中文书目资源关联数据集构建与实现*

常娥[1,2] 牛永骎[1] 孙文佳[1]
[1]东南大学经济管理学院 南京 211189 [2]东南大学图书馆 南京 210096

摘要: [目的/意义] 探讨构建与发布中文馆藏书目资源关联数据集的方法与路径,以推动其朝网络化、语义化和国际化方向发展。[方法/过程] 在阐释以 RDA 和 BibFrame 为引导的国际编目领域全面转型基础上,从中文书目资源关联组织模型设计、CNMARC 元素的自动拆分与转化等 4 个层面深入探讨中文书目资源关联数据化转型中存在的核心问题,并以国鼎图书室书目资源为例,构建发布中文书目资源关联数据集。[结果/结论] 在利用 BCOM 模型进行资源关联组织基础上,提出中文书目资源关联数据集的构建与发布流程。

关键词: 书目资源 关联数据集 数据关联模型 开放获取协议
分类号: G254
DOI: 10.13266/j.issn.0252-3116.2017.19.003

1 背景与意义

当前,万维网正由传统文件网络快速转向各类实体与实体、概念与概念以及实体与概念之间互相连通的数据网络。关联数据被公认为是万维网实现数据化转变的最佳工具之一[1]。截止 2016 年 12 月 30 日,开放数据网络中心(Datahub)已收录数据集 1 142 个[2],形成了以维基百科数据集为核心的高度连通而又自治的数据网络集群,其功能超越了学术研究范畴,覆盖了生活的方方面面,包含地理信息、生命科学、各类文献出版物、词表、元数据、政府信息、演员、导演、电影、音乐以及饭店信息等众多领域数据。

图书情报界纷纷将馆藏书目数据、各种词表、元数据方案等发布成关联数据集,以期在开放数据网络环境中占有一席之地。继美国、英国、瑞典、德国、丹麦等各国国家图书馆馆藏书目关联数据集发布后,世界最大联机编目中心(Online Computer Library Center,OCLC)于 2012 年 6 月宣布将其编目数据宝藏 WorldCat 发布成关联数据集[3]。2016 年,OCLC 和美国国会图书馆(Library of Congress,LC)同时宣布停止对机读目录(Machine Readable Cataloging,MARC)的支持,全面转向对资源描述与检索(Resource Description and Access,RDA)数据的支持[4]。由此可见,图书馆领域书目格式的转型与关联数据化进程已成为不可逆转之时代潮流。图书馆作为各类文献信息的集散地,承载了知识之传承与发展的重要职能。置身于开放数据网络时代,图书馆理应确立“大”资源组织观,成为开放数据网络中的交通枢纽或核心节点。

近年来图书馆领域关联数据集发展迅速,但在开放数据网络中总体占比仍然较少,其中中文图书情报领域关联数据集则更少。相比于 DataHub 中诸如媒体类、社交网络类、地理类和生命科学类等众多领域的开放关联数据集而言,隶属于出版类的图书馆领域关联数据集虽总体出入度排名不低[5],但除了美国国会图书馆主题词(Library of Congress Subject Headings,LCSH)、虚拟国际规范文档(Virtual International Authority File,VIAF)等少数几个数据集被高度链接外[6](链接入度大于 20),大部分数据集只是稀疏关联,链接入度小于 10。总体而言,图书馆领域关联数据集均以特定馆藏资源对象,缺乏整体的数据关联框架,各数据集关联的范围和程度有限,有些仅是各自数据集内部关联,与非图书馆领域数据集的联合度并不高[7]。因此,图书馆距离成为开放数据网络中的交通枢纽或核心节点还存在很大的努力空间。

我国图书情报学界对于关联数据的研究目前仍处于跟踪研究阶段,中文编目资源如何变革与发展,以及关联数据化发布等有关政策及具体实施的报道不多。在开放数据网络时代,置身于国际编目体系全面转型

* 本文系国家社会科学基金青年项目“图书馆资源组织中的数据关联机制研究”(项目编号: 14CTQ005)研究成果之一。
作者简介: 常娥(ORCID: 0000-0001-6865-2118),副研究馆员,博士,硕士生导师,E-mail: chang_e@seu.edu.cn; 牛永骎(ORCID: 0000-0002-8790-7072),硕士研究生; 孙文佳(ORCID: 0000-0002-3601-2623),硕士研究生。
收稿日期: 2017-06-20 修回日期: 2017-08-04 本文起止页码: 22-31 本文责任编辑: 王善军

与创新发展之中，国内图书馆不能一味地等待与观望，需主动从各个层面与角度探索包括各类知识组织工具在内的中文馆藏资源关联数据化构建与发布之方法与策略，积极推动中文编目资源朝网络化、语义化和国际化方向发展。

2 国内外研究现状

图书馆领域数十年积累起来的海量编目数据是不可多得的宝贵财富。将书目资源发布为关联数据集可采用两种方式：其一为直接将 MARC 数据映射为资源描述框架(Resource Description Framework，RDF)格式进行发布，前期瑞典、英国、德国、丹麦等各国国家图书馆发布的书目数据集均属于这一类型。该方法属于元数据格式向 RDF 模式的简单映射，缺乏整体关联组织模型控制，因而导致书目数据集链入度不高，缺乏影响力。其二为在研究构建书目资源关联组织框架基础上，将 MARC 数据转换为 RDF 格式进行发布。该方法不仅可将书目数据发布为关联数据集，而且为馆藏资源数据的整体关联与发现奠定了基础。近年来，新型书目资源关联组织框架成为国内外图书情报领域重点研究内容之一。

