DOI:CNKI:11-2746/G2.20110927.1323.001

关联数据开放应用协议*

张春景　刘　炜　夏翠娟　赵　亮

摘　要　关联数据的应用正迅速发展并涉及到商业、媒体、出版、政府等领域。在关联数据发布、消费和再创造的过程中,必然涉及参与者的利益问题,需明确不同类型和归属的数据(或数据集)的所有权、发布权、使用权、收益权等。关联数据的发展迫切需要相关的许可协议。目前关联数据用到的协议主要有 CC(知识共享)家族的 CC BY-SA、CC0、CC BY、CC BY-NC 和 ODC(开放数据共用组织)的 PDDL(公共领域奉献和许可),ODbL(开放数据许可)、ODC-By(开放数据共享署名许可),以及 OGL(Open Government License,开放政府许可协议)等。对开放数据相关许可协议在国内的应用提出建议。表 1。参考文献 15。

关键词　关联数据　应用协议

分类号　G254

The Open Application Licenses of Linked Data

Zhang Chunjing, Liu Wei, Xia Cuijuan & Zhao Liang

ABSTRACT　For the past few years Linked Data has been increasingly used in business, media, publication and governmental areas. In the process of publishing, consuming and re-creation of Linked Data, issues as to the ownership, the rights to use and publish, and who can benefit from Linked Data need to be appropriately addressed. Currently, the main agreements are Creative Commons family including CC BY-SA, CC0, CC BY, and CC BY-NC; the Open Data Commons family including Public Domain Dedication and License, Open Data License, Open Data Commons Attribution License and Open Government License and so on. Some suggestions on the application of Linked Data in China are put forward. 1 tab. 15 refs.

KEY WORDS　Linked Data.　Application license.

1　高速增长的关联数据

关联数据的应用正以惊人的速度发展,很多知名公司加入其中,示范性应用此起彼伏、层出不穷,其中被人谈论较多的如美国和英国的政府信息、英国广播公司、纽约时报、路透社、百思买等[1]。其所涉及的数据类型和学科领域也迅速扩展,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、政府信息和图形图像等,几乎无所不包。带来的结果是数据总量呈爆炸式增长,从 2007 年 5 月开放关联数据(LOD)云图中仅有的 12 个开放数据集,增长到 2011 年 9 月①的 295 个开放数据集②,其中包含 310 多亿个 RDF 三元组,5 亿多个关联链接[2],Chris Bizer 等人对这些数据作了详细统计[3](见表 1)。

*　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:张春景,Email:zhangchunjing@gmail.com

①　www.linkeddata.org 上最新数据的统计日期为 2011 年 9 月 19 日

②　参见:http://richard.cyganiak.de/2007/10/lod/

表 1 关联数据的学科领域分布及数量[4]

领域	数据集数量	三元组数量	百分比	(外部)链接	百分比
多媒体	25	1,841,852,061	5.82	50,440,705	10.01
地理	31	6,145,532,484	19.43	35,812,328	7.11
政府	49	13,315,009,400	42.09	19,343,519	3.84
出版物	87	2,950,720,693	9.33	139,925,218	27.76
跨领域	41	4,184,635,715	13.23	63,183,065	12.54
生命科学	41	3,036,336,004	9.60	191,844,090	38.06
用户产生的内容	20	134,127,413	0.42	3,449,143	0.68
	295	31,634,213,770		503,998,829	

数据的发布是为了利用。关联数据提供了一种强大的网络应用方式,通过规范的 RDF 描述并以标准的 HTTP 协议发布,从而使大量的数据集能够随时为人所用,关联数据提供的语义接口,能够使数据无缝嵌入到万维网上的其他应用中,看起来就像是使用本地数据一样[5]。同时由于以 HTTP 作为标准 API,其通用性使得任何数据服务商都能以极低的成本整合数据,整个 Web 可以看成一个由无数数据提供商共同建设的数据库服务器,Mashup 变成一种最基本、最常用的构建应用的方法。这些都是目前正在兴起的"关联数据的消费"所研究和讨论的话题[6]。

在利用关联数据进行发布、消费和再创造的过程中,必然涉及参与者的利益问题,必须明确不同类型和归属的数据(或数据集)的所有权、发布权、使用权、收益权等,需要制定成本可行、操作简便的框架,建立起一套保证关联数据可持续发展的制度和机制。好的社会运行制度和政策法律框架能够明确利益相关者的权利义务,调节其关系,从而促进新技术的普及。

