DOI:10.13530/j.cnki.jlis.190022

文献编目:从数字化到数据化

胡小菁

摘　要　近十年来,文献编目领域从理论模型、标准规范到实践应用,均发生了自机读目录问世以来的最大变化。这个变化与关联数据技术的应用直接相关,可以概括为从数字化到数据化,也就是书目数据由机器可读走向机器可操作,进而融入互联网全球数据库。在此过程中,编目界经历了观念上的重要变更(从记录到数据),厘清了混淆的概念(实体及其名称与描述),重新对书目数据建模,并展开了一系列实践。其中,作为应用的重要组成部分,数据基础设施在数据化中发挥着重要作用。图 1。参考文献 43。

关键词　文献编目　关联数据　数字化　数据化

分类号　G254

Cataloging from Digitization to Datafication

HU Xiaojing

ABSTRACT

In the past decade, the great change has taken place in the field of cataloging from theoretical models and standards to applications since the invention of Machine Readable Cataloging (MARC). This change is directly related to linked data technology and can be summarized as cataloging from digitization to datafication, i.e., bibliographic data from machine-readable to machine-actionable for integrating into the Web. Cataloging community experienced important changes in concepts (from records to data), clarified confused concepts (entities and their names and descriptions), re-modeled bibliographic data, and engaged in various experiments and programs.

First, the focus of cataloging transforms from records to data. In the theoretical model, IFLA paid attention to "the basic level national bibliographic record" in Functional Requirements for Bibliographic Records (FRBR). But Functional Requirements for Authority Data (FRAD) "focuses on data, regardless of how it may be packaged". A recordless environment is gradually being formed. In cataloging rules, Resource Description & Access (RDA) emphasized the core elements, but the new RDA (Toolkit Beta Site) abandons the core elements. In metadata format, BIBFRAME and RDA vocabularies clearly identify different data which are confused in MARC.

Second, concepts between entities and their names and descriptions are clearly distinguished. IFLA Library Reference Model (LRM) defines Nomen as an entity. Authority control becomes entity management and no longer relies on the uniform form of a name. To distinguish between entities (Real World Objects) and their descriptions (such as authority records), MARC21 adds new subfield $1 that records the identity

通信作者:胡小菁,Email:xjhu@library.ecnu.edu.cn.ORCID:0000-0002-1703-9724(Correspondence should be addressed to HU Xiaojing,Email:xjhu@library.ecnu.edu.cn.ORCID:0000-0002-1703-9724)

of the entity itself.

Third, data are modeled as RDF vocabularies. Different vocabularies have different classes and properties. Although BIBFRAME vocabulary and RDA vocabulary are very different in class or entity identification, BIBFRAME can use with RDA as a content standard.

Finally, datafication is in practice. Library of Congress (LC)'s Bibliographic Framework Initiative is in its final stage after several rounds of pilots. The Swedish National Library launched LibrisXL in June 2018, which is the first linked data system to replace the core cataloging model in integrated library system. Consumption of existing linked data is a very important part of practice. Casalini, LC's Providers vocabulary, and the LD4 Community Working Group on Reconciliation are all committed to this. MARC 21's new 758 field is also for this. At the same time, the library community participates in the construction of linked data infrastructure to facilitate data sharing and consumption. Programs include LC's linked data service (id.loc.gov), Virtual International Authority File (VIAF), RDA Reference Vocabulary and LD4's extension of BIBFRAME vocabulary, etc. 1fig. 43 refs.

KEY WORDS

Cataloging. Linked data. Digitization. Datafication.

进入21世纪,以美国国会图书馆副馆长Marcum在2004年所作报告“编目的未来”[1]为重要标志,编目工作发生了自机读目录(MARC)问世以来最大的变化,笔者曾总结为四大发展趋势,即编目格式简单化、编目外包普遍化、数据来源多样化及联合目录本地化[2]。十年后回头再看,这四个方面仍反映国际编目工作的发展方向,联合目录本地化更是以基于云的图书馆自动化系统的方式产生着显著影响。但是,这个总结也存在一个根本性的不足:没有触及编目工作最底层的书目数据格式。编目的彻底变革必然涉及书目数据格式,这需要理论指导、标准支撑。而在当时,本世纪最重要的编目标准规范都还没有发布:或尚在研制中,如《资源描述与检索》(RDA);或还未酝酿,如《国际图联图书馆参考模型》(IFLA-LRM,简称LRM)和书目框架(BIBFRAME)。书目数据格式的变革,则以2002年Tennant“MARC必须死”一文为象征[3],而真正的标志则是2008年瑞典国家图书馆率先以关联数据发布其联合目录Libris[4]。随后前述编目领域的指导性文件和主要标准陆续问世,编目理念正经历着根本性变化,与之相应的编目工作必将随之而变。这种变化,可以概括为从数字化到数据化。