以 OCLC 和 LC 为首的国际编目领域成功推行了新的内容标准 RDA 和编码标准书目框架计划(The Bibliographic Framework Initiative，BibFrame)，以支持传统编目数据的关联数据化转型，进而为馆藏资源数据的深度聚合与知识发现提供服务[8-9]。世界各国的编目实践千差万别，制定本地的 RDA 政策声明是实施统一的 RDA 编目之前提。2010 年 10 月，美国国会图书馆率先发布了 RDA 政策声明，并不断更新与完善，成为其他国家制定本地 RDA 政策声明之蓝本[10]。随后，澳大利亚、德国、奥地利、瑞士和英国相继公布了 RDA 政策声明。2015 年 5 月，中国高等教育文献保障系统(China Academic Library and Information System，CALIS)公布了 RDA 政策声明，是目前中国唯一一家发布了 RDA 政策声明的机构，但该声明局限于西文编目范畴。作为一个国际编目内容标准，RDA 包含众多的交替、可选、附加以及省略规则，以适用于全世界所有国家。

无论世界各国的 RDA 政策声明及其实践如何本土化，其核心元素都是以书目记录功能需求(Functional Requirements for Bibliographic Records，FRBR)家族模型(包括 FRBR、规范数据功能需求(Functional Requirements for Authority Data，FRAD)和主题规范数据功能需求(Functional Requirements for Subject Authority Data，FRSAD))作为依托。FRBR 模型的“作品(work)-内容表达(expression)-载体表现(manifestation)-单件(item)”四层结构设计，不仅增加了该模型的复杂性，而且使得普通编目人员难以理解，进而在其推进过程中产生了阻力。有机构和学者认为作为 FRBR 模型忠实实践者的 RDA 元素过于复杂，对其推广和应用前景十分担忧。2011 年 LC 设计并推出了仅包含“作品(work)-实例(instance)”两层结构的新型书目框架 BibFrame，并增加“规范(authority)”和“注释(annotation)”两个核心类对模型功能进行扩展。其中，规范类功能在于对资源描述实施规范控制，注释类功能在于容纳前三项核心类无法包含的描述项，例如包含点评、下载、转引等各种新兴的网络元素，以及传统馆藏信息等。

然而，BibFrame 工作组对于规范类和注释类的设置一直争议不断。2014 年 BibFrame 词表基本稳定后，LC 联合其他机构进行了一年多的实验，经过不断的实践和论证，最终于 2015 年 10 月推出了 BibFrame 2.0 版本，明确取消规范类和注释类，增加单件类(item)，至此奠定了 BibFrame 的“作品(work)-实例(instance)-单件(item)”三层模型结构。与前一版模型相比，BibFrame 2.0 与 FRBR 模型保持了高度一致性，但又不完全相同。对于 BibFrame 的测试、讨论与研究仍在进行着，距离实际应用还会有一段时间[11]。由此可见，书目格式的转型任务非常艰巨，道路十分曲折。但学者们相信 BibFrame 和 RDA 等标准经过不断协调与发展，在经历一个比较漫长而渐进的完善过程中，终将完成取代 MARC 格式的历史重任。

继 FRBR、RDA 和 BibFrame 等标准推出后，近年国内图书馆界针对馆藏资源组织与服务的未来发展进行了持续的理论探索和实践研究。在理论探索方面，国内学者不仅系统引介了国外关联数据的概念、技术框架[12]、发布与消费技术[13-14]、开放应用协议[15]、典型数据集[16-17]，以及新型书目组织框架与新一代编目标准等相关研究[18-19]，而且深刻揭示了万维网时代的规范控制机制[20]，比较分析了 FRBR 与 BibFrame 模型的异同，阐释了 BibFrame 核心类演变的过程和原因[11]。除引介国外新型书目组织框架 FRBR 和 BibFrame 外，国内学者亦经过多方探索构建了具有多维聚合功能的开放式馆藏资源关联组织模型[21-23]。这一系列理论研究不断加深了国内图书馆界对于关联数据、RDA、BibFrame 等技术与标准的理解，进而推动了国内图书馆

利用关联数据重组馆藏资源开展知识服务的理论与实践研究。

在实践应用方面，虽然目前关联数据中枢 Datahub 中收录的中文关联数据集极少，亦未见中国图书馆界发布的书目或知识组织工具关联数据集，但相关实验研究项目已经开展。上海图书馆于 2015 年底发布了家谱关联数据集[24]，家谱知识本体采用了 BibFrame 技术框架进行设计[25]。2016 年 3 月，上海图书馆又首家率先推出家谱关联数据开放平台[26]，其中收录了家谱、中国历史纪年表、地理名词表和机构名录等在内的多种关联数据集，后期将陆续发布各种术语词表、规范文档以及书目关联数据集。该平台提供各种数据的消费接口供研究人员调用，以促进数据的开放获取、共享和重用。

这一系列实践项目标志着我国图书馆界对于关联数据的研究已经进入起步应用阶段。然而遗憾的是，对于与国际标准接轨的书目格式转型研究仍然停留在理论探索、培训交流和跟踪报道层面，鲜有实践研究项目。作为我国图书馆界的“龙头”，国家图书馆一直在密切关注 RDA，但还未发布 RDA 政策声明，RDA 也只有在西文编目中有涉及，且多为套录数据，国内进行 RDA 原编的数量微乎其微[10]。经过 RDA 翻译工作组多方努力，目前已成功出版中文版《资源描述与检索(RDA)》[27]，为后续研究与实践工作奠定了基础。

3 中文书目资源关联数据化转型之核心问题

3.1 中文书目资源关联组织模型设计

以 MARC 为代表的传统封闭式、一维线性的书目组织格式已无法适应开放数据时代的馆藏资源组织需求，如何设计融合关联数据技术的新型馆藏资源组织框架是国内外图书情报领域的核心研究内容之一。目前，国外图书馆界重点推出了 FRBR 家族模型和 BibFrame 模型。经过多年的讨论、实验测试与协调，这两个模型已经从最初的竞争与分化状态，转变为统一与融合的状态，主要原因在于 BibFrame 2.0 版本在该模型核心类部分做出的重要调整，彻底转变了 BibFrame 模型的思考模式与关注重点，淡化了 MARC 格式对于 BibFrame 词表的影响。论文《BibFrame 核心类演变分析》[11] 深入阐释了 BibFrame 模型取消了规范类和注释类，并增设了单件类的历史背景与原因，笔者十分赞同专家的观点，本文不再赘述。

