

中文关联书目数据发布方案研究

夏翠娟, 许磊

(上海图书馆, 上海 200031)

摘要: 本文在对国际关联书目数据调研的基础上, 论述图书馆将MARC数据发布为关联书目数据的意义, 并以上海市文献联合编目中心的中文书目数据为例, 探讨关联书目数据发布中的数据建模、本体词表设计、CNMARC与关联书目数据模型的映射设计、中文关联书目数据发布中的内容整合方案和技术实现方案, 最后总结现存问题和不足之处。

关键词: 关联书目数据; CNMARC; BIBFRAME

中图分类号: G254

DOI: 10.3772/j.issn.1673-2286.2018.01.002

1990—2010年, 互联网和大数据技术的发展使得托夫勒于1980年预言的“第三次浪潮”成为现实, Web成为“无处不在、无时不在”的信息中心, 信息获取变得前所未有的便利, 谷歌、百度、脸书、推特成为信息获取和信息交流的平台。作为知识保存、传承和传播中心的图书馆受到强烈冲击, 在互联网时代, 图书馆的业务流程和服务模式须重新思考和定义。

起源于20世纪50年代的书目控制, 是现代图书馆的核心职能之一。“全面记录并保存人类所有文献”和“通过记录达到对文献的有效检索和广泛利用”被认为是书目控制的两个基本任务^[1], 制订国际化的编目条例和实施全球范围内的联合编目, 成为书目控制的方法和手段。20世纪80年代, 在一批目录学家的推动下, 我国引入西方现代书目控制的思想, 随后中国机读目录格式(CNMARC)、《中国机读规范格式》《规范数据款目规则》《中国文献编目规则》陆续推出, 全国联合编目中心、CALIS联合编目中心、上海市文献联合编目中心相继成立。

随着全球范围内如火如荼的数字图书馆建设, 互联网成为工作、学习和生活的主要环境, 大量纸质文献被数字化, 文献种类不再限于纸质资料而扩展到多媒体资源, 大量原生数字化资源在网络上呈爆炸式增长态势。卡片目录时代以纸质文献为描述对象的MARC, 及基于MARC的书目格式和编目规则疲于应付。研究数字化文

献书目控制的“数字目录学”等议题得到高度重视^[2], 试图改善MARC封闭性和复杂性的DC元数据及其方法论得到广泛深入的应用, 并成为科技部、国家图书馆推出的数字图书馆系列标准规范的基础。在功能需求的层面, 国际图书馆协会联合会(International Federation of Library Associations and Institutions, IFLA)提出FRBR模型, 以面向对象的“实体-关系”分析方法来重新构建书目控制的功能需求框架, 以便于当前计算机技术的应用。

国内学者不断尝试将多媒体资源和网络信息资源纳入传统书目控制内容中, 或将图书馆的书目控制方法延伸到互联网应用。2008年, 美国国会图书馆提出“书目控制未来将是合作性的、分散的、世界性的和基于Web的”^[3]。因此, 在编目规则层面, AACR2升级为RDA, 充分借鉴FRBR, 基于概念来区分书目对象相关实体、属性及关系, 并对各类规范取值词表进行规定, 提高书目数据的一致性和规范性, 以更好地实现互联网环境下大规模数据的互操作。在数据模型和数据格式层面, 充分利用语义Web的本体、RDF及关联数据技术, 欧洲数字图书馆推出EDM、美国国会图书馆推出BIBFRAME, 以作为MARC的替代。与此同时, 美国、英国、瑞典、芬兰、韩国、日本的国家图书馆以及全球最大的书目控制机构OCLC, 将书目数据发布在Web上, 成为开放的“关联书目数据”, 努力为互联网时代

的书目控制夯实基础。

关联数据技术使得图书馆的书目数据从封闭、专业性强的MARC格式和编目系统中解放出来,不是仅将静态的书目数据发布到门户网站,为用户提供被动的文献查阅服务,而是在Web上提供一致、开放、标准化的开放数据服务。这样的书目数据可被搜索引擎抓取,被第三程序调用,并可方便地与互联网上的数据融合,成为互联网的有机组成部分。上海图书馆(以下简称“上图”)一直在关注关联书目数据的方法、技术和案例,本文以上海市文献联合编目中心(下文简称“上海联编中心”)的书目数据为例,探讨中文书目数据发布为关联数据的技术实现方案。

1 国际关联书目数据发布项目概述

书目数据是图书馆开展服务的基础,也是图书馆最宝贵的数据财产。随着关联数据技术、工具和流程日渐成熟,越来越多的图书馆开始将其馆藏书目数据发布为关联书目数据。2015年,OCLC开展关联数据项目调查,共得到90个机构156个项目反馈,在有详细描述112个项目中有56个关联书目数据项目^[4]。由Zepheira公司主导的Library.Link项目,参与的机构有上千家。