数字化是广为大家熟知的概念,指将信息转为计算机可读的数字格式的过程[5]。图书馆界从1960年代就开始了图书馆目录的计算机可读过程,其成果机读目录(MARC)成为信息数字化的先行者,MARC格式用于书目信息及相关的规范信息等,至今已逾半个世纪。数据化即把信息变为数据,“指一种把现象转变为可制表分析的量化形式的过程”[6]。数据化实践有很长历史,但此概念直到2013年才由迈尔-舍恩伯格和库克耶在《大数据时代》一书中总结提出,所谓“可制表”指对信息进行结构化标识,让数据从不可能的地方提取出来。如《大数据时代》所述,“数据代表着对某件事物的描述,数据可以记录、分析和重组”[6]。机读目录作为结构化数据,既是数字化成果,也天然具有数据化属性,不仅记录书目、规范等信息,也可用以进行一定的量化分析。但因为受MARC格式限制,书目信息以“记录”形式作为一个整体存在,机读目录中的单个“数据”不具有重组能力,因而离真正的数据化还有不小的距离。

为摆脱MARC格式的束缚,国际图书馆学界在新兴的互联网技术中选择了关联数据,力

图使图书馆的书目数据由机器可读(machine-readable)走向机器可操作或可执行(machine-actionable),从而融入互联网全球数据库。关联数据2006年由Tim Berners-Lee提出,是一种在万维网(Web)上发布结构化数据的方法,以便通过语义查询将其相互链接并变得更有用,不仅为人类读者提供网页阅读,而且可由计算机通过自动读取方式共享信息[7]。关联数据是适合书目信息数据化的技术,其在书目信息或编目领域的应用,经历了转换现有书目数据、以关联数据形式发布的最初尝试,最终确定从底层书目数据制作开始改变。虽然前行过程中也曾伴随一些质疑甚至否定,但目前图书馆编目领域的关联数据应用经历了开发、试验,正面临实践推广的最后阶段,书目信息的数据化成为可见的发展方向。

1 记录与数据

一直以来,图书馆目录基本以条目为单位,一件在编文献(如一本或一套书)对应一个条目或书目著录(Bibliographic Description),在MARC格式中就是一条"记录"。MARC(2709格式)以记录为基本单位,记录是不可分割的整体,一条记录中保存书目数据内容的"数据区",只在有"头标区"以及"目次区"配合使用的情况下才有意义。对于书目信息,数据化的要求,就是由"记录"细化到记录中包含的"数据"或书目数据,此处书目数据指"描述并提供对书目资源检索的数据元素"[8]。

1.1 从记录到数据

数据(或数据元素)原本存在于记录中。在编目领域,关注点从记录到数据的转变,发生在21世纪的前十年,在国际图联(IFLA)"功能需求"系列概念模型的发展过程中,可以清晰地看到这种变化。1997年发布系列第一种《书目记录的功能需求》(FRBR)后,IFLA在1999年和2005年先后准备编制《规范记录的功能需求与编号》(FRANAR)和《主题规范记录的功能需求》(FRSAR),从题名可知,3个模型的中心都是"记录"。但此后确定模型将"着重于数据本身,不注重如何打包数据"[9]为记录,最终2个报告在发布时题名分别改为《规范数据的功能需求》(FRAD)和《主题规范数据的功能需求》(FRSAD)。

实际应用中,新ISSN门户采用基于BIBFRAME的应用配置文件,其发布的关联数据以"ISSN资源"为中心,而不是围绕"ISSN记录"[10],也是着重点从记录到数据的体现。国际性的合作编目项目(PCC)在《2018—2021年战略方向》中更直接地将应用关联数据后的状况称为"无记录环境"[11]。