在实际关联数据网络中，链接的除了真实世界对象实体外，还包括概念、观点、原理、事实等各种数据类资源。无论是 FRBR 家族模型，还是 BibFrame 模型均为完全数据化的细粒度馆藏资源组织预留了发展空间，只不过当前研究重点停留在书目级的馆藏资源组织与描述上，以致力于解决传统 MARC 书目格式的转型问题。所以，现有 BibFrame 原编数据中的资源更多的是指向某本书这一实体对象，书的名称或标签只作为书的属性之一，书中知识内容采用属性标签进行粗粒度的揭示与描述，例如 bf: Classification 和 bf: Topic。如何将书中所有知识点抽取出来，进而进行描述与组织，是未来馆藏资源组织框架的研究重点。能否在 FRBR 或 BibFrame 框架中进行扩展，如何扩展，还是脱离 FRBR 或 BibFrame 模型重新构建组织模型，重新构建后，如何与 FRBR 或 BibFrame 模型进行对接等一系列问题都十分值得探索。以 FRBR 和 BibFrame 为代表的新型馆藏资源组织模型无疑属于关联数据和知识本体范畴。在本体模型的设计与研究中，概念和实例的区分与选择是非常困难与纠结的，一直是本体研究的重要问题，至今未见合适的选判标准。在实际的本体应用中，往往依赖具体应用，来判断某个对象究竟是概念还是实例。

3.2 CNMARC 元素的自动拆分与转化

早在关联数据技术提出之初，新一代编目标准 RDA 还未正式推出以前，瑞典国家图书馆于 2008 年率先将瑞典联合目录发布为关联数据集，随后英国、德国和丹麦等国家均发布了国家馆藏书目关联数据集。目前，虽然在 DataHub 中收录有 RDA、BibFrame 等词表关联数据集[28]，但图书馆领域已发布的书目关联数据集绝大多数仍是未经 FRBR 或 BibFrame 建模的 MARC 数据的直接 RDF 化格式转换。虽然 RDA 和 BibFrame 还需要经历一个漫长而渐进的完善过程，距离世界范围内的全面实践与应用还有一段时间，但图书馆界领域长期积累起来的编目资源是不可多得的宝贵知识财富，不可能舍弃。所以，将历史编目数据转变为 FRBR 或 BibFrame 建模的关联数据，抑或是直接发布，将 MARC 格式转换成 RDF 格式，是不可逾越的研究任务，亦是非常关键的一步。

国外，RDA 官方站点(http://www.rda-jsc.org/，http://www.rdatoolkit.org/) 和 BibFrame 官方站点(http://www.bc.gov/BibFrame) 均已推出 MARC21 数据转换为 RDA 或 BibFrame 格式的工具。由于中国机读目录格式(China Machine-Readable Catalogue，CNMARC) 与 MARC21 两者在内容方面基本一致，但在信息资源

划分、字段指示符赋值、子字段元素设置及赋值方面存在差异，这使得在两者之间建立映射既存在可能性，同时又面临挑战。国内学者已就关联数据中 CNMARC 到 MARC21 的映射以及 CNMARC 到 RDF 的映射问题进行了研究[29-30]，并借助国外已有的 MARC21 数据转换工具，着手研制 CNMARC 数据转换为 BibFrame 的工具和平台[25]。作为国内首家成功推出关联数据技术开放数据的图书馆案例，上海图书馆在其家谱关联数据集发布过程中，认为将已有的家谱书目元数据映射和转换成 BibFrame 框架下的 RDF 数据是最大的困难，同时还需要克服数据不一致的问题。

在中文馆藏资源关联数据集构建与发布的过程中，针对历史编目数据的处理，主要涉及编目各元素项的自动拆分和转换映射两个环节。对于第一个环节，由于编目数据采用了元数据编码方式，其本身具有非常清晰的逻辑结构，借助软件开发工具，将其分解成单个的元素项完全可实现，难度不大。对于第二个环节，主要是将拆分打撒后的独立元数据项一一转换映射成新型馆藏资源关联组织模型的类和属性关系。在历史编目元素项与新型馆藏资源关联组织模型转换与映射的过程中，会存在无法对应的情形，需要对原有的编目元素项进行修改、增补或删除等操作。无论是 MARC 格式的转换，还是 CNMARC 格式的转换都遵循这样的步骤，即需要完全解构并按照新型馆藏资源关联组织框架重组历史编目元数据项。

鉴于国外图书馆界研究先行一步，已完成了 MARC21 格式到 RDA 和 BibFrame 格式的转换与映射，因此国内在研究 CNMARC 数据如何进行转换映射的过程中，可以直接按照新型书目组织框架进行转换与映射，对于有争议而不太清晰的元数据项的转换，可以参考 MARC21 的转换方法，但无需将所有 CNMARC 转换成 MARC21，然后再映射成 RDA 或 BibFrame 格式。

3.3 书目资源 URI 的确定与生成

统一资源标识符(Uniform Resource Identifier，URI)和资源描述框架 RDF 是关联数据的两大核心技术。然而，学者们的研究重点往往集中在关联数据 RDF 三元组及其关联模型的构建与生成上，对于资源 URI 标识的确定与生成的讨论不够深入。URI 是包含关联数据在内的互联网领域的重要基础技术之一，是在网络虚拟空间标识和定位事物的基本方式。URI 是统一资源定位符(Uniform Resource Location，URL)的上位概念，它不仅可以标识真实世界中的网页、声音、动画、文件、视频等各种实体事物，还可以标识概念、名称、事件、术语、时间等各种虚拟事物，统一将它们作为网络资源进行标识，进而使其可以被管理、存储、跟踪和调用。