2 关联数据发布和利用协议

在关联数据出现以前,就已有若干的开放协议,例如现有的开源软件协议达到 68 个之多,其中由知名机构发布的、使用较广泛的主要有 Apache License 2.0、BSD 开源协议、GPL、LGPL、MIT license 和 Mozilla Public License 等,与开放存取相关的协议有知识共享许可协议(Creative Commons License)、免费文献许可协议(GNU Free Documentation License)、开放内容和开放出版物学科协议(Open Content and Open Publication License)、设计科学许可协议(Design Science License)、共享文件许可协议(Common Documentation License)等,其中知识共享许可协议因其灵活的授权机制而得到广泛应用[7]。

正如关联数据的定义中提到的一样,关联数据主要是实例数据和类数据,因此已有的开源软件协议和开放存取协议并不适用于对关联数据进行授权和声明,而需要一种全新的适用于关联数据自身特点的协议。

LOD 的 207 个开放数据集中,目前已经有 18 个数据集公开了其协议信息,这些数据集中用到的协议主要有 CC BY-SA、CC0、CC BY、CC BY-NC 和 PDDL[8]。

CC BY-SA、CC0、CC BY、CC BY-NC 都属于 CC(Creative Commons,知识共享)下辖的授权协议,CC 是由劳伦斯·莱斯格创立的一个非盈利组织所运营的协议①,主要用于创作性作品权利的声明。PDDL(Public Domain Dedication and License,公共领域贡献和许可)是开放知识基金会(Open Knowledge Foundation)所管理和运行的 ODC(Open Data Commons,开放数据共用)协议中的一种,主要针对的是公开发布的数据和数据库[9]。除了以上两种协议外,不久前,英国政府

① 参见:http://cn.creativecommons.org/

还发布了 OGL(Open Government License,开放政府许可协议),主要用于政府拥有的公共数据。

2.1 开放数据共用(Open Data Commons)①

大部分关联数据是以公开的 Web 形式存在的,且希望通过开放数据运动公开更多的数据,关联开放数据(Linking Open Data,LOD)小组就是以开放数据来展示关联数据技术,因此业界已有的针对开放数据的协议 ODC 非常适用于关联数据中的公开数据。

开放数据共用(Open Data Commons,ODC)是一种开放数据的知识产权声明,用以规范、约束、明确数据拥有者、发布者、使用者在获取、传播、利用、再生产数据时的权利和义务。开放数据共用是开放知识基金会的一个项目,由其咨询理事会运行。和基金会一样,ODC 是一个非盈利的组织,于 2008 年 3 月发行了第一个开放数据许可,即公共领域的贡献和许可证(PDDL)。ODC 家族目前有三种形式,分别是 PDDL、ODbL(Open Data License,开放数据许可)和 ODC-By(Open Data Commons Attribution License,开放数据共享署名许可)[10]。

PDDL 主要针对公共领域的数据和数据库。PDDL 从一开始就声明开放数据共用不是一个法律机构,不提供任何形式的法律服务。PDDL 由 Jordan Hatcher 和 Dr Charlotte Waelde 合作撰写,目前的版本为 1.0。

由于该协议下的数据处于公共领域的状态,因此对于接受方没有任何限制与要求。相同地,贡献方毫无保留地放弃任何权利,并且在已经使用 PDDL 的同时,不能进行其他类似版权或数据库权利的声明。

该协议所涵盖的范围包括:①法律效力,包括对于版权和数据库权利的弃权声明和在司法管辖区不允许放弃版权和数据库权利的许可声明。②合法权利,包括的内容有版权和数据库两方面,版权一般包括数据库模型,数据库表单和表单索引,数据条目,输出表单和内容的字段名。数据库权利包括对内容的提炼和再利用。该协议涵盖的范围不包括创建或操作数据库的电脑程序,数据或数据库的专利,还有与数据库相关的商标[11]。

如果所有人放弃对作品的所有权利,这种将行为延续到其继承人或继任者;如果和相关的司法管辖区的法律违背,则有可能会再次拥有作品的版权和数据库权利。

ODbL 是一种允许用户无偿共享、修改和使用数据的许可协议,类似 GPL 协议和 CC 的相同方式共享协议,要求对于修改后的数据以相同的方式共享并且署名,即用 CC 相同方式共享协议来定义数据库的权利。