1.2 记录从完整到可选

从关注记录到关注数据的这种变化,也体现在编目标准最近的更新上。一直以来,编目条例、著录规则都强调记录的完整性,著录详简级次也是评价书目数据质量的重要指标,如《中国文献编目规则》[12]和《英美编目条例(第2版)》(AACR2)都规定了由简到详的第一、二、三级著录,就是典型例证。FRBR的研发初衷之一,是在因经济压力需要降低编目成本的环境下,为合作编目提出基本或核心级记录标准[13],也就是"基本级国家书目记录"。2011年发布的《国际标准书目著录》(ISBD)统一版,依据FRBR的基本级国家书目记录,在"ISBD概要"部分标记了各著录单元的必备状态[14]。

然而,2017年发布成为IFLA标准的《国际图联图书馆参考模型》(LRM),不再有与FRBR中"国家书目记录基本需求"对应的内容,并声称"尽管实体作品、内容表达、载体表现和单项之间的结构关系是模型的核心,但在实施中不需要模型中声明的这些属性和其他关系。假设特定应用中因不需要而省略某些属性或关系,生成的系统仍可被视为IFLA LRM的实现"[15]。

最新编目规则RDA在2010年发布时仍延续传统,强调核心元素,也就是一条记录中的必

备元素。然而,自 2017 年起历经一年多修订于 2018 年发布的测试版(被称为“新 RDA”)中,RDA 遵循 LRM、放弃核心元素概念,也就是说在规则中所有元素都是可选的,这是“数据化”在编目规则中的首次体现。对新 RDA 来说,它只关心每一个数据,甚至数据形式也不强求统一,可以用不同方式表达,它规定了具有普适性的 4 种元素记录方法,被称为 4 路径(4-fold path),即记录非结构化描述、结构化描述、标识符和 IRI(国际资源标识符)①。至于如何选择记录方法、如何集成数据形成一条“记录”,则由数据制作机构通过应用配置文件来决定。

1.3 数据从混同到明确

传统编目规则在关注记录的同时,也注重对数据元素给予充分揭示。但在具体应用中,有时会受制于元数据格式而不能得到明确标识。以公认数据元素最丰富的 MARC 格式为例,在编目规则中明确的某些数据元素,有可能会混在同一个子字段中不予区分,典型的如“其它形态细节”即 MARC21 的 300 $b 子字段或 UNIMARC 的 215 $c 子字段,包含图书的插图及其颜色,影音资料的色彩、播放速度、凹槽特征、有声无声、声道、放映格式,实物的材质等诸多不同特征的数据。

数据化要求明确区分不同数据元素,给予不同标识,在元数据格式中必须予以体现。为此,RDA 依据规则为上述特征注册了不同的元素,BIBFRAME 词表也采用了不同的类和属性,比如实物不同部位所用材料分别为 baseMaterial(基底材料)、appliedMaterial(应用材料)和 mount(底座)等。

上述从关注记录到关注数据的变化,都契合应用关联数据的背景。可以说,国际上编目相关的模型、规则和格式,伴随着近年图书馆相关领域关联数据应用的探索,在 21 世纪的第 2 个十年,已经为数据化升级做好了准备。

2 实体及其名称与描述

书目信息中,绝大部分要作独立处理的数据元素,在书目记录中都存在,因而其识别与标记原本并不是问题,或至少不是很大的问题。无论是书目数据模型 LRM 还是编目规则 RDA,虽然整体上与其前身 FRBR 或 AACR2 相比变化很大,但一旦深入到具体的元素,规定做法大多没有太大变化,完全保留原有做法的也不在少数,因而也曾有“新瓶装旧酒”之讥。实际上这是因为更新模型与规则的主要目的并非改变书目信息的内容,而是改变书目信息的表示与处理方式,尽管内容的改变有时也不可避免。

数字化(如书目记录)的主要目的是供人类阅读使用,而数据化的目的在于让机器处理。当细化到机器处理层面,原来隐藏在书目记录中被忽视的“名”与“实”问题,此时就浮上了水面,需要加以解决。这些问题涉及名称与实体的表示,明确区分实体及其名称、实体及其描述、实体描述以及对描述的描述,是数据化过程中的重要关注点。

2.1 名称与实体

所谓实体指任何事物(Thing)。人们通常以名称指代实体,如“鲁迅”指原名为周树人的作家、“北京”指中国的首都。但名称与实体不能混为一谈,比如我认识“鲁迅”(这两个字),但不认识鲁迅(这个人)。通常可以用引号括起表示名称,但也可能不加标识,人类在阅读时也能够区分名称和实体,但要计算机处理则必须有不同的标识,明确告诉机器哪个是名称、哪个是实体,才不会造成机器解析错误。