如果从 URI 组成来看，它是由一组按照特定语法规范构成的字符串，提供了一种简单且可扩展的标识网络资源的方法[31]。由于 URI 地址在网络世界中的统一性与唯一性，经其标识的资源地址，不仅可获得关于被标识资源的有用信息，而且可关联到一组其他相关资源，因为同一机构在发布网络数据时会采用统一的 URI 设计原则和模式。法国、英国、美国以及澳大利亚等各国政府和图书馆在构建发布关联数据集时，在遵循 Cool URI 的基本原则和统一模式基础上[32]，从域名结构、构成模式、命名约定等不同角度，制定了各自的 URI 设计原则和模式[33-35]。总体来看，国外 URI 设计模式主要包含了数据集 URI、本体 URI、词汇集 URI 等不同类型。上海图书馆在遵循无变量、稳定性、使用 HTTP URI、可读性和国际化的设计原则基础上，制定了 URI 设计模式，包含数据集 URI、本体 URI、规范词表 URI、取值词表 URI、非信息资源 URI 和信息资源 URI 这 6 种类型[36]。

由于 URI 的使用是全网域的，被标识资源种类繁多，信息资源和非信息资源、实体资源和概念资源等交叉融合，因此从形式上对不同类型资源的 URI 标识进行规范，以保证 URI 标识的永久性、稳定性和唯一性是十分迫切需要的。但 URI 的确立与生成研究仅止于此还不够，还需要针对某种馆藏资源关联组织模型，更加细致深入地探讨 URI 标识选判标准，原因在于：在馆藏资源关联组织模型中，各类型资源数据是融合的。换言之，实体数据和概念数据共存，URI 链接的资源可能是真实世界的对象，也可能是概念类的各种名称，因此需要判断何为实体资源的名称，何为概念类的名称。这一问题看似简单，实际上非常容易引发纠结，属于前文所讨论的本体模型中的概念与实例的区分问题。对于实体资源的名称，宜将其作为资源数据的属性值加以表达，无需赋以 URI 标识，而对于概念类的名称，则适合作为概念资源加以描述，需要赋以 URI 标识。

有鉴于此，深入研究 RDF 三元组中资源项的选判标准，进而按照不同资源类型的 URI 设计规范生成 URI 标识串，是馆藏资源关联数据集构建中不容忽视的研究任务。以当前重点讨论的 BibFrame 和 FRBR 模型为例，在“作品”“实例”“单件”“内容表达”“载体表现”类中均设有 URI，其中“作品”代表的是独特的知识或艺术创造，“内容表达”代表的是作品知识以字母、

数字、音乐或舞蹈、声音、图像、实物或移动等形式及其组合来表达的方式，两者的 URI 指向的都是一种虚拟的概念类资源，而“载体表现”和“实例”虽然代表的是作品内容呈现的物理体现，但由于代表作品物理实体样例或实例的“单件”类的存在，所以“载体表现”和“实例”的 URI 指向的仍是一种虚拟的概念类资源，而“单件”的 URI 指向的则是实体馆藏资源，即图书馆中存在的可得见、摸得着的实体文献资源。

无论是 BibFrame 模型，还是 FRBR 模型，“单件”除了指实体馆藏资源，从书目框架模型的整体架构来看，它还包含了馆藏文献的复本资源概念，即一模一样的两本书。且不论古籍文献资源，由于现代书籍完全是工业化产品，因此在其出版印刷过程中，会存在大量的完全一样的书籍，两本或多本完全一样的书籍会同时被某个图书馆购买并拥有。传统 MARC 文献编目中，一般采用索书号区分复本资源以方便排架和查找，但如果转向 RDA 编目，这一问题该如何处理呢？若采用相同的 URI 标识“单件”则无法区分复本资源，若采用不同 URI 标识“单件”，则又意味着本来一致的复本资源变成了不同的资源。由于工业化产品存在大量复制品的原因，这一问题除了存在于馆藏资源关联组织领域外，在其他领域亦会出现，具有普遍性，亟待解决。但对于人或机构的关联化描述，则不会存在类似问题，因为即使是同卵双胞胎也会被看作为是两个不同的个体，由于克隆技术是被禁止的，所以克隆人不在考虑范围内，由于机构是由人组成的，所以两个完全相同的机构也不会存在。

此外，虽然 BibFrame 和 FRBR 模型都考虑到了书籍版本问题，但由于该问题非常复杂，存在着“同一种文献”“同一版本文献”以及“同一种文献的不同版本”等各种不同版本情形[53]，如何在 RDA 编目数据中进行区分以融合同一种书的各种不同版本或进行文献版本关系发现，都有待进一步深入研究，其中书目资源 URI 的确立与生成是关键之所在。

3.4 内容标准与开放获取协议

馆藏资源数据一旦在数据网络中公开，不可避免地将涉及知识产权问题，图书馆如何制定开放数据许可协议以促进馆藏资源数据的有效传播与利用是亟待解决的问题。图书馆界作为开放数据运动的拥护者，在开放数据网络中贡献了数以亿计的 RDF 三元组以及各种组合关系。为了更好地让开放关联数据服务于公众，除美国、英国等各国政府部分制定了详细的开放数据获取协议外，以德国国家图书馆、大英图书馆和欧洲数位图书馆等为代表的国际图书馆领域均制定了各自的开放数据获取协议。