大英图书馆(the British Library, BL)于2011年7月开始将图书和连续出版物数据发布为关联数据,在2017年又新增CIP数据。BL发布关联数据,一方面是为响应英国政府的开放数据政策,另一方面是为探索关联书目数据应用的可能性。BL在复用Bibo、DC、Event等本体词表基础上自定义本地词表,并设计了适用于所发布3种类型文献的大英图书馆数据模型,链接到外部的数据集有ISNI、VIAF、LCSH、GeoNames等^[5]。其消费方式有SPARQL endpoint和RDF数据包下载,支持JSON、XML、Turtle等序列化格式。在2014年4月—2015年4月的关联数据利用统计中,来自于搜索引擎与关联数据爬虫的访问已达到了4 300万次,数据包下载平均每月40次^[6]。

法国国家图书馆关联数据项目(Bibliothèque nationale de France, BnF)发布了关联书目数据集,还通过data.bnf.fr为普通读者提供统一的检索服务。BnF利用FRBR模型建立作品、主题、作者间的关系,并复用skos、foaf、dcterms、RDAgroup2elements、rdvocab,自定义bnf-onto词表。由于其书目数据采用一套ARK永久标识符,且挂接到规范数据,解决了从MARC到RDF数据转

换中实体识别的难点。BnF链接的外部数据集有VIAF、DBpedia、Agrovoc、id.loc.gov等。BnF的关联书目数据以HTML、JSON、RDF等格式发布,通过SPARQL endpoint、FTP下载对外提供数据服务^[7-8],同时在HTML页面嵌入Schema.org和Opengraph Protocol词表以提高其对搜索引擎的友好性^[9]。data.bnf.fr在2012年底上线,一年内个人访问量增长了600%,页面点击量增长了400%,显示出关联数据的优势。80.6%的data.bnf.fr访问者是通过搜索引擎跳转而来,其中又有71.9%的访问者会跳转到图书馆其他服务页面^[10]。

美国国会图书馆(Library of Congress, LC)自2009年开始将其规范数据、术语词表、MARC代码和取值词表发布为关联数据。LC于2012年启动“书目框架转换行动”,开发BIBFRAME关联数据模型、词表、应用纲要、编码规范等以取代MARC,使其成为下一代图书馆数据格式,并于2017年6月在Library.Link发布^[11]。在此过程中,LC开发数据转换工具Marc2Bibframe、数据编辑工具BIBFRAME Editor等,并发布BIBFRAME应用纲要、MARC21字段与BIBFRAME词表映射等文档。BIBFRAME在国际图书馆界得到广泛关注,德国国家图书馆、匈牙利国家博物馆图书馆、意大利联合目录平台SHARE、美国国家医学图书馆、关联数据项目LD4P及上海图书馆都开展了基于BIBFRAME关联数据项目的研究和应用^[12-13]。

韩国国家图书馆(National Library of Korea, NLK)、日本国立国会图书馆(National Diet Library, NDL)也实施了本国关联书目数据发布项目^[14]。NLK于2011年开始将其书目和规范数据发布为关联数据,以SPARQL endpoint和API的形式提供RDF、JSON等格式的数据服务;NDL于2012年发布关联数据服务平台NDL Search,开放的数据包括NDL-OPAC、日文期刊索引数据、NDL数字资源及其他图书馆的电子档案等^[15]。NDL还利用API接口,为开发者提供基于DC-NDL词表描述的RDF和JSON格式的数据^[16]。

OCLC WorldCat作为世界最大的联合目录,2014年开始利用Schema.org词表对亿级书目数据进行语义化改造。其在2015年的使用统计中,平均每天的用户请求量达160万次。OCLC旗下另一个达到百万级访问量的关联数据项目是VIAF。VIAF不仅链接Dbpedia、ISNI、Getty、id.loc.gov等多个数据集,也被其他关联数据集链接。VIAF提供MARC和RDF两种类型的数据格式。除这两个项目外,OCLC还利用MARC数据中的规范档

发布作品实体 (WorldCat Work Entity, Work Entity) 和主题词的分面应用 (Faceted Application of Subject Terminology) 两个关联数据集。

欧洲数字图书馆 (Europeana) 是数字图书馆领域成功的项目之一。它是由欧洲各国图书馆、档案馆、博物馆等文化遗产机构共同参与的大型合作联盟。Europeana通过关联数据模型 (Europeana data model, EDM) 对跨机构、跨领域的多类型资源进行语义描述, 以关联开放数据发布, 并以批量下载、API、SPARQL endpoint等方式提供数据服务^[17-18]。

2 中文关联书目数据发布方案研究

关联数据的四原则对于书目数据在互联网环境下的开放、共享和融合至关重要, 当书目数据转换为RDF格式, 而书目数据所描述的文献被作为对象化的“东西”赋予HTTP URI, 即可在全网域范围内被唯一标识和定位。当在Web上访问文献的HTTP URI时, 返回RDF格式的书目数据, 为书目数据在互联网环境下跨平台和跨领域的开放共享奠定基础。在数据底层 (三元组) 关联不同网域的资源对象, 从而为互联网环境下多源数据融合带来可能。传统的OPAC为读者提供文献服务, MARC以文献为描述对象, 以记录为基本单位, 一条MARC记录通常对应单个文献; RDF以三元组为基本单位, 用以反映事实和知识, 为细粒度的数据服务和知识服务奠定基础。

关联数据技术包括数据的发布和消费两个方面, 关联书目数据采用关联数据发布技术发布书目数据, 通过关联数据消费技术提供开放数据应用开发接口, 在互联网上提供面向机器、可被第三方便调用的开放数据服务。