① IRI 是统一资源标识符 URI 的一般化、可使用非 ASCII 字符。本文按关联数据领域的通常习惯,以 URI 代指 URI 和 IRI(引用 RDA 时除外)。

1990 年代 IFLA 为书目数据建模时(FRBR),没有考虑名称问题,只确定了书目、责任、主题三类实体。21 世纪为规范数据建模时(FRAD),由于需要定义规范记录的对象与其所用名称,区分实体及其多种不同名称成为不可回避的问题,因而提出了名称(name)、标识符和受控检索点等 3 个新实体。其后在对主题规范数据建模时(FRSAD),进一步将主题的名称抽象为一个单独的实体,为与 name 作出区别,采用拉丁文 nomen 表示。功能需求模型统一版 LRM 在确定模型时,没有选择把名称作为实体的一个属性,而是沿用 FRSAD 的做法,把名称(nomen,或译为"命名")作为普遍适用的高层实体,与其他实体并列[16],如图 1[15]17上部所示。

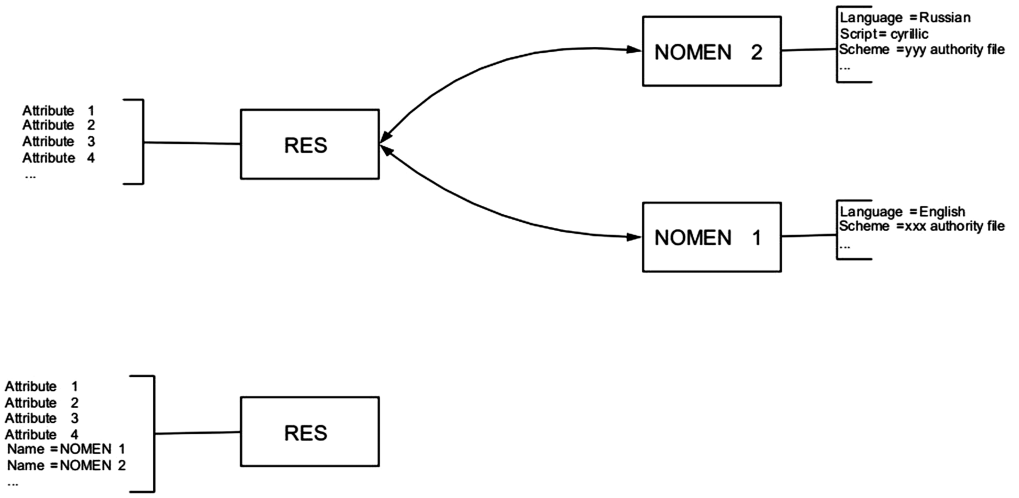

图 1　Nomen 的两种实体—关系模型

对于"名称"使用属性还是实体,LRM 的选择是考虑到在实体—关系模型中,无法声明属性间关系,如果名称作为属性(图 1 下部),则无法表达多个名称间关系,也无法对名称本身的特性作进一步揭示。LRM 以 nomen 作为实体(图 1 上部),可以为其定义语言、文字、来源等属性。LRM 将 nomen 与其他实体之间的关系定义为"称谓"(appellation),另外还定义了 nomen 与 nomen 之间的 3 种关系,即:等同(两个名称是同一资源的称谓)、部分(一个名称取自另一名称的一部分)和派生(一个名称来自另一名称)。

以往编目规则中绝大部分是描述元素,要求提供检索点的情况极其有限,仅题名、责任者、分类主题等;对文献的出版负有责任的出版者,一般也不作为检索点[12]333。定义 nomen 为实体后,可以使描述与检索点不再那么截然不同,检索点有可能扩大到所有元素,新 RDA 就是这么做的[17]。新 RDA 根据 LRM 新增 nomen 元素,其定义是:指代 RDA 实体的标识,包括名称、题名、检索点、标识符、分类号和主题标目。

2.2　字符串与标识符:名称与实体的标识

无论名称还是实体,都需要以一定方式加以标识。关联数据建立在资源描述框架(RDF)之上,在 RDF 中名称以文字(literal)或字符串(string)来表达,而实体或对象作为"资源",以 URI 标识。形式上,URI 直接引用,而字符串则以引号括起表示。

