

BBC关联数据实现研究*

贾君枝^{1,2} 梅玥¹

(1. 山西大学经济与管理学院, 太原 030006; 2. 中国人民大学信息资源管理学院, 北京 100872)

摘要: 英国广播公司 (British Broadcasting Corporation, BBC) 是媒体领域内最早应用关联数据技术的成功案例。对BBC关联数据的实现步骤进行研究, 以期为我国媒体机构的关联数据化提供发展思路。本文在大量相关文献调研基础上, 结合关联数据发布原则, 描述BBC数据模型及其本体, 从实体抽取、实体归类、使用URI命名实体、建立CIS与DBpedia链接、发布数据五方面研究BBC发布关联数据的步骤。BBC关联数据在降低制作成本、及时更新数据以及改善用户体验等方面具有明显的应用优势, 对于实现媒体领域的关联数据化具有深远且重要的意义。

关键词: BBC; 关联数据; 媒体

中图分类号: G254 **DOI:** 10.3772/j.issn.1673-2286.2020.09.003

引文格式: 贾君枝, 梅玥. BBC关联数据实现研究[J]. 数字图书馆论坛, 2020 (9): 18-23.

互联网被称为继报纸、广播、电视三大传统媒体之后的“第四媒体”。网络媒体传播速度快、内容多、范围广, 极大地影响和改变了人们的生活。但其信息数据量庞大, 信息载体种类丰富且信息更新速度快, 不仅增加了媒体机构制作及维护网站的工作量, 且手工制作的网站之间相互独立, 导致节目之间的相互关联变得极为困难。同时因缺乏单一、共同的综合数据来源和有效的数据发布机制使媒体机构无法及时对其网站内容以及新形式的信息载体 (移动设备、游戏机等) 上的内容进行更新, 导致媒体机构的工作效率低下且实际支出远远高于预算成本。

关联数据是国际互联网协会推荐的一种规范, 用来发布和联接各类数据、信息和知识, 最早在2007年5月提出, 旨在构建一张计算机能够理解的语义数据网络, 而不仅仅是人能读懂的文档网络, 以便在此之上构建更智能的应用^[1]。其能够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设施设施和运算能力, 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。一定程度上关联数据可帮助媒体机构解决上

述困境。当前链接开放数据 (LOD) 云图包含1 239个数据集 (截至2019年3月), 其中作为九大子领域之一的媒体领域数据集有31个, 在LOD云图中占比约2.5%^[2]。

BBC作为媒体领域应用关联数据技术的成功案例, 它自动将内容分类系统 (CIS) 与DBpedia建立连接, 使其门户网站能够呈现从关联数据云自动更新的丰富内容, 在控制成本的前提下生成动态聚合的页面而无须人工交互, 将其内部资源与外部资源有机地整合起来, 以方便用户导航、改善用户体验、为用户提供更好的服务。

实现其关联数据化不仅能够促进媒体领域关联数据的发展, 同时也对媒体行业其他组织实现关联数据化发展提供了重要的借鉴意义。基于此, 本文对BBC数据模型及其本体、关联数据实现步骤及应用实践进行分析, 为其他媒体机构的关联数据化提供一定的参考。

1 BBC数据模型及其本体

BBC制作了大量面向用户的媒体资源, 而构建数据模型是关联数据实现的关键。数据模型旨在定义该类

*本研究得到国家社会科学基金项目“数据开放环境中的词表重用问题研究” (编号: 19BTQ023) 资助。

型资源的主要实体对象及其关系,并运用本体对其进行描述,以实现数据的管理、存储和共享。

1.1 BBC通用数据模型

Thing是BBC创造并谈论的所有事物的最高级类,包含4个数据属性,即短标签、首选标签、消歧提示以及以编程方式使用且人类可读的关键字,均用字符串表

示。其中事件、机构、人物、地点、主题都是其子类,并且每个子类所需的最少元数据都由core: Thing继承,并定义了各自的属性,如人物有性别、职业、出生死亡日期与地点等。Thing指代宽泛,等同于owl语言中的Thing,可以表示Web文档的主要话题,是内容管理系统(CMS)的定位符所需的表示形式,也是BBC创意作品的标签(见图1)。