上图于2017年启动关联书目数据发布项目, 初步以上海联编中心的150万余种普通图书为例, 探索关联书目数据的发布流程、数据建模 (包括本体模型与词表、本体与CNMARC的映射)、内容组织和技术实现方案。

2.1 关联书目数据发布流程

关联书目数据发布流程分为数据建模、数据清洗、数据转换、数据发布四个步骤 (见图1)。

(1) 数据建模是MARC数据转换为RDF数据的基础, 为RDF数据定义主体所对应的类、谓词所对应的

属性、客体的取值约束, 并用形式化的术语词表规范表达, 形成书目本体。数据建模包括模型设计、词表设计、映射设计三个部分。

(2) 数据清洗是为书目数据中的字符串转换为实体对象作准备, 根据映射设计将CNMARC中的字段映射到书目本体的类和属性。首先, 要对CNMARC格式的数据进行解析, 将“记录-字段-子字段-值”的结构解析为“类-属性-值”的结构; 其次, 根据书目本体定义的类提取人、地、机构等规范数据, 取值词表及各类书目实体, 并进行纠错、去重、消歧、合并等清洗工作。

(3) 数据转换是生成规范数据、取值词、书目实体等实体对象并赋予HTTP URI, 根据书目本体的定义和CNMARC数据的情况对每个实体的属性进行赋值, 并以RDF序列化格式编码, 同时在各类实体间建立起关联关系的过程。

(4) 数据发布主要解决的是RDF数据存取和发布, 在Web上提供检索、展示、统计等面向用户的服务和面向机器的开放数据服务。

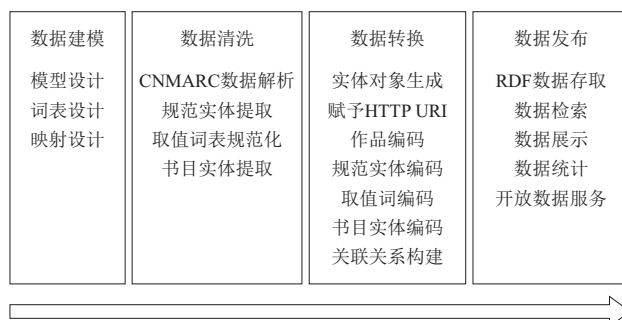


图1 关联书目数据发布流程

2.2 基于BIBFRAME的中文关联书目数据模型

2.2.1 本体设计

本体设计包括模型设计和词表设计。模型设计是通过分析书目数据中的实体关系, 来提炼抽象概念, 定义概念间的关系。如FRBR是最早融入面向对象思想的书目数据模型, 提出“作品”“内容表达”“载体表现”“单件”“人”“机构”“主题”等概念, 并定义相互间的关系。词表设计是用规范的术语词表来表达模型中概念、概念特征和概念间的关系, 概念用类来表示, 概念特征和概念间的关系用属性表示, 需要对每个类和属性进行明确定义。如类的继承关系, 属性所描述的

类,属性的取值范围和约束等,并用机器可理解的形式化语言编码。

国外的关联书目数据项目所采用的书目数据模型主要有EDM、BIBFRAME和SchemaBIBEx。其中,EDM是适用于图书馆、档案馆、博物馆等文化继承机构的关联数据模型,强调数字化对象与实物的关系;BIBFRAME是为关联书目数据设计的双目数据模型,是对FRBR的继承和简化;SchemaBIBEx是基于schema.org的书目扩展。由于BIBFRAME的一个重要目的是替代MARC,在设计过程中与MARC的映射和转换作为重点考虑,发布了一系列MARC21与BIBFRAME的映射。CNMARC虽然与MARC21有诸多不同,但仍可作为参考和借鉴,因此选择BIBFRAME 2.0作为中文关联书目数据的基础框架,并根据CNMARC的特殊性在BIBFRAME 2.0框架下进行扩展。

词表设计是将模型形式化,使人和机器可读,其用规范的术语表示类和属性。本研究所用的词表由82个类和87个属性组成,其中大部分来自BIBFRAME,少量复用自DC、MADS、foaf、Bibo,也有自定义术语(前缀为“shlib”)。

2.2.2 映射设计

自20世纪60年代,MARC逐渐发展成格式复杂、字段众多的数据格式。字段可分为必备字段、特定资源必备字段、有则必备字段和选择使用字段。一方面,编目机构在制定本地著录细则时,会根据需要将部分选择使用字段重新设定为必备、特定资源必备或有则必备字段;另一方面,某些沿用自磁带技术的字段及其取值代码,在映射到本体词表后不再有意义,需要排除在映射表外或规定其取值规则。本研究以CNMARC规则为基础,结合《上海联编中心字段汇总表》,对上海联编中心CNMARC字段的使用情况进行统计分析,明确字段意义,以确定字段映射范围和优先级。

BIBFRAME 2.0将书目数据分为“作品-实例-单件”的核心实体层,每层实体有不同的属性,而每条CNMARC记录均是按照标识块、编码信息块、著录信息块、附注块、款目连接块、相关题名块、主题分析块、知识责任块、国际使用块和国内使用块10个功能块依次展开的层次结构。每个功能块中混有作品、实例或单件的属性和关系。BIBFRAME 2.0同一个核心实体的属性关系可能映射到不同功能块的字段,不同核心实体的属