在编目规则中,描述或著录的是字符串(含实体名称),检索的是资源(实体或对象)。由于传统编目规则对名称与实体没有明确区分,实体也以字符串表达,只不过对其形式有受控要求,即唯一的规范标目形式或称首选检索点等。

在数据化环境中,规范标目形式也不过是一种字符串,不具有特殊性,只有 URI 才是实体作为资源的标识。由于 URI 主要供机器使用,直接显示可能对人类不太友好,因此可以为资源定义一个专用属性,作为供人阅读或关键词检索的标签,如 BIBFRAME 的 preferedTitle(首选题名)。通过应用关联数据,编目界的规范控制可以由从不同的名称形式中选择唯一的规范形式,走向接受不同机构采用不同"首选形式"的实体管理[18],如虚拟国际规范档(VIAF)采用的方式。

在 1.2 节所述新 RDA 的 4 路径中,非结构化描述、结构化描述和标识符都属名称,IRI 则表达实体。某些元素可选择所有 4 种方法,如 agent(施事者);其他元素则可能有若干方法"不适用"的情况,比如新增的载体表现说明(Manifestation Statements)系列元素,由于其本身性质为客观照录或转录,只能选择非结构化描述,采用被描述资源上的语言形式。

2.3 真实世界对象及其描述

从资源描述框架(RDF)角度理解,任何谈论的事物(实体)都是"资源"。资源有两种,一种是信息资源,可以电子方式传递;一种是非信息资源,不能以电子方式传递,包括真实世界对象或抽象概念、想象中的东西。书目信息中绝大部分是非信息资源,除了分类、主题中有一些概念或想象中的东西外,其他基本上属于真实世界对象(RWO)。

图书馆的规范记录,包含资源(实体)的信息,如个人的姓名、生卒年、使用语言等,作品的题名、创作者、创作年代等,也包含制作规范记录的信息,如编制机构、记录编制与更新日期等。前者是对资源本身的描述,后者是对描述的描述,作为混合记录,本质上是对资源的描述。但在实践中,规范记录曾被当作资源本身的代表,如同 2.1 节所述的名称与实体不分。如 MARC21 的控制子字段 $0 规范记录控制号,曾被作为实体的标识符。在关联数据应用之后,这个问题被发现,美国国会图书馆(LC)关联数据因而为 RWO 与规范记录设置了不同的 URI,在规范记录中,RWO 显示为"附加信息"(Additional Information)。

为容纳真实世界对象(RWO),MARC21 在 2017 年新增控制子字段 $1,以区别于记录规范记录控制号和标准号的子字段 $0[19]。尽管随着关联数据在编目实践中应用加速,MARC21 预计会逐渐被 BIBFRAME 取代,但 MARC21 与关联数据相关的修订从未停止。这类修订,可视为 MARC21 的数据化改造。MARC21 持续修订的主要原因,一是在转向 BIBFRAME 的过程中还需要多次进行与 MARC21 之间的双向转换,二是预计在很长时期内还会有机构继续使用 MARC21。

3 数据化建模

编目相关元数据标准可分为 4 类,即数据结构标准、数据内容标准、数据取值标准和数据交换标准[20]。MARC 或都柏林核心等属于结构标准,编目规则属于内容标准,分类法、主题词表、名称规范档等属于取值标准,2709 格式或 XML 等属于交换标准。在 RDF 中,结构标准和取值标准都被称为词表(Vocabulary),后者也称取值词表;前者使用 RDFS(RDF 模式)定义时也称 RDF 词表,使用 OWL(Web 本体语言)定义时也称本体,但有时也并不严格区分。本文对结构标准通称"RDF 词表"或简称"词表",取值标准称"取值词表"。

在从记录到数据的变革中,数据内容标准变化是一方面,数据结构标准转变是相伴的另一方面。如前引《大数据时代》所称,"数据代表着对某件事物的描述"[6]。应用关联数据,首先要确定模型和词表的设计,即确定如何描述事物,其中最基本的就是确定需要哪些类(Class),以及与之相应的属性(Property)是用对象属性还是数据类型属性。

3.1 类与属性

关联数据中,数据的最小单元即语句(State-

ment),包含对资源的一个基本描述或称 SPO 三元组,即主体、谓词、客体。需要定义主体与客体的类型或类,以及谓词或属性,从而构成词表,用于描述资源。

如 FRAD 最终报告所称,"将某一事物定义为一项属性还是一个独立的实体是设计概念模型的关键"[9]7。2.1 节所引 LRM 对于 nomen 作为实体的考虑,也涉及建模时对属性与关系的选择。