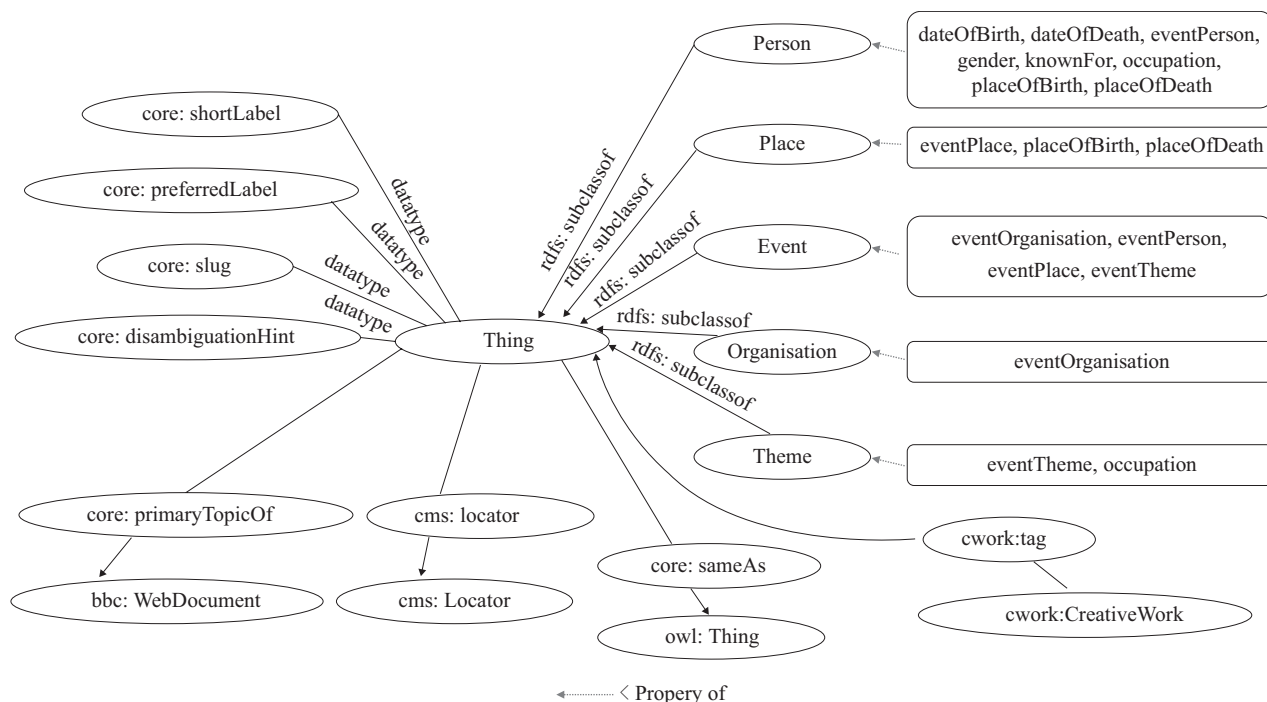


图1 BBC数据模型

1.2 BBC本体

BBC根据其当前业务需求逐步构建起多个本体,以描述所创建的内容,成为BBC关联数据平台提供关联数据服务的基础。目前平台发布了14个本体^[3],如图2所示,主要分为内容、领域、应用程序本体三大类型^[4]。

内容本体包括创意作品本体与核心概念本体。创意作品本体(Creative Work Ontology)定义了描述BBC制作的创意作品及其相关元数据所需的术语,是表达创意内容所需最少元数据的模型。核心概念本体(Core Concepts Ontology),即针对人、地点、事件、组织、主题的通用BBC本体,代表了整个BBC有意义的事物。