ODbL 管辖的内容仅仅针对数据库权利,不包括数据库中的内容。其具体内容如下:①法律效力,包括对使用版权和邻接权的许可、对数据库权利的许可、许可人所签订的协议。②合法权利,包括版权、数据库权利和合约三部分。版权和数据库权利规定的内容和 PDDL 一致。合约是使用者和许可人对于存取数据库签订的一种协议。③使用者的权利包括对内容的提炼和再利用,可以创建衍生数据库,创建数据库集合和创建再生品等。④权利的终止条件是,如有违背协议中的条款,则协议自动终止,并不会通知接受方。若停止违背行为,则协议 60 天后自动恢复。

协议不涵盖的内容和 PDDL 的规定相同,即不包括创建或操作数据库的电脑程序、数据或数据库的专利、与数据库相关的商标。

ODC-BY 是 2010 年 6 月 24 日新发布的一种数据库的特定许可协议,需要署名数据库。它类似于创作共用的署名许可协议,但只是为数据库设立的。ODC-BY 的大部分内容和 ODbL 相同,不同之处在于它明确提出了对于数据库和数据的署名问题,不必揭示相同方式共享的需求。

开放数据共用的声明也非常简单,只需要在网页的显著位置进行标注就可以,所标注的内容包括协议名称和协议全文的 URL 地址。

关联协议(Connected Commons)是 Talis 公司发布的一种协议,目的是直接支持公共领域

① 参见:http://www.opendatacommons.org/

关联数据的发布和再利用。Talis 公司提供一个平台,数据提供者可以不需任何花费就可将数据发布在 Talis 平台上,同时数据的提供者和用户都可以无偿使用 Talis 的数据服务。Talis 作为一个开放数据的服务商,通过这种方式可以吸引更多的数据提供商,从而开拓 Talis 的 Web3.0 服务和云计算服务。目前 Talis 上的内容已经有 5 亿个 RDF 三元组和 10G 的内容,并支持对于 RDF 的 SPARQL 查询语言[12]。

Talis 的关联协议其实就是对 PDDL 或者是 CC0(放弃版权的一种声明协议,见后文)的一种应用,进入 Talis 平台的所有数据和内容都必须遵守 PDDL 或者 CC0。

目前已有大量的项目和数据集在使用 ODC 的相关协议,例如 ShareGeo Open① 是 2010 年由英国政府发起的一个开放共享数据项目,属于 JISC(Joint Information Systems Committee,联合信息系统委员会)下辖的一个空间数据资料库。不过它只供英国高等教育和继续教育机构(UK Higher Education and Further Education bodies)的成员上载数据,但任何人都可以下载和使用数据。

2010 年开始的 OpenfMRI 项目致力于免费公开共享功能磁共振成象数据集,其中还包括原始数据。该项目的主要目的有两个:一个是为想要公开共享 fMRI 数据的研究者提供必要的基础设施和支持,另一个是用该数据库作为高级数据分析和高性能计算方法的实验平台。因此该项目使用 ODC 的 PDDL 作为其许可协议,目前完全开放的有五个共享数据集,其中包括 82 个主题,还有六个共享数据集即将发布。

从 2008 年 PDDL 发布以来,经过一年多的时间,该协议被广泛应用[13],除了以上提到的关联协议,ShareGeo Open 和 OpenfMRI 以外,还有 An organization ontology 项目②, open library 项目③,biblios 项目④等。

由于 PDDL 只是针对数据库权利的一种协议,因此这些项目并非仅使用 PDDL 协议,一般都是和其他协议结合使用,例如 Talis 是将 PDDL 和 CC0 结合使用,ShareGeo Open 项目是 PDDL 和 CC BY 3.0 协议的组合,而 OpenfMRI 项目中则是 PDDL 和 ODC-BY 的组合。

2.2 开放政府许可协议(Open Government License)⑤

公共数据是指那些处于公共领域的数据,包括政府公开发布的数据和已超过版权保护期限的数据。公共数据理所当然也可以发布为关联数据。关于公共数据方面的协议主要有 OGL(Open Government License,开放政府许可协议),目前版本为 1.0,由英国政府于 2010 年发布,目的是将政府拥有的公共数据提供给普通大众及企业使用,英国政府的数据现在也可在维基百科上使用。

OGL1.0 的内容近似于知识共享的以相同方式共享协议 CC BY 和 ODC-BY,涵盖的信息范围很广,包括皇家版权、数据库和源代码,并且授权使用者不限于英国本土。它还提供机读格式供人们使用工具自动发现可以使用的数据。