性关系也可能映射到相同功能块的同一个字段。

(1)作品是被编目资源的概念本质,是抽象的存在。与作品相关的属性或关系包括责任者、主题、语言、类型、体裁等。MARC记录以文献的物质载体为主要对象进行编目,在映射作品的相关属性和关系时,须从物质对象的字段中提取相关信息,包括著录信息块的200字段、附注块的311和313等字段、款目连接块的4XX字段、相关题名块的500等字段、主题分析块的6XX字段以及知识责任块7XX字段。

(2)实例是作品的具体化表达,反映作品多样的表现形式和载体形态。基于文献描述的MARC数据有大量的字段都是与实例有关的属性或关系,包括标识符、版本、出版发行、载体题名、载体附注等。除主题分析块和知识责任块字段外,其他8个功能块都有大量字段映射到实例层的属性和关系。

(3)单件是实例的物理或电子版的单一复本,包括获取条件、借阅政策、访问网址、条码、索书号、流通状态等。支持流通业务的馆藏信息不是本次实验项目关注点,因此单件映射以最简化的方式处理,对索书号和电子资源访问地址进行映射,即国际使用块的856字段和国内使用块的905字段。

在普通图书类CNMARC字段映射中,第一种是字段与BIBFRAME 2.0词汇是一一映射的,即一个字段或子字段只映射一个RDF三元组,如205版本字段映射到BIBFRAME 2.0是“Instance-editionStatement-Literal”这一个三元组。第二种是一对多映射,即一个字段或子字段可映射到多个RDF三元组,如200题名字段,同一个字段需要映射到“Work”和“Instance”两层不同的三元组。第三种是多对一映射,多个字段或子字段可映射到相同的三元组,即含义重复的字段。在这类映射中,需要判断映射优先级,如出版发行时间字段的210\$d\$h与100\$a第9—16位,作品题名500与200等。第四种是特殊的一对多关系,包括两层含义。(1)字段在指示符不同时,含义不同,同一个字段会映射到不同的三元组。如团体责任者字段,当第一指示符是0时,表示团体,映射后三元组的宾语是“Organization”;当第一指示符是1时,则为“Meeting”。(2)定长字段同一栏位根据取值需要映射到不同或多个三元组。如106字段根据取值不同,会有不同的映射三元组,当取值是d,含义是大型印刷本,映射三元组是“Instance-fontSize-FontSize”;当取值是e,表示文献是报纸形式,三元组就是“Instance-genreForm-GenreForm”。

2.3 数据的规范、整合和关联

2.3.1 书目实体组织

基于BIBFRAME 2.0的核心书目实体主要包括作品、实例、单件,明确地识别和表达这3类实体间的关系是关联书目数据发布的关键。

书目实体中最重要的是作品。1847年, Panizzi在为大英博物馆图书馆制定编目规则时,第一次使用“作品”这个术语,并提出“一个读者可能知道他所需要的作品,但不能期待他知道所有的版本; 这些信息他有权利从目录中获取”^[19]。随后OCLC在FRBR中第一次正式定义“作品”,即独有的知识或艺术的创作。自此,作品成为构建层次性的图书馆数据结构和目录的基础。

内容组织的主要难点是作品的认定和作品相关属性的提取。虽然BIBFRAME 2.0的作品内涵包含FRBR的作品和内容表达,但在实践中,某些属性不建议放在作品层,而是作为实例的属性发布; 某些属性既放在作品层又放在实例层。如101\$d提要语种子字段属于FRBR内容表达的属性,在BIBFRAME数据中可以只放在实例层; 7XX责任者字段是针对作品及其内容负责的实体对象,属于FRBR作品和内容表达层,在本研究中主要责任者放在作品层,实例层包含所有的责任者关系。当多条MARC书目记录属于同一个作品时,作品的描述属性须从所有记录中提取。这就涉及属性值的来源问题,研究通过自定义属性“shlibib:source”保留CNMARC记录的唯一标识号以说明属性值的来源。

本研究将作品分为普通图书、汇编文献、年鉴和集刊、丛书4种,在表达作品与实例的关系时,根据文献类型的不同进行不同的处理,尤其是除普通图书外的3种特殊文献。对于无总题名的汇编文献(即一个实例有多个作品),需要分成同一责任者汇编和不同责任者汇编两种情况考虑。对于年鉴和集刊这两类定期连续出版的图书,使用“超级作品”的概念。超级作品是一个没有相应实例的抽象实体,主要起聚类的作用。如《中国历史学年鉴》为超级作品,2000年版《中国历史学年鉴》与《中国历史学年鉴》超级作品的关系用“bf:partOf”来表达。对于丛书,本研究只在丛编项的题名和责任者同时被著录时,才作为作品而赋予HTTP URI发布,便于识别该丛书,并将该丛书的所有子目实例通过这个作品聚集在一起,如果只有题名没有责任者,则只将丛书的题名作为丛书子目实例的属性值,虽然无法