领域本体是根据BBC的用户需求所构建的特定领域范围内的本体,包括节目、体育、故事、课程、野生动物、政治、食品、新闻、商业新闻本体。

应用程序本体编码了应用逻辑,即帮助BBC存储、管理关联数据的本体。内容管理系统本体(CMS Ontology)是一个表示CMS与三元组存储库交互的本体,定义了关联数据平台与生成内容的系统相交互的术语。关联数据平台包含创意内容的语义元数据以及BBC生产的内容。CMS本体定义了这些事物和内容如何与表示同一事物的其他BBC实例相关联。如曼联的URI应该如何链接到提供有关曼联的更多信息(如体育统计数据)的外部内容管理系统。或者,三元组存储库中的创意作品URI应该如何引用另一个CMS。CMS本体为关联数据平台客户提供指向其他系统中某个事物附加信息的链接。起源本体(Provenance Ontology)是用于捕获有关RDF三元组存储库中数据源的数据本体。BBC本体(BBC Ontology)可用于描述网络文档、BBC产品和平台之间的逻辑联系。目前,在未来所有

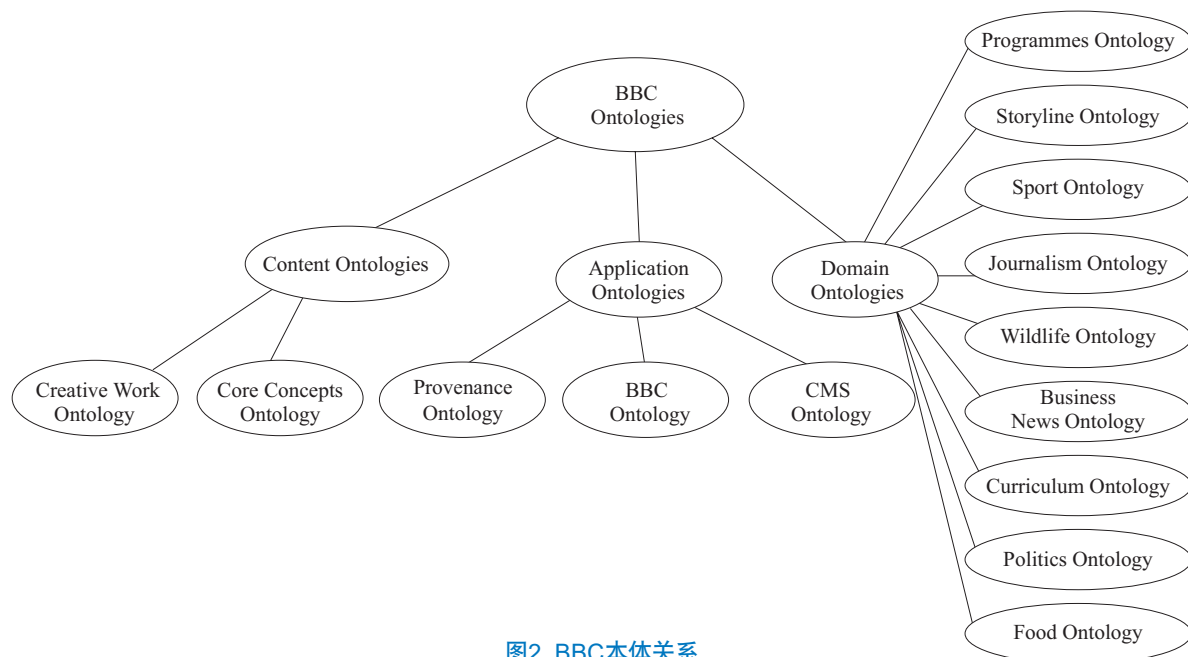


图2 BBC本体关系

的媒体产品中,以BBC在线制作为主要内容的产品共有10种,包含News(新闻频道)、Sport(体育频道)、Weather(气象频道)、CBBC(少儿频道)、CBeebies(儿童频道)、K&L(教育频道)、TV(电视频道)、Radio(电台频道)、Home(主页)、Search(搜索服务)^[5]。产品部门发布关联数据并与三元组存储库交互,这些平台制作并且发布与BBC内容相关的网络文档。