OGL1.0 的内容主要包括:①需标明信息提供者提供的信息来源及声明,还需提供本协议的网页链接。如果信息提供者没有提供声明,或者信息来源署名的提供者很多,但多数提供者的信息并没有运用在产品或应用中,则可以仅列此文字“Contains public sector information licensed under the Open Government License v1.0”,但绝不可声称这些信息已得到信息提供者的许可,也不得误导或扭曲信息本身或来源,且要合乎 Data Protection Act 1998 或 the Privacy and Electronic Communications 法令。②授权终止条款:如果未符合以上规定,授权自动终止。③

① 参见:http://www.sharegeo.ac.uk/docs/licenceFAQ.pdf
② 参见:http://www.epimorphics.com/public/vocabulary/org.html
③ 参见:http://openlibrary.org/
④ 参见:http://www.biblios.net
⑤ 参见:http://www.nationalarchives.gov.uk/doc/open-government-licence/

信息的授权范围不包括个人资料;在法规下不允许公布或揭露的信息;LOGO、图示(除非已整合至文件或数据集本身);军事图标;第三方人授权于信息提供者的权利;知识产权,包含专利、商标、设计权;身份信息,如英国护照。④授权用户不限于英国[15]。

OGL 的声明也非常简单,只需要给出一段可视的文字表述,并提供 OGL 的 URI 地址。另外,对于任何信息,OGL 都需要协议许可人提供关于署名的声明,这种声明一般是信息提供者指定的。

应用 OGL 的有英国政府科学办公室所发布的"未来的食物和农业:全球可持续发展的挑战和选择"的行动计划书①,米尔顿凯恩斯委员会(Milton Keynes Council)的花费信息,英国教育部发布的"支持和希望:一种满足特殊教育和残疾教育需求的新方法"征询意见稿等,大多是政府公开发布的一些研究报告和政务公开信息。

2.3 知识共享(Creative Commons)

虽然大部分关联数据是以公开数据或者公共数据的形式存在的,但是没有什么技术原因阻止私人的、私有的或订购的数据成为关联数据。两个或两个以上的企业或私人团体可以合法地通过 HTTP 在私有网络上交换私有的关联数据。关联数据可以在不同部门之间利用内联网进行交换。这种私有的关联数据必然需要完善的协议和机制来保护不同数据拥有者的权利。

知识共享组织成立于 2001 年,其唯一的目标就是在默认的限制性规则日益增多的今天,构建一个合理、灵活的著作权体系。值得注意的是知识共享许可合同不是为软件设计的,而是针对其他种类的创作性作品创设的,比如网站、学术、音乐、电影、摄影、文学、教材等作品。

CC 提供六种主要的许可协议,这六种协议规定了他人根据许可协议可以享有的一系列基本权利,这些协议从最严格开始至最宽松结束,分别是署名—非商业使用—禁止演绎(BY-NC-ND)、署名—非商业性—相同方式共享(BY-NC-SA)、署名—非商业性使用(BY-NC)、署名—禁止演绎(BY-ND)、署名—相同方式共享(BY-SA)、署名(BY)[16]。

知识共享组织还推出了 Public Domain Mark 和 CC0 两种协议,前者是针对那些非常古老的、已经丧失版权的作品的一种协议,作此协议声明的目的是使上述作品更容易被发现,从而更广泛地被应用。CC0 是对自己所拥有的作品的版权和其他权利放弃的一种声明协议,一旦某作品被声明为 CC0,则任何人可以以任何方式和任何目的使用该作品。和其他 CC 协议不同的是,CC0 和包括 GPL 在内的许多软件协议是兼容的,也可以用于软件领域。

CC 在其网站提供了一种简单的填写表格的形式使用户能非常便捷地来声明自己的作品,表格填写完成后,CC 网站会给出一段代码,用户将该代码复制至自己的网站即可。

CC 的应用非常广泛,例如 LOD 的 18 个公开其协议信息的数据集中有 12 个数据集使用的是 CC 协议,其中使用 CC-BY-SA 协议的有 3 个,使用 CC-BY 协议的有 4 个,使用 CC0 协议的有 4 个,使用 CC-BY-NC 协议的有 1 个。

3 结语

关联数据运动在欧美等一些发达国家已形成规模,与之配套的相关协议也紧随其后,虽然还没有成为正式的法律文本,但已能够通过"协议"方式起到规范利益相关者的作用,为关联数据的发展提供了一定的法律基础和保障。我国目前还没有机构组织推出任何具有影响的关联数据资源,但跟踪与应用研究早已开始,例如 2010 年 8 月,上海图书馆学会就召开了有关关联数据的专题研讨会,中国科技信息研究所、中科院国家科学图书馆等机构也发表了一些研究成果。相信这些研究都将推动关联数据在我国