建立关联,但可支持字符串匹配的检索。

2.3.2 规范实体

除作品的认定外,关联书目数据发布的一个重要任务是提取书目数据中的人名、地名、机构团体名称等规范数据,作为对象实体而非字符串。由于上图在数字人文平台建设中,已经建立了基于关联数据的人名规范库、地理名词表、中国历史纪年表等规范库和术语词表,用于家谱、古籍、手稿、档案的人名、地名、朝代的规范控制^[20]。因此,本研究尽量利用和完善已有的规范数据,同时,也利用了中国国家图书馆名称规范档(以下简称“国图规范档”)的数据。将国图规范档中上海联编中心书目数据中的人名融合到上图人名规范库,使上海联编中心书目数据的责任者可直接参引人名规范库中的人物实体,地名可直接参引上图地理名词表,机构团体名称、会议名称为上海联编中心书目数据所独有,另行发布。

2.3.3 取值词表

在关联书目数据中,为改善数据的规范性,避免编目过程中的录入错误,促进数据共享,避免数据冗余,须对一些重要属性的取值进行规范。通过整理属性的常用取值,为每个取值词赋予一个HTTP URI,使其便于在编目中参引,在检索中聚类 and 分面。本研究为上海联编中心的书目数据整理了发行方式、读者对象、文献类型、图表类型、语种5种取值词表。

3 中文关联书目数据发布的技术实现方案研究

3.1 数据清洗、编码、转换

关联书目数据建立在书目本体的基础上,本体中的属性通过定义域(domain)和值域(range)来约束和规范属性的描述主体和取值客体,生成一条“主-谓-宾”结构的RDF三元组。其中主体是一个由HTTP URI唯一标识的实体对象,客体可以是文本型的属性值,也可以是另一个由HTTP URI唯一标识的实体对象,这些实体对象和属性值从书目数据中提取。因此需要对CNMARC数据进行清洗,目的是实现对各类书目实体和规范实体

的提取。首先,要找到对同一个实体对象描述的所有字段;然后,对书目实体(作品-实例-单件)、规范实体(人物、地名)和取值词表进行去重、合并、消歧;最后,提取各种实体并赋予HTTP URI,为属性赋值,与其他实体建立关联并以RDF序列化格式编码,实现从CNMARC到BIBFRAME的转换。

HTTP URI设计模式是关联书目数据发布的基础,一套合理的HTTP URI设计模式有助于实体的识别

和管理。首先,定义关联书目数据发布的根URI(base URI)为“http://bib.library.sh.cn”;其次,选取需要发布HTTP URI的实体及使用空节点的实体,一般原则是,需被其他实体多次重复关联或可在Web上被参引的实体需要发布HTTP URI,只在本地系统内作为某个RDF三元组中谓词的值,并不会作为其他RDF三元组谓词值的实体,用空节点表示;最后,为各类实体设计HTTP URI模式(见表1)。

表 1 HTTP URI设计模式

实体类型	URI 模式	说 明
作品	http://bib.library.sh.cn/resource/work/{随机码}	-
实例	http://bib.library.sh.cn/resource/instance/{随机码}	一条CNMARC书目记录对应一个实例
单件	http://bib.library.sh.cn/resource/item/{随机码}	-
人	http://data.library.sh.cn/entity/person/{随机码}	沿用人名规范库HTTP URI设计模式
机构	http://data.library.sh.cn/entity/organization/{随机码}	包括责任团体和出版机构
会议	http://data.library.sh.cn/entity/meeting/{随机码}	-
地点	http://data.library.sh.cn/entity/place/{随机码}	沿用地理名词表的HTTP URI设计模式
取值词表	http://bib.library.sh.cn/vocab/issuance/{英文缩写}	-

对于人物实体,先将上海联编中心150余万书目数据中涉及的人名(701字段)与国图规范档中155余万人进行匹配,发现有77万余人与国图规范档重合。将77万余人与上图人名规范库进行匹配,发现有7万余人重复,合并去重后得到84万余人。对于7万余重复的数据,直接获取人名规范库中的HTTP URI;对于人名规范库中不存在的人,按照人名规范库人物实体HTTP URI生成规则赋予新的HTTP URI,作为书目数据中相应的责任者属性值。

将人物视作实体而非字符串意味着一个人的信息不仅是一个人名,还包括生卒年、籍贯、生平事件、任职经历、著述、小传、与其他人物的关系等信息。上图人名规范库基于人物本体模型建设(见图2),对于人名规范库中不存在的人,需要根据该本体模型对人物信息进行清洗、编码和转换。国图规范档中有关于人物的简短小传,在小传中有其生卒年、籍贯、人物的各种别名、著述等简单说明,却是一段非结构化的文本,因此数据清洗的主要工作是从小传中提取结构化数据。对于人名规范库中已存在的人,需要进行合并。合并的原则是将生卒年、籍贯等客观唯一的属性,合并去重,只保留一个值;对于生平事件、任职经历、小传等客观不唯一的属性,保留多个值并注明其来源。如茅盾就有来