2 BBC发布关联数据的步骤

为捕捉媒体新闻报道的关键事件、人物、活动等,需提取出报道中的重要单词及短语,将之与所定义的概念建立链接,通过丰富的语义关系构造以实现数据的结构化展示及后期的聚合应用。

2.1 实体抽取

基于命名实体识别技术,从BBC文章中提取出主要实体。当前有较多的实体识别器,如OpenCalais、Twine和Zemanta等帮助用户从文本中提取实体。BBC与Rattle Research合作创建了Muddy Boots特征检索系统处理BBC文章,结合Yahoo公司的术语抽取API,从文档内容中提取出重要单词或短语的列表以进行内容分析^[6]。这些提取出的实体只是文本,并且此时没有附加语义或分类。

2.2 实体归类

对提取出的文本进行内容分析,找到其对应的概念类。BBC通过内容分类系统完成此项工作。CIS分类体系总共包含15万个术语,涵盖4个领域,即BBC品牌、地点、人物和一般主题。每个领域都有自己的一个术语层次结构,并采用SKOS表示^[7]。通过实体分析,找到与CIS对应的类,如节目概要“展望北京奥运会,英国拳击的希望”将被归类为“北京”(地点),“英国拳击”(主题)。通过“北京”一词链接到与北京有关的BBC新闻报道,有效地建立该节目与其他关于“北京”的节目之间的联系。既实现不同BBC服务的连接,又同时保持各自独立开发。

2.3 使用URI命名实体

URI是所描述事物的全局唯一名称,通过HTTP URI,很容易在线链接到每个描述并访问它们的内容。实体对象明确后,必须采用URI命名,且最好用可用的关联数据集命名,这是重用现有Web标识符的原则。BBC开发人员使用最广泛的来源包括DBpedia、MusicBrainz(一种开放式的音乐数据库)和节目本体。DBpedia为广泛的概念提供URI,还提供有关这些概念和关系的丰富结构化数据。2009年开始,BBC中提到的地点、人员、主题或组织都用DBpedia的URI进行

标记。这些标签让BBC的观众很容易就能找到特定主题的节目。而且使用链接数据URI作为标签的一个优点是不断地丰富这些标签的信息。

BBC重用DBpedia的标识符,使用DBpedia查找搜索服务^[8],即给定一个字符串(包含一个或多个单词),通过关键词搜索API来查找给定字符串的关联DBpedia资源。以字符串“berlin”为例: <http://lookup.dbpedia.org/api/search/KeywordSearch?QueryClass=place&QueryString=berlin>,得到5个查询结果,如表1所示。得到查询结果后,BBC编辑再根据上下文消歧选出最合适的URI。

表1 “berlin”的查询结果

Label	URI
Berlin	http://dbpedia.org/resource/Berlin
Axis powers	http://dbpedia.org/resource/Axis_powers
Berlin Wall	http://dbpedia.org/resource/Berlin_Wall
West Berlin	http://dbpedia.org/resource/West_Berlin
Kinshasa	http://dbpedia.org/resource/Kinshasa

2.4 建立CIS与DBpedia链接

将数据与其他数据集互联有助于数据被用户发现,也有助于将数据集发布到关联数据网络。BBC将CIS与DBpedia链接起来,依靠DBpedia标识符链入关联数据网络。与DBpedia相链接具有许多优点:其提供了丰富的关联和层次关系,可用作自动分类系统的补充材料;其已建立了与BBC领域特定数据(如MusicBrainz)的互操作,维护成本较低等。在BBC应用程序中可以使用和显示链接数据,并提供实体之间的丰富关系。

通常通过基于术语的上下文相似性聚类算法实现链接。链接过程分为两步:一是查找DBpedia标签,二是基于上下文消歧^[9]。通过DBpedia关键词和前缀搜索API,与CIS的输入字符串进行匹配。为了找到给定术语的最可能匹配,使用维基百科文章之间的链接数作为总体相关性的权重。如维基百科关于威廉·莎士比亚的文章有超过6 000个指向它的链接,而尼古拉斯·莎士比亚的文章只有18个,CIS输入的术语“莎士比亚”最有可能的匹配是“威廉·莎士比亚”。为了消除可能匹配的歧义,继续在DBpedia中通过聚类和上下文查找来识别CIS术语的相似上下文。链接算法为所有CIS术语创建集群,集群由CIS分类层次结构和括号文本组成。如