国内学者对于开放数据领域的研究较多，但对于开放数据许可协议的研究则较少，主要集中在国外各种开放数据获取协议的定义、内容说明、基本特点以及不同协议间的区别上，如知识共享(Creative Commons，CC)、开放数据共用(Open Data Commons，ODC)、开放政府许可协议(Open Government License，OGL)等常用协议的介绍和比较，但大多未进行深入探讨，尤其是针对国内图书馆界在开放数据许可协议的制定与使用方面所呈现的独特性研究不够系统和深入[54]。另外，目前国内图书馆界发布的开放关联数据集屈指可数，直接导致了开放数据的获取协议和内容标准关注度不够。

知识共享家族协议是目前参引最多的开放数据许可协议，此类协议的核心在于明确了数据提供者与使用者之间的平衡关系，同时明确了数据的共享范围、使用权限等问题，并考虑了数据隐私问题。在数据提供者与使用者之间的平衡关系中，重点考虑的是数据版权冲突与收益问题。当数据的冠名、收益或其他相关知识产权发生时，图书馆需要考虑如何减少版权冲突和避免损失。一般而言，图书馆不支持带有商业性质的数据拷贝、复制、编辑加工等一系列行为，而对非商业性的研究则可以进行数据拷贝与复制。开放数据许可协议不具备法律效力，而是基于使用者自觉遵守的规则约定。因此，除了制定开放数据许可协议外，图书馆还会考虑通过只支持浏览、禁止复制等技术手段减少版权冲突，对于科研目的的数据使用，则通过实名认证申请下载的方式来进行使用。

4 中文书目资源关联数据集构建与发布实证——以国鼎图书室藏书为例

4.1 中文书目资源关联数据集的发布流程

综上所述，新型中文书目资源关联组织框架、CNMARC 元素自动拆分与转换，以及书目资源 URI 确定与生成等是将中文书目资源发布为关联数据集的核心之所在。本研究在国家社科基金资助下，在 FRBR/FRAD/FRSAD 三者集成与扩展的综合概念框架之上，通过引入包含研究背景、材料方法、模型假设、实验数据、结果讨论等学术元数据框架的基础上，构建了由“资源层-中间层-知识层”三层架构的图书馆资源底层通用的整体数据关联模型(Bottom Common Organization Model of the Whole Library Knowledge Resource，

BCOM)[7]。以 BCOM 模型为基础，根据中文书目资源发布为关联数据集所遵循的一般步骤和方法，笔者尝试提出中文书目资源关联数据集构建与发布流程，具体如图 1 所示：

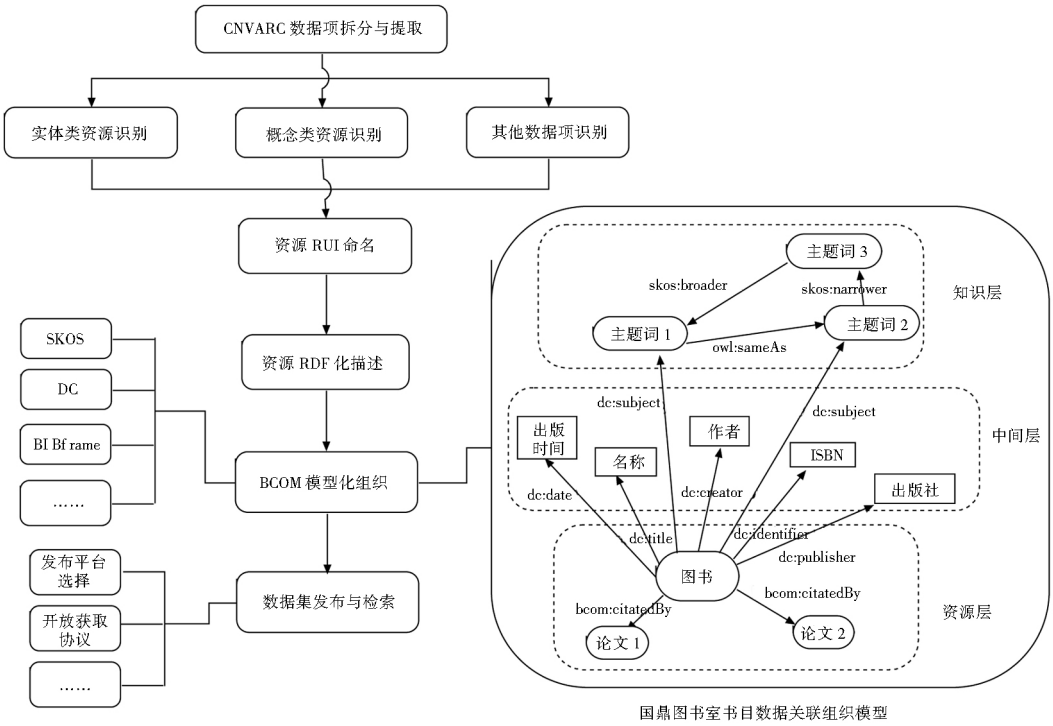

图 1　中文书目资源关联数据集发布流程

首先，CNMARC 数据项拆分与提取。这一步关键任务包含两个方面：其一是将特定格式的 CNMARC 数据转换为易于人工解读的 EXCEL 或文本数据，便于后期数据清洗和整理，一般借助自动 CNMARC 字段分析程序完成；其二是将拆分所得元数据项分为实体类资源、概念类资源和其他元数据三类。

实体类资源指具有物理实体依托的资源类，对应于 BCOM 模型的资源层，是实际存在的图书、论文、人、机构等；概念类资源指具有实质含义的知识概念名词，一般没有实体依托，对应于 BCOM 模型的知识层，一般由 CNMARC 数据的关键词元数据项获得；其他元数据项指除前两类资源项之外的书目元数据项，对应于 BCOM 模型的中间层，包括名称、出版时间、ISBN 号等。