RDA 本属编目规则,但对其定义的元素提供配套的 RDA 注册(https://www.rdaregistry.info/),期望可以作为关联数据词表使用(以下称 RDA 词表)。RDA 遵循 LRM 概念模型,RDA 词表也基本沿用 LRM 的 11 个类,只在集体施事者下增加了 2 个下位类:家族和团体。与此同时,RDA 词表定义了一千多个属性,属性与类的数量完全不成比例。这是相对传统的做法。

新兴词表通常会定义更多的类,如美国国会图书馆(LC)编制的 BIBFRAME。BIBFRAME 模型与词表在 2014 年最初发布时设置了 54 个类、270 多个属性[21],属性与类数量比约为 5:1。2016 年发布 2.0 版,对先前版本做了大刀阔斧的修订,类增加到 175 个,属性则减少至不足 200 个。虽然也有少量类重新定义为属性,但扩展定义了大量的类,包括属性重新定义为类和增加与属性对应的类[22],这使得属性与类的数量比几乎接近 1∶1。BIBFRAME 的这种变化,是基于如下认识,即对于机器处理而言,用类相比属性有 3 个优点:可重用、易查询、可柔性降级[23]。

PCC 的子项目"合作连续出版物项目"(CONSER),在完成 RDA 元素到 BIBFRAME 映射表后提交报告,进一步提出从机器可执行性角度,希望 BIBFRAME 能够将连续出版物的首末期编号和年代也定义为 URI[24]。未来或许 BIBFRAME 还会进一步增加类的数量。

在实践上,开发中的图书馆服务平台 Folio,有一个元数据模型 Codex[25],用来集成图书馆收藏的各种资源的异构元数据,目前只设计了 5 个对象(类),即实例、单件、馆藏包(Package)、馆藏地和收录范围(Coverage)。其中把"收录范围"设计为类,这是在馆藏层面与 CONSER 上述建议相似的设计。

3.2 对象属性与数据类型属性

RDF 三元组即"主体 谓词 客体",其中谓词即属性,基本形式可以举例如下:

1) <作品 A><作者><作者 URI>

2) <载体表现 A><题名>"ABC"

两者的不同在于前者的客体为"对象"(使用 URI),后者的客体为字符串(用引号括起)。两种属性分别称为对象属性和数据类型属性。在实体—关系模型如 LRM 中,前者为实体间关系,后者为实体的属性或特性(Attribute)。定义对象属性或关系,也意味着需要定义更多的类。但在实际应用中,会有一定的灵活性。

以 BIBFRAME 为例,对象属性原本要求客体为"对象"即属于某个类,但 BIBFRAME 2.0 没有对所有对象属性都做出严格的定义。BIBFRAME 2.0 没有采用严格定义的定义域(Domain)和值域(Range),而是分别采用"用于"和"期望值"对主体和客体的使用做出规定。这样做的理由是更灵活,既不必因需要用于多个类而构造专有类,同时也方便使用其他命名空间中定义的资源[26]。

以 RDA 为例,RDA 认为 LRM 允许属性和关系互换使用,这支持 RDA 对所有元素采用 4 种记录方法,即非结构化、结构化描述、标识符和 IRI,其中前 3 个用于属性,最后 1 个用于关系。"这些发展将扩展 RDA 覆盖的元数据抓取场景的范围,从印刷或手写目录卡片,到机器转录、规范控制、关系和关联数据"[27]。

如 3.1 节所见,BIBFRAME 与 RDA 词表在类或实体认定上有很大差异。不过这并不影响 BIBFRAME 与作为内容标准的 RDA 共用。事实上 LC 编目用的 BIBFRAME 编辑器中,各录入项的提示标签就是 RDA 元素或条款号[28]。对使用 RDA 规则的编目员来说,无需了解 RDA 元素