“Mary (1985情景喜剧)”归属于“电视”类。该算法基于多个可能的DBpedia匹配来识别每个集群的匹配DBpedia类别和临时值。之后,这些识别出的上下文用于消除CIS概念匹配的冗余。如“苹果”本身是模棱两可的,考虑到“微软”和“谷歌”的背景,“苹果”的含义将变得清晰。最后,系统结合前面步骤的结果,完成与DBpedia的链接。

2.5 发布数据

发布数据并不复杂,只需将数据存储在不公开访问的DNS并设置合适的权限即可。BBC通过两种方式发布关联数据,一是将数据集和本体作为关联数据发布到Web上,供用户获取;二是BBC已经开发了一种自动建模算法,用于将RDF三元组转换为带有基于主题树的注释的HTML页面。这样一来存储在数据库中的文章内容以及与之关联的链接数据将发布在HTML页面上。同时BBC已加入LOD云图,现已发布了4个数据集,分别是BBCMusic、BBC Programmes、BBC Wildlife Finder以及DBTune.org John Peel Sessions RDF server,并在关联开放词表(LOV)平台上注册了8个词表,即bbc、bbccore、bbcprov、bbccms、wlo、sport、cwork、stories^[10],这在一定程度上提高了数据集的影响力。

3 BBC关联数据的应用优势

目前,BBC开发了一个全新的关联开放数据网站BBC Things^[11],该网站包括出现在BBC节目和在线内容中的地点、人物和组织等重要概念的数据。起初只能从BBC内部访问,现在任何人都可以通过这个新网站来访问存储在BBC平台上的数据。作为BBC关联数据平台的一部分,BBC Things网站可帮助在技术或编辑层面处理BBC数据。从编辑的角度来看,网站可以使内容编辑和生产者很容易通过搜索事件、地点、人物和其他概念快速找到与内容相关的引用;从技术的角度来看,BBC以外的开发者现在可以使用BBC提供的数据创建他们自己的新网站和应用程序。BBC Things体现了BBC未来媒体的许多指导原则,它开放了整个BBC的在线数据,代表了BBC向公众开放其关联数据平台的第一步。BBC现已成功地应用关联数据技术到其节目、音乐、新闻、出版、教育等众多领域。这种成功源于在

BBC门户网站中使用关联数据,使网站能够从关联数据云自动更新并强化自己的内容,集成其他网站的数据并支持他人重用数据,这样可确保聚合页面尽可能包含更加丰富的内容。

纵观国内新闻媒体,央视网是我国重点新闻网站之一,在全国有着较大的影响力。央视网和BBC在地位和定位上均存在相似之处。央视网作为中央电视台的官方网站,拥有国内最大的网络新闻视频库,汇聚了150多个央视和地方新闻栏目。可以说,BBC和央视网都是依托本国最著名的广电平台而发展起来的全国性权威新闻网站。这些相似的媒介特点使得对二者进行比较具有了可能性。对比央视网,BBC关联数据的应用优势主要包括以下4点。

(1) 制作成本低。央视总共拥有45个频道,在中国和海外均可收看到其电视节目。每一个频道下又包含多个节目,这些节目的页面均为编辑手工制作,不仅效率低下而且网站之间的代码重用量有限,不仅浪费时间更造成了巨大的成本开销。随着移动设备的普及,央视网开通了手机电视和移动传媒服务,需要在移动平台上提供相对应的产品。平台产品的增加直接导致了工作量的冗余,需要大量的人力物力去完成。BBC Programme (BBC节目)于2007年夏季启动,利用关联数据来解决这些问题。它为BBC的每个节目提供持久的网络标识符,每个网络标识符都有多种内容协商表示,确保跨多个设备(如桌面和移动设备)推出一致的产品,并且用于生成页面的数据可以以不同格式(RDF/XML、JSON和普通模式)重复使用,以构建节目支持应用程序^[12]。BBC内的其他团队可以将这些节目页面直接纳入新的或现有的电视频道和广播电台节目页面,以及食品、音乐和自然历史等跨节目类型网站,节省了大量的人力成本和时间成本。