其次，资源 URI 命名。考虑到在 BCOM 模型中资源具有不同的类型，在制定资源 URI 生成模式时，除了使其具有永久性、稳定性和唯一性外，还需要增加 URI 的可识别性，即在 URI 域名构成中加入特定区分字段，以区分实体类资源和概念类资源。此外，针对图书资源而言，图书复本采用同一 URI 标识，而同一种图书的不同版本则采用不同 URI 标识。

再次，资源 RDF 化描述与 BCOM 模型化组织。这一步的任务可分解为两个层面：一是采用“资源-属性-值”方式对实体资源进行属性描述，以完成 BCOM 模型中间层映射，并联通资源层和知识层；二是利用“资源-关系-资源”方式对实体资源和概念资源分别进行关联组织，以完成 BCOM 模型资源层和知识层映射。由于 BCOM 模型旨在提供馆藏资源底层数据整体关联的通用框架，因此诸如属性元素选择、资源与概念关系构建等模型的具体化还有赖于实际应用。

考虑到书目数据集与其他数据集的融合性，可优先选择已注册的各种词汇表进行资源属性与关系描述，例如 DC、SKOS、OWL、BibFrame 等。此外，为了更好地揭示学术论文的主题内容，BCOM 模型设计了包含研究背景、材料方法、模型假设、实验数据和结果讨论等在内的学术元数据，并且在资源层利用文献间的引用关系对实体资源进行拓展。通过查阅关联开放词表(Linked Open Vocabularies，LOV)网站发现目前文献间的引用关系以及学术元数据还无相关注册词汇表可借鉴，因此本研究注册发布了 BCOM 模型词汇表，其中引文关系定义为 bcom: citedBy。

最后，数据集发布与检索。目前创建与发布关联数据集主要有静态发布、批量存储发布和调用时生成这 3 种方式。考虑到中文书目数据量较大且存在更新

问题，建议选择调用时生成 RDF 文件这一较为灵活的发布方式，常用工具包括 Drpual、D2R、Virtuoso、Triple 等工具。开放获取协议方面，建议参引知识共享家族协议，同时采用禁止复制、支持实名认证下载等技术手段保护数据版权。

4.2 国鼎图书室书目关联数据集的构建与发布

李国鼎先生(1910.1.28-2001.5.31)长期从事台湾经济发展工作，贡献卓著，被誉为“台湾经济快速发展的建筑师”和“台湾科技之父”。东南大学图书馆专设国鼎图书室以收藏国鼎先生生平所著所有图书和文稿，其他学人研究国鼎先生所产生的各种资料，包括照片、视频等，以及经济方面的台版期刊、图书等。国鼎图书室珍藏丰富，无论是对研究国鼎先生本人，还是研究台湾经济与社会问题，都具有非常重要的文献价值。

目前，国鼎图书室仅对部分图书资料进行了编目与揭示，主要包括民国时期台湾金融方面的统计资料、政策资料、李国鼎本人所著专著，以及后人撰写的国鼎传记类资料等，其他文献资料未做加工整理。囿于传统资源编目技术的樊篱，已入编的图书资料仅限于校内用户查阅和使用，这极大影响了国鼎图书室特藏资源的传播与利用，降低了其在业界的影响力。笔者以在编国鼎图书室书目资源为基础，按照上文所述中文书目资源数据集构建与发布流程，发布了国鼎图书室书目资源关联数据集，以促进国鼎图书室特藏资源的传播与利用。

本研究借助 MySQL 数据库，主要创建了图书表(book)、论文表(paper)和概念表(concept)这 3 个资源类表，利用 D2R 平台发布了国鼎图书室书目资源数据集。在数据发布平台选择方面，考虑到后期数据量较大且存在更新问题，因而选择了支持调用时生成 RDF 文件的 D2R 平台。原因在于，相比于 Drpual、Virtuoso 和 Triple 等工具，D2R 提供了独立而专门的映射语言及表达规则，可支持关系数据库显性或隐性数据关联关系的提取和表达，从而完成复杂关系结构的灵活映射，因此其 RDF 语义映射和转换能力更为突出[39]。

为了充分利用 BCOM 模型对国鼎图书室书目资源进行关联化组织，笔者利用文献间的引用关系对资源类进行拓展，即利用 CNKI 数据库查找出国鼎图书室图书资料被引用的信息，并记录施引论文信息，构建 paper 资源表，从而丰富了资源层数据。由于图书资源是本文研究重点，而非论文资源，因此中间层不包含研究背景、材料方法等学术元数据描述，主要借助书名、作者、出版时间和关键词等 6 项书目元数据对图书资源进行描述。知识层主要由图书资源的主题概念组成，重点揭示主题概念间各种逻辑关系，包括等级、等同、相关和并列这 4 种基本类型。国鼎图书室书目资源关联组织模型详见图 1。为了区分实体资源和概念资源，笔者根据上海图书馆提出的 URI 设计原则，制定 URI 设计模式，且在 URI 域名中加入特定区分字以区分不同资源类，即用“…/Object/…”标识实体类资源，用“…/Concept/…”标识概念类资源。

值得注意的是，作者和出版社可简化为图书类资源的属性值进行简单揭示，也可以作为资源类，放在 BCOM 模型资源层进行详细描述。例如，作者类资源可通过名称、出生年龄、性别、职业等进行描述；出版社可通过名称、地点、创办时间等进行描述。笔者采用了第一种方案，即将作者和出版社简化为资源属性值，主要考虑到国鼎图书馆藏书范围有限，涉及的作者和出版社有限，因此进行了简化处理。

在词汇表选择方面，考虑到国鼎图书室书目关联数据集与其他数据集的融合性，优先选择了已注册的各种属性词汇表进行资源属性描述，例如，在 BCOM 模型中间层尽可能采用 DC 词汇表进行描述，在知识层则采用 SKOS 和 OWL 词汇表进行描述。