源于上图名人手稿数据库的小传和来源于国图规范档的小传。

为人物实体赋予HTTP URI,按照本体组织属性、属性值、与其他人物实体间的关系,需要以RDF序列化格式编码,在浏览器上访问“http://data.library.sh.cn/entity/person/t3qypozz7y13mfdt.json”,可查看“茅盾”的RDF/JSON格式数据。

作品提取的关键点在于作品集信息键的构建,即构建责任者/题名信息键^[21]。责任者是对作品负有第一责任的个人或团体,在CNMARC字段中,是所有的701字段或711字段;题名信息是作品的首选信息,取自CNMARC的500字段或245字段,500字段的优先级高于245字段。因此,常规的作品信息键构建公式是{500>245}/{701*+711*}。如果不同书目数据构建的题目信息键取值相同,说明它们是同一个作品的不同实例,从而实现将不同作品的所有实例聚集。

上海联编中心在2017年借助国图规范档实施规范数据与书目数据的挂接工作,在书目数据7XX字段新增\$3记录国图规范档唯一标识号。因此,信息键的责任者部分只需提取\$3参与匹配即可。题名信息包含500/200的\$a、\$h和\$i子字段,题名信息键从3个子字段取值拼接,要保留原始的顺序参与匹配。

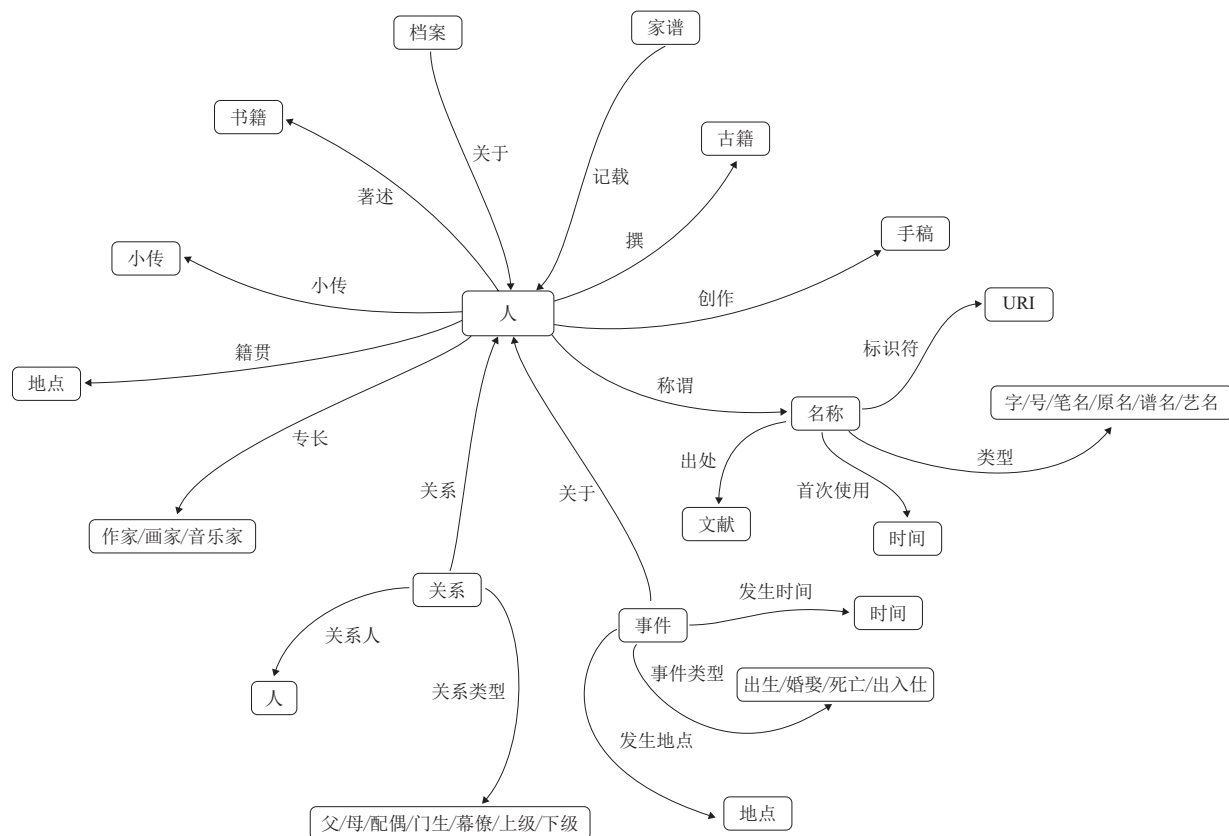


图2 上图人名规范库的本体模型

对于书目实体的提取,本研究先从CNMARC书目记录的相应字段抽取所有作品集信息键,去重合并后生成作品,可能存在多条CNMARC书目记录共同生成一个作品的情况。在为所有的作品赋予HTTP URI的同时,记录每个作品所合并的CNMARC书目记录,即建立一张作品URI与CNMARC唯一书目记录号(039\$b)的映射表。然后,从每条CNMARC书目记录中抽取一个实例实体,通过映射表找到对应的作品实体,用“bf:InstanceOf”属性建立实例与作品的关联。最后,生成实例对应的单件,用“bf:itemOf”建立单件与实例间的关联。