(2) 数据丰富且更新快,易于整合多种数据来源,并提供推荐服务。目前,央视网使用的数据资源基本上都来自政府部门公开和商业机构提供的数据。不仅数据来源单一且数据实时更新较慢。针对这样的问题,以BBC Music(音乐)为例,BBC音乐网站的目的是提供全面的BBC音乐内容指南,将艺术家的信息与播放他们的BBC节目联系起来,为音乐领域的艺术家、出版商和评论等主要对象提供持久的网络标识符,并将其与观众感兴趣的其他BBC领域信息集成。BBC音乐集成包括3种信息来源,即Musicbrainz、维基百科和BBC^[13]。Musicbrainz提供艺术家之间的关系以及外部

网站链接等数据,维基百科提供艺术家传记信息,BBC提供了曲目的音频片段、图像、专辑评论等额外信息。因此用户可以获取到艺术家广泛全面、及时更新的信息,其所提供地理项目查找服务,能够访问艺术家所在的地理位置信息或者给出与该地点相关的艺术家,并实现音乐推荐。通过艺术家可以访问许多与音乐相关的数据集,跟随一个艺术家到另一个艺术家的链接,获取到艺术家之间的联系(如这位艺术家的第一个音乐视频由与其他艺术家相同的人指导),可以用来推荐相关艺术家的有序列表。

(3) 参与编辑社交媒体,利用更广泛的网络作为内容管理系统。央视网有内部的内容管理系统,对于像央视这样业务规模较大的媒体,需要耗费大量的人力与成本,占用了大量的内存。对于BBC而言,BBC Music利用外部网络作为其内容管理系统,采用Musicbrainz和维基百科为BBC音乐网站提供基础数据,使其能够覆盖更广泛的艺术家。由于资源有限无法为BBC的每位艺术家保留传记,充分利用外部网络作为内容管理系统的构成,不仅节约内存,还可以确保数据保持最新并且不会过时。它也比维护内部内容管理系统更有效,因为内部内容管理系统需要开发和集成成本,而且很难引导、策划和维护最新的内容管理系统。如当一位艺术家去世时,这位艺术家的主页将在几小时内被外部社区更新,这一变化将直接在BBC的网站上反映出来。BBC编辑将直接为Musicbrainz和维基百科的用户作出贡献,BBC音乐将展示这些信息的汇总视图放在其背景中。

(4) 改善用户体验,利于记者创作。BBC新闻组织利用关联数据(如百科、政府类数据)以添加其上下文和进行导航。用户可以轻松从一个特定新闻页面链接到与之相关的新闻页面,也能根据主题导航轻松找到关于某特定事物的所有新闻^[14]。这不仅可以改善用户体验,而且有利于记者的创作。如果一个记者正在研究关于政治家的故事,通过确定感兴趣的政治家,利用关联数据可以合并有用的数据和文档,以提供有关政治家的背景信息。而央视网的新闻页面中缺乏相关链接,用户无法直接在当前新闻页面中通过特定关键词的链接轻松找到自己所感兴趣的内容。要想找到所感兴趣内容的全部信息需要用户自己不断搜索与翻阅历史页面。

4 讨论

以BBC为代表媒体领域的关联数据实现具有深远且重要的意义。这种实现方法在实践中也被证明是一种有效的方法, 对其他新闻媒体机构实现关联数据化有一定的借鉴意义。BBC通过实体抽取、实体归类、使用URI命名实体、建立CIS与DBpedia链接、发布数据等, 成功将关联数据技术广泛应用到众多业务领域。通过上述研究也可发现, 新闻机构利用关联数据是大势所趋, 其具有相当明显的应用优势, 制作成本低、数据丰富且更新快、易于整合多种数据来源、改善用户体验等。尽管应用关联数据有以上诸多优势, 但不可忽视的是关联数据的可靠性成为重要问题, 央视网的数据质量来源可靠, 因为政府数据比百科数据信任度更高。如何在保证数据质量的前提下有效地借鉴BBC实践, 是我们未来研究中需要进一步探讨的问题。

