www.w3.org/TR/swbp-vocab-pub/>, 2012
- 7 D2R server. <http://d2rq.org/d2r-server>, 2012
- 8 D2R platform. <http://d2rq.org/>, 2012
- 9 Open-link software. <http://virtuo-so.openlinksw.com>, 2012

10 Triplify. org: overview. <http://triplify.org/>, 2012

11 Volz J, Bizer C, Gaedke M, *et al.* Silk-A link discovery framework for the web of data. Proceedings of LDOW 2009,

Madrid, Spain, 2009

12 Oktie Hassanzadeh. Linkage Query Writer, PVLDB, 2009

13 Datasets. <http://data.gov.uk>, 2012

Research on Linked Open Government Data and Application

Yuan Yuanming^{1,2}, Wu Chanle^{1,2,3}, Ai Haojun^{1,2}

(1.Computer School of Wuhan University, Wuhan 430072, China; 2. National Engineering Research Center of Multimedia Software, Wuhan 430072, China; 3. Wuhan Donghu University, Wuhan 430212, China)

Abstract Open government data (OGD) is growing steadily and it has been a vital communication channel between government and citizens. Government data published by linked data technology can promote transparency and even improve reuse rate of data. The paper describes the related concepts, publishing principles and advantages of linked government data. And then it deeply analyzes technology architecture which includes expression layer, compiling layer, interlinking, and browse/query layer on the basis of related concepts. Following it summarizes applications of linked government data in USA, UK and so on developed countries. Finally, it gets a view of applications and challenges of linked government data in China smart city construction.

Key words linked data, open government data, semantic web

(收稿日期:2012-07-10)

• 简讯 •

首次全球最大规模 IPv6 过渡技术测试 11 月在中国举办

目前,IPv6 过渡,这一世界难题一直困扰着全球各国政府、标准化组织、运营商、设备厂商和互联网服务商,在全球范围内对过渡技术进行全面评测已迫在眉睫。

作为全球 IPv6 商用的主要推动者,IPv6 Forum 联合 IETF、BBF 提议开展全球范围的 IPv6 过渡技术测试活动。经过中华人民共和国工业和信息化部(简称工信部)电信研究院和北京天地互连信息技术有限公司(以下简称天地互连)的积极争取,全球首次过渡技术测试大会定于 2012 年 11 月 12 日-15 日在北京举行。

本次大会由中国下一代互联网示范工程(CNGI)专家委指导,工信部电信研究院和天地互连全球 IPv6 测试中心联合承办。2012 年 11 月 12 日-14 日在北京邮电大学开展测试,11 月 15 日召开测试情况总结会议。IETF、BBF、IPv6 Forum 等国际组织为本次测试活动的联合指导单位,欧洲 ETSI、美国 IOL、日本 WIDE 等区域组织为协办单位。中国电信、中国联通、中国移动、中国教育科研网、NTT、法

国电信、Yahoo 等单位为支持单位。

据悉,本届大会是迄今为止全球最大规模的 IPv6 过渡技术测试活动,会议预计将吸引来自中国、美国、日本、欧盟等 30 多个国家和地区的 600 多位业内专家参加,本届会议将成为全球下一代互联网产业发展的里程碑事件。

本次测试大会旨在加强国际交流,推动 IPv6 过渡技术的发展,促进下一代互联网产业的成熟,避免在过渡关键阶段大规模使用 NAT,从而导致将来二次改造的尴尬。测试活动为终端厂商、设备厂商、业务应用厂商提供展示商用过渡设备和解决方案的国际平台,为高校和研究机构提供了用于验证创新性想法与原型系统的测试实验环境。测试活动将为运营商、互联网公司、大型企业网络过渡和业务迁移提供建议。测试活动充分利用全球优势资源引入先进的测试体系和标准,对国际主流的双栈、隧道、翻译技术进行测试,也对我国提出的 Softwire、IVI、SAVI、Smart6、Laft6、PNAT 等国际标准进行测试。据悉,测试报告由工信部电信研究院、天地互连全球 IPv6 测试中心、欧洲 ETSI、美国 IOL、日本 WIDE 联合签发,未来将纳入“IPv6 Ready”和“IPv6 Enabled”国际测试认证体系。