3.2 数据存储、发布、服务

将所有的规范实体、书目实体、取值词表转换为RDF数据,存储在RDF存储库中后在Web上发布,即可为人提供检索、浏览、展示服务,为机器提供开放数据服务。

本文采用专用的RDF存储容器RDF Store而非关系数据库,来存储书目实体、部分规范实体和取值词表

的RDF数据,人名规范库和地理名词表的数据也存储在RDF Store。一方面,RDF Store可直接导入RDF数据,并支持RDF专用查询语言SPARQL对RDF Store中的RDF数据进行查询和更新;另一方面,RDF Store支持Sparql Endpoint向Web开放RDF数据,允许在Web的任何服务器或客户端上编写SPARQL查询语言,查询和获取RDF Store中存储的RDF数据。

如图3所示,书目实体中涉及的人(责任者)、地(出版地)直接关联到其他命名空间中的人名规范库(<http://names.library.sh.cn>)和地理名词表(<http://data.library.sh.cn>);涉及的机构、会议和取值词表则在同一个命名空间中关联。

数据发布和服务层包括为人提供的检索、浏览、展示和书目控制服务以及为机器提供的开放数据服务。

与传统OPAC相比,关联书目数据可提供“作品-实例-单件”分层聚类浏览功能,同时书目数据中涉及各类实体间丰富的关联关系也得到展示。由于作品、实例、单件、人、地、机构等都作为实体,因而可以提供基于概念的检索。根据实体的各种属性检索到该实体,或根据实体间的关系检索到该实体。

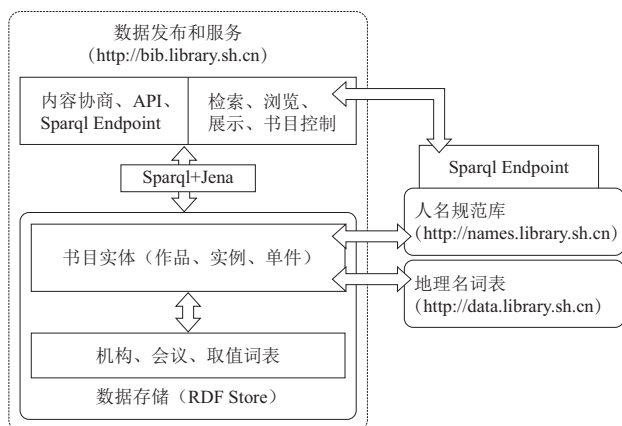


图3 中文关联书目数据存储、发布、服务框架

以人物实体为例,当在责任者检索框输入人名,系统会通过人名规范库中提供的Sparql Endpoint检索到对应的人物实体,得到其HTTP URI,再到书目实体库中检索责任者的值即该HTTP URI的书目实体。由于匹配的是HTTP URI而不是字符串,因而输入人物的任何一个名称得到的书目结果一致。这样,不仅实现了对一个作者所有作品的书目控制,也是对查准率和查全率的改善。

关联书目数据的主要目的是将书目数据从封闭的MARC格式和专业的图书馆领域中释放,超越领域、平台和系统的限制,在Web上为机器提供开放、规范的书目数据,为第三程序提供应用开发接口,鼓励其他领域应用图书馆数据。本研究以三种方式为机器提供开放数据服务:(1)所有实体都有HTTP URI,访问实体的HTTP URI,即可返回关于该实体的RDF数据;(2)通过开放SPARQL Endpoint和本体词表,来提供跨网域的RDF数据查询和获取;(3)提供各种封装的API,为使用不同程序语言的程序员提供方便的数据访问接口。

4 结语

本研究通过将上海联编中心的部分书目数据发布为关联书目数据,完成本体模型、词表和映射的设计,将150万余CNMARC格式的中文书目数据转换为基于BIBFRAME 2.0的RDF数据,重构传统书目的数据模型、结构和格式,在数据清洗过程中发现编目过程的数据问题并进行纠正,改善书目数据的一致性和规范性;发布在Web上,初步实现基于互联网的书目控制和规范控制,实现“作品-实例-单件”三层模型的展示,并以内

容协商、SPARQL Endpoint等方式提供开放数据服务,以促进书目数据在全网域范围内的开放利用,实现互联网时代书目数据的价值,重塑图书馆在网络环境中的形象。但还存在以下问题。

(1)在数据建模方面,BIBFRAME 2.0词表的很多属性词只用了“Used with”说明其建议的定义域,而没有用“rdfs:domain”严格定义。如很多属性既可用于“bf:Work”,也可用于“bf:Instance”“bf:Item”,这为本体的设计带来一定困扰。

(2)关联书目数据的发布先要解决作品的认定问题,然而编目规则在具体实施时的多样性和复杂性,导致CNMARC数据的复杂和不一致,为作品认定带来一定困难,尤其是汇编文献。本研究采用“题名+责任者”为匹配项,来实现作品的去重与合并存在一定误差,有待在未来的工作中进一步完善算法,甚至引入“众包”的模式,允许用户人工认定作品。

(3)在数据编码和转换方面,基于BIBFRAME 2.0的书目数据模型定义大量的“类”。如将变异题名作为“类”而非“属性”来处理,须大量使用空节点,为数据的管理、传输、交换和查询带来问题。