到底被定义为类还是属性、对象属性还是数据类型属性,只需按本地应用配置文件的要求,在4种记录方法中选择使用一种或多种即可,三元组的表达由计算机软件在后台处理。

4 数据化实践

以书目模型、编目规则、RDF词表为代表的元数据标准规范,为图书馆目录的数据化实践做了很好的准备。与此同时甚至更早,关联数据探索也在图书馆目录、规范库的关联数据发布上率先进行。自2008年瑞典全国联合目录Libris以关联数据形式发布以来,OCLC、匈牙利国家图书馆等不少机构采用不同词表分别进行了发布尝试。另外LC的id.loc.gov、OCLC的WorldCat作品、虚拟国际规范档(VIAF)、RDA取值词表等,更多地作为数据基础设施,与其他外部数据源如DBPedia、Wikidata等一起,为近年书目信息的关联数据实践打下应用与数据基础。

4.1 编目试验与实施

在转换现有图书馆目录发布为关联数据之后,先行者们开始考虑从源头开始,将整个编目转向基于关联数据的工作流程。经过近十年探索,欧美一些国家的关联数据编目工作流程的实施已经或即将开启。

美国国会图书馆(LC)2011年启动"书目框架转变行动",准备放弃MARC格式。此后LC持续进行BIBFRAME开发[29],开展了多轮试验。①2012—2013年由多个图书馆参与早期实验,2013年发布BIBFRAME词表。②2015—2016年第1阶段试验(Pilot 1),采用BIBFRAME编辑器作为编目界面,实时查询LC规范库,用BIBFRAME词表取代MARC进行原始编目;2016年发布BIBFRAME词表2.0版,2017年把LC所有馆藏目录和规范目录由MARC转换至BIBFRAME2.0,作为编目来源库。③2017—2018年第2阶段试验,采用更新后的BIBFRAME编辑器,增加参加试验的编目员,在真实的BIBFRAME编目环境中对多种文献类型进行原始编目。④2019年正进行扩展试验,解决由MARC切换到BIBFRAME编目的各种问题,其中包括进行BIBFRAME到MARC的反向转换[30],为的是在LC完全转向BIBFRAME编目后,可以继续为那些暂时没有转向BIBFRAME的图书馆提供MARC数据。

在2008年率先发布全国联合目录为关联数据的瑞典国家图书馆(KB),继续走在实施前列。2012年,KB启动基于关联数据的LibrisXL项目,用以取代基于MARC21的Voyager图书馆自动化系统的核心编目部分。LibrisXL系统于2018年6月启用,采用基于BIBFRAME 2.0的KBV词表,成为首个实际使用的基于关联数据的系统[31]。

美国高校配合BIBFRAME开发,开展了多项关联数据研究,影响较大的有加州大学戴维斯校区的"编目再发明:未来图书馆运作模式",2017年发布最终报告《BIBFLOW路标》[32]。项目进行了转变编目工作流程的试验,涉及套录和原编,特别针对连续出版物和规范记录。最终报告总结了从MARC生态系统中的关联数据(为MARC批插入URI)到原生关联数据生态系统(关联数据原生编目)的二阶段路径。

另一持续时间更长的是LD4系列(https://www.ld4l.org/),共4个项目,由安德鲁·梅隆基金提供资助,承担机构为斯坦福、哈佛、康奈尔等美国著名高校。2014—2016年首个项目LD4L(图书馆关联数据)为初步探索。2016—2018年两个项目LD4P(关联数据生产)和LD4P-Lab(关联数据生产实验室)进入实践,包括开发工具、探索工作流程、针对不同特藏资源对BIBFRAME词表进行扩展。2018—2020年项目LD4P 2副标题"实施之路",终级目标是实现以关联数据来描述图书馆资源,已招募十多所图书馆作为合伙人,与PCC、LC协作,准备从2019年4月开始在云编辑环境中生成关联数据[33]。

关联数据实施离不开厂商、尤其是图书馆自动化系统和书目数据供应商的支持。LC 在每年的美国图书馆协会冬夏两次年会上,都会举办 BIBFRAME 更新论坛,邀请包括厂商在内的实践者分享他们与 BIBFRAME 相关的工作,其中最富有成效的厂商当属意大利编目外包商 Casalini Libri 和图书馆自动化系统商@ Cult。它们是欧美很多 BIBFRAME 项目背后的支持厂商,与十多家北美研究图书馆合作开发的 SHARE VDE 环境,以外部来源 URI 强化 MARC 记录(记录于子字段 $0 或 $1),将 1 亿条书目和规范记录转换为 BIBFRAME 2.0 后发布,并支持数据自动更新[34]。LD4P 2 项目也采用 SHARE VDE 环境,作为图书馆合作生成