4.3 国鼎图书室关联数据集的检索与利用

国鼎图书室书目资源数据集包括图书资源 139 个，论文资源 217 个，概念资源 119 个，RDF 三元组共 1 260 个，发布首页面如图 2 所示。该数据集以 BCOM 模型为基本关联组织框架，因此图书、论文和概念资源表三者相互连通，可以选择任意资源类表作为浏览检索起点。选择图书类资源，点击某本书的超链接可看到该书详细的 RDF 化描述信息，以及被其他文献引用和包含主题概念的情况。选择论文或主题概念类资源，则可以了解某论文引用了国鼎图书室中的哪些图书或某主题概念存在于国鼎图书室的哪些图书之中。

例如，以图书类资源表为检索入口，点击图 2 中的 book 链接，然后选择点击链接 http://localhost:2020/resource/book/100，可获得该书详细的 RDF 描述信息(见图 3)。由图 3 可知，该书由李国鼎先生所著，书名为《台湾经济高速发展的经验》，由东南大学出版社于 1993 年出版，并且包含两个主题概念，被 22 篇其他文献所引用。

点击相关引文链接即可进一步获得某施引文献引用了国鼎图书室的哪些图书资料。例如，点击图 3 中的引文链接 http://localhost:2020/resource/paper/126(《试析 1949 年后国民党的治台政策及其变迁》，林震

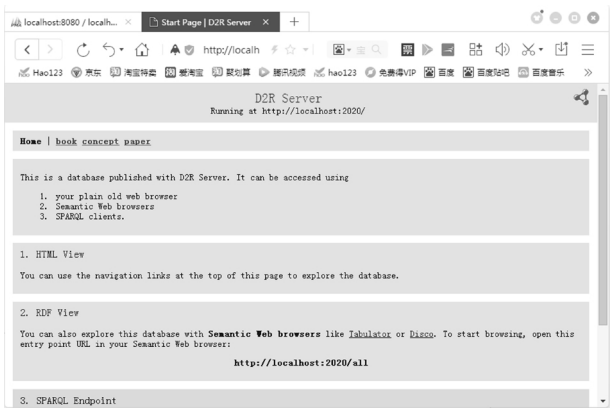

图 2　国鼎图书室特藏数据集发布页面

图 3　图书类资源示例

著)，发现该文献不仅引用了《台湾经济高速发展的经验》(http://localhost:2020/resource/book/100)，而且引用了国鼎图书室的另外两本著作，分别为《台湾的经济计划及其实施》(http://localhost:2020/resource/book/86)和《台湾经济发展背后的政策演变:修订本》(http://localhost:2020/resource/book/95)，详见图 4。

图 4　关联引文资源示例

点击相关主题概念链接即可进一步获得该主题概念存在于国鼎图书室的哪些图书之中。例如，点击图 3 中的概念链接 http://localhost:2020/resource/concept/48(经济发展)，发现该主题概念不仅存在于《台湾经济高速发展的经验》(http://localhost:2020/resource/book/100)一书中，还存在于《二〇〇六年海峡两岸经济科技发展趋势研讨会纪要》(http://localhost:2020/resource/book/24)、《加工出口区与经济发展》(http://localhost:2020/resource/book/33)、《经济政策与经济发展》(http://localhost:2020/resource/book/43)、《台湾的现代农业》(http://localhost:2020/resource/book/92)等 7 本著作中。如图 5 所示:

图 5　图书主题概念示例

D2R 平台不仅支持数据集的浏览检索，亦支持 SPARQL 查询检索。限于篇幅，本文仅以图书资源作为浏览检索入口，展示了国鼎图书室书目资源整体关联组织与利用状况，其他资源类检索入口和 SPARQL 查询将不再赘述。

5　结语

置身于开放数据网络与国际编目改革浪潮中，我国图书馆界需紧跟国外实践研究步伐，积极探索中文编目资源关联数据集构建与发布之方法，以推动其朝网络化、语义化和国际化方向发展。本文在讨论中文书目资源关联模型设计、CNMARC 数据自动拆分与提取、书目资源 URI 确定以及内容标准与开放获取协议的基础上，以国鼎图书室特藏资源为例，探索了构建与发布中文馆藏书目资源关联数据集的方法和路径。囿于研究时间和精力，本文着重探索了国鼎特藏室中图书类资源及其相关概念类资源的关联组织问题，今后的研究工作将以此为基础进一步拓展和整合更多的资源类，包括作者类、出版类、时间类、地点类等相关类资源。此外，国鼎特藏室除了包含图书外，还包含期刊、论文、照片、手绘稿、图片、音像等各种形式的资源，如何将所有形式的特藏资源及其相关概念统一进行关联