参考文献

- [1] 沈志宏, 张晓林. 关联数据及其应用现状综述 [J]. 现代图书情报技术, 2010 (11): 1-9.
- [2] Linking Open Data Cloud Diagram [EB/OL]. [2020-08-08]. <http://lod-cloud.net/#media>.
- [3] BBC本体主页 [EB/OL]. [2020-08-08]. <https://bbc.co.uk/ontologies>.
- [4] Sofia Angeletou. Linked Data: new ontologies website [EB/OL]. [2020-08-08]. <https://www.bbc.co.uk/blogs/internet/entries/78d4a720-8796-30bd-830d-648de6fc9508>.
- [5] Ralph Rivera. Connected storytelling-one service, ten products, four screens [EB/OL]. [2020-08-08]. <https://www.bbc.co.uk/blogs/aboutthebbc/entries/>.
- [6] Yahoo Term Extraction [EB/OL]. [2020-08-08]. <https://www.programmableweb.com/api/yahoo-term-extraction>.
- [7] MILES A, MATTHEWSB, WILSONM, et al. SKOS core: simple knowledge organisation for the web [C] //International Conference on Dublin Core and Metadata Applications. Dublin Core Metadata Initiative. 2005: 3-10.
- [8] DBpediaLookup [EB/OL]. [2020-08-08]. <https://github.com/dbpedia/lookup>.
- [9] KOBILAROV G, SCOTTT, RAIMONDY, et al. Media meets semantic web-how the bbc uses dbpedia and linked data to make connections [C] //European semantic web conference. Springer, 2009: 723-737.
- [10] LOV关联开放词表: 词表&本体查询工具 [EB/OL]. [2020-08-08]. <http://catwizard.net/posts/20160501221330.html>.
- [11] Opening up the BBC's Linked Data with /things [EB/OL]. [2020-08-08]. <http://www.bbc.co.uk/things>.
- [12] RAIMONDY, SCOTTT, OLIVERS, et al. Use of semantic web technologies on the BBC web sites [M] //Linking Enterprise Data. Boston: Springer, 2010: 263-283.
- [13] ENOB. BBC music John Peel lecture [J/OL]. British Library, 2015 [2020-09-01]. <https://www.bbc.co.uk/programmes/p033smwp>.
- [14] SHEARERM. BBC News Lab: Linked Data [J/OL]. BBC News, 2013 [2020-09-01]. <https://www.bbc.co.uk/programmes/p033smwp>.

作者简介

贾君枝, 女, 1972年生, 博士, 博士生导师, 教授, 通信作者, 研究方向: 信息组织与信息检索、关联数据, E-mail: junzhij@163.com。
梅玥, 女, 1994年生, 硕士研究生, 研究方向: 信息组织与信息检索、关联数据。

Research on BBC Linked Data Implementation

JIA JunZhi^{1,2} MEI Yue¹

(1. School of Economics and Management, Shanxi University, Taiyuan 030006, China; 2. School of Information Resource Management in Renmin University of China, Beijing 100872, China)

Abstract: The British Broadcasting Corporation is one of the largest news broadcasting organizations in the world and the earliest successful case of applying linked data technology in the media field. This article studies the implementation steps of BBC linked data to provide development ideas for linked data in Chinese media organizations. Based on the investigation of a large number of related literatures and the principle of publishing linked data, this paper describes the BBC data model and its ontology, and studies the steps of BBC publishing linked data from five aspects: entity extraction, entity classification, entity naming by URI, establishing CIS and DBpedia link, and publishing data. BBC linked data has obvious application advantages, and the realization of linked data in the media field has far-reaching and important significance, which will reduce production costs for media organizations, update and enrich data faster, and further improve user experience.

Keywords: BBC; Linked Data; Media

(收稿日期: 2020-09-03)