本研究是对中文关联书目数据发布方案的初步探索,同时充分利用上图已有的数字人文平台建设成果,并与之有机结合,互为补充。但由于本文在文献类型上主要是普通图书,将来需进一步拓展到连续出版物,并继续探索针对上述遗留问题的解决方案。

参考文献

- [1] 杨光.网络环境下的书目控制[J].图书馆,2002(5):28-30.
- [2] 柯平.数字目录学——当代目录学的发展方向[J].图书情报知识,2005(3):18-22.
- [3] 顾犇.关于《书目控制未来报告》草案[J].国家图书馆学刊,2008,17(1):76-78.
- [4] YOSHIMURA K S.Analysis of international linked data survey for Implementers[J/OL].D-Lib Magazine,2016,22(7/8).[2017-10-01].http://www.dlib.org/dlib/july16/smith-yoshimura/07smith-yoshimura.html.
- [5] DELIOT C.Publishing the British national bibliography as linked open data[EB/OL].[2017-11-23].www.bl.uk/bibliographic/pdfs/publishing_bnb_as_lod.pdf.
- [6] DELIOT C,WILSON N,COSTABELLO L.The British national bibliography:Who uses our linked data?[EB/OL].[2017-11-23].http://dcevents.dublincore.org/IntConf/dc-2016/paper/download/420/471.

- [7] Retrieving data.bnf.fr data[EB/OL].[2017-11-23].<http://data.bnf.fr/en/semanticweb>.
- [8] SIMON A,WENZ R,MICHEL V,et al.Publishing Bibliographic Records on the Web of Data:Opportunities for the BnF(French National Library)[M]/The Semantic Web:Semantics and Big Data.Springer Berlin Heidelberg,2013.
- [9] Understanding data.bnf.fr data model[EB/OL].[2017-11-23].<http://data.bnf.fr/en/pendata>.
- [10] Stamford Prize for Innovation in Research Libraries(SPIRL)[EB/OL].[2017-11-23].<https://library.stanford.edu/sites/default/files/Bibliothèque%20nationale%20de%20France.pdf>.
- [11] 编目精灵III.LC提供2500万书目记录免费批下载(附LC在Library.Link)[EB/OL].[2017-11-29].<http://catwizard.net/posts/20170721153805.html>.
- [12] BIBFRAME 2.0 Implementation Register[EB/OL].[2017-11-29].<https://www.loc.gov/bibframe/implementation/register.html>.
- [13] 夏翠娟,刘炜,张磊,等.基于书目框架(BIBFRAME)的家谱本体设计[J].图书馆论坛,2014,34(11):5-19.
- [14] 夏翠娟,李颖.互联网环境下韩国、日本书目控制发展动向——“东亚地区书目控制动向国际论坛”述评[J].图书馆论坛,2016,36(4):91-97.
- [15] NAGAI Y,HASHIZUME A,FUKUYAMA J.Use and connect: linked open data of the national diet library,Japan[EB/OL].[2017-11-23].<http://www.ndl.go.jp/jp/aboutus/cooperation/news/2015/dc2015.pdf>.
- [16] メタデータ[EB/OL].[2017-11-23].<http://iss.ndl.go.jp/information/metadata/>.
- [17] 王萍,黄新平.基于关联开放数据的数字文化资源语义融合方法研究——欧洲数字图书馆案例分析[J].图书情报工作,2016,60(12):29-37.
- [18] Europeana Data Model Primer[EB/OL].[2017-11-29].https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf.
- [19] PCC SCS/LDAC Task Group on the Work Entity:Preliminary White Paper[EB/OL].[2017-11-29].<https://www.loc.gov/aba/pcc/documents/PoCo-2017/WorkEntity%20Preliminary%20White%20Paper-2017-09-27.pdf>.
- [20] 夏翠娟.以关联开放数据服务为基础的数字人文平台建设方案研究[J].图书馆学与资讯科学,2017,43(1):47-70.
- [21] HICKEY T B,TOVES J.FRBR Work-Set Algorithm.Version 2.0[EB/OL].[2017-11-22].<http://www.oclc.org/research/activities/past/orprojects/frbralgorithm/2009-08.pdf>.

作者简介

夏翠娟,女,1979年生,硕士,高级工程师,研究方向:元数据、本体、知识组织、关联数据、数字人文,E-mail: cjxia@libnet.sh.cn。
许磊,男,1989年生,助理馆员,研究方向:中文编目、关联书目数据。

Research and Implementation on Chinese Linked Bibliographic Data

XIA CuiJian, XU Lei
(Shanghai Library, Shanghai 20031, China)

Abstract: Based on the literal review of linked bibliographic data research and implementation of national libraries in the world, the paper analyzes the necessity of publishing MARC data as linked bibliographic data on the web. The paper also discusses the main functions of linked bibliographic data. Then it explores the modeling of Chinese linked bibliographic data, the mapping from CNMARC to BIBFRAME 2.0, the methods, process, technologies of content integration, data publishing, open data services for programming, and the difference from OPAC while providing services to users. Finally it summarizes the problems and deficiencies during publishing Chinese linked bibliographic data.

Keywords: Linked Bibliographic Data; CNMARC; BIBFRAME

(收稿日期: 2017-12-02)