数据化加工，以构建出更加丰富多元的国鼎特藏关联数据集亦是本研究未来之工作重点。

参考文献：

[1] BIZER C. The emerging web of linked data [J]. IEEE intelligent system, 2009, 24(5): 87-92.
[2] Datahub [EB/OL]. [2017-06-10]. https://datahub.io/dataset?tags=lod.
[3] 胡小菁. WorldCat 也关联数据了 [EB/OL]. [2017-06-10]. http://catwizard.net/posts/20120621214628.html.
[4] 王景侠. 书目格式的关联数据化发展及其启示：从 MARC 到 BIBFRAME [J]. 图书馆杂志, 2016(9): 50-56.
[5] 贾君枝，寇蕾蕾. 关联数据云图中出版类数据集特点分析 [J]. 国家图书馆学刊, 2016(1): 59-68.
[6] Linked data cloud [EB/OL]. [2017-06-10]. http://lod-cloud.net/.
[7] 常娥，华苏永. 馆藏资源底层通用整体数据关联模型研究 [J]. 图书馆论坛, 2016(8): 7-12.
[8] About RDA [EB/OL]. [2017-08-03]. http://www.oclc.org/en/rda/about.html.
[9] Bibliographic Framework [EB/OL]. [2017-08-03]. http://www.loc.gov/BibFrame/.
[10] 梁红，姜化林，涂颖哲. CALIS、美、英、澳 RDA 的政策声明比较分析 [J]. 大图图书馆学报, 2016(1): 25-34.
[11] 胡小菁. BIBFRAM 核心类演变分析 [J]. 中国图书馆学报, 2016(5): 20-26.
[12] 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): 5-11.
[13] 夏翠娟，刘炜，赵亮. 关联数据的发布技术及其实现 [J]. 中国图书馆学报, 2012(1): 49-57.
[14] 夏翠娟，刘炜. 关联数据的消费技术及其实现 [J]. 大学图书馆学报, 2013(3): 29-37.
[15] 张春景，刘炜，夏翠娟. 关联数据开放应用协议 [J]. 中国图书馆学报, 2012(1): 43-48.
[16] 赵蕊菡. 政府类开放关联数据集调查研究 [J]. 图书与情报, 2016(4): 102-112.
[17] 涂志芳，吴丹. 医学相关领域开放关联数据集调查研究 [J]. 图书情报工作, 2015, 59(9): 14-23.
[18] 刘炜，胡小菁，钱国富，等. RDA 与关联数据 [J]. 中国图书馆学报, 2012(1): 34-42.
[19] 胡小菁. RDA：从内容标准到元数据标准 [J]. 图书馆论坛, 2014(7): 1-7.
[20] 刘炜，张春景，夏翠娟. 万维网时代的规范控制 [J]. 中国图书馆学报, 2015(3): 22-23.
[21] 欧石燕. 面向关联数据的语义数字图书馆资源描述与组织框架设计与实现 [J]. 中国图书馆学报, 2012(11): 58-71.
[22] 王忠义，周杰，黄京. 数字图书馆多粒度关联数据的创建与发布 [J]. 情报学报, 2016(8): 885-896.
[23] 夏立新，陈晨，王忠义. 基于多维度聚合的网络资源知识发现框架研究 [J]. 情报科学, 2016(5): 3-8.
[24] 黄启哲. 上海推出“家谱知识服务平台” [N]. 文汇报, 2016-02-18(9).
[25] 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(11): 5-19.
[26] 上海图书馆开放数据平台 [EB/OL]. [2017-06-10]. http://data.library.sh.cn/.
[27] RDA 发展联合指导委员会. 资源描述与检索(RDA) [M]. RDA 翻译工作组，译. 北京：国家图书馆出版社, 2014.
[28] HILLMANN D, COYLE K, PHIPPS J. RDA vocabularies: process, outcome, use [J/OL]. [2017-01-16]. http://www.dlib.org/dlib/january10/hillmann/01hillmann.html.
[29] 贾君枝，白林林. 关联数据中 CNMARC 到 MARC21 的映射实现 [J]. 国家图书馆学刊, 2015(4): 80-93.
[30] 白林林，贾君枝. 关联数据中 CNMARC 到 RDF 的映射实现 [J]. 国家图书馆学刊, 2015(4): 94-102.
[31] Uniform Resource Identifier(URI): generic syntax [EB/OL]. [2017-06-10]. http://www.ietf.org/rfc/rfc3986.txt.
[32] SAUERMANN L, CYGANIAK R. Cool URIs for the semantic web [EB/OL]. [2017-06-10]. http://www.w3.org/TR/cooluris/.
[33] WILLIAMS S. URI patterns [EB/OL]. [2017-06-10]. http://github.com/UKGovLD/URI-patterns-core/blob/master/URI%20Patterns.md#reference.URISetsV1.
[34] Australian Government Linked Data Working Group. URI guidelines for publishing linked datasets on data.gov.auv0.1 [EB/OL]. [2017-06-10]. http://github.com/AGLDWG/TR/wiki/URI-Guidelines-for-publishing-linked-datasets-on-data.gov.au-v0.1.
[35] British Library URI patterns [EB/OL]. [2017-06-10]. http://www.bl.uk/bibliographic/pdfs/british_library_uri_patterns.pdf.
[36] 许磊，夏翠娟，刘炜，等. 关联数据 URI 设计规范探讨 [J]. 国家图书馆学刊, 2016(5): 22-32.
[37] 梁美宏，曾建勋. 基于书目关联数据的文献版本关系发现研究 [J]. 图书情报工作, 2016, 60(5): 123-130.
[38] 杨敏，夏翠娟，徐华博. 开放数据许可协议及其在图书馆领域的应用 [J]. 图书馆论坛, 2016(6): 91-98.
[39] 白海燕，梁冰. 利用 D2R 实现关系数据库与关联数据的语义模式映射 [J]. 现代图书情报技术, 2011(7): 1-7.

作者贡献说明：

常娥：构建论文总体框架，论文的写作与修改；

牛永骥：数据收集与处理；

孙文佳：数据收集与处理。

Design and Implementation of Linked data Based on Chinese Bibliographic Resources

Chang E[1,2]　Niu Yongqin[1]　Sun Wenjia[1]

[1] School of Economics & Management, Southeast University, Nanjing 211189

[2] Southeast University Library, Nanjing 210096

Abstract: [Purpose/significance] Libraries in China should actively research the design and implementation of linked data based on Chinese bibliographic resources in order to promote its network, semantization and internationalization. [Method/process] The paper discussed the reform of the international cataloging field led by RDA and BibFrame. Then, it analyzed the core problem of the construction of the linked data of Chinese bibliographic resources from four aspects such as data association model design, CNMARC elements transformation etc. Taking the bibliographic resources of GuoDing Library as an example, a Chinese bibliographic resource linked data set was published. [Result/conclusion] The paper proposes the process of the design of linked data based on Chinese bibliographic resources with the bottom common organization model (BCOM) of the whole library knowledge resource.

Keywords: bibliographic resource　linked data set　data association model　open access protocol