① 参见:http://www.bis.gov.uk/assets/bispartners/foresight/docs/food-and-farming/11-683-future-of-food-and-farming-action-plan.pdf

的发布和利用。因此及早开展有关关联数据开放应用协议的研究和制定,能够从一开始就规范人们涉及知识产权的行为,促进这方面的应用快速发展,以满足互联网数据共享的需求。本文所介绍的这些相关协议并不能涵盖开放关联数据应用的所有协议,相信今后还会衍生出更多的协议规范,但最终能够得到普遍认可和应用的应该只集中于少数几个。总而言之,无论何种协议都是为了促进人们积极参与关联数据运动,保障关联数据运动不断向前发展。

参考文献:

[1] 刘炜. 关联数据:意义及其实现[OL]. [2011-04-10]. http://www.kevenlw.name/archives/1435. (Liu Wei. Linked Data: What for and how to[OL]. [2011-04-10]. http://www.kevenlw.name/archives/1435.)

[2] 白海燕. 关联数据及DBpedia实例分析[J]. 现代图书情报技术,2010(3). (Bai Haiyan. Linked Data and DBpedia case analysis[J]. New Technology of Library and Information Service, 2010(3).)

[3] Linked Data community. Linked Data-Connect Distributed data across the Web[OL]. [2011-04-25]. http://www.linkeddata.org.

[4] 黄永文. 关联数据在图书馆中的应用研究综述[J]. 现代图书情报技术,2010(5). (Huang Yongwen. Research on Linked Data-driven library applications[J]. New Technology of Library and Information Service,2010(5).)

[5] 刘炜. 关联数据:概念、技术及应用展望[J]. 大学图书馆学报,2011(2):5-12. (Liu Wei. Overview on Linked Date: Concept, technology and implementation[J]. Journal of Academic Libraries,2011(2):5-12.)

[6] 李佳佳. 关联数据问答[OL]. [2011-05-04]. http://www.kevenlw.name/archives/1153. (Li Jiajia. Linked Data FAQ[OL]. [2011-05-04]. http://www.kevenlw.name/archives/1153.)

[7] 李佳佳. 国外开放数据许可及相关机制研究[J]. 情报理论与实践,2010(8). (Li Jiajia. Research on open data license and the related mechanism[J]. Information Studies: Theory & Application,2010(8).)

[8] Open Knowledge Foundation. The data hub[OL]. [2011-04-25]. http://ckan.net/package/search?q=tags:license-metadata+AND+groups:lodcloud+AND+-tags:lodcloud.needsfixing+-tags:lodcloud.nolinks+AND+-tags:lodcloud.unconnected.

[9] 黄永文. 关联数据驱动的Web应用研究[J]. 图书馆杂志,2010(7):55-59. (Huang Yongwen. Research on Linked Data-driven web application[J]. New Technology of Library and Information Service,2010(7):55-59.)

[10] Paul Miller, Rob Styles, Tom Heath. Open date commons, a license for Open Data [OL]. [2011-04-18]. http://events.linkeddata.org/ldow2008/papers/08-miller-styles-open-data-commons.pdf.

[11] Eoin McCarney, Caleb Derven. Open data licensing: Trojan horse or sunken treasure? [OL]. [2011-05-10]. http://irserver.ucd.ie/dspace/handle/10197/2763.

[12] The Talis Community License (draft) [OL]. [2011-05-06]. http://www.talis.com/tdn/tcl.

[13] Western Economic Diversification Canada. Licensing open water data on water and environmental Hub [OL]. [2011-05-18]. http://www.letthedataflow.ca/content/open-data-license.

[14] The National Archives. Open government license for public sector information [OL]. [2011-05-25]. http://www.nationalarchives.gov.uk/information-management/uk-gov-licensing-framework.htm.

[15] Creative commons [OL]. [2011-04-20]. http://www.creativecommons.org.

张春景 上海图书馆研究室助理研究员。通讯地址:上海市淮海中路1555号。邮编:200031。

刘 炜 上海图书馆副馆长,研究员。通讯地址同上。

夏翠娟 上海图书馆系统网络中心研究开发部工程师。通讯地址同上。

赵 亮 上海图书馆系统网络中心副主任。通讯地址同上。

(收稿日期:2011-09-05;修回日期:2011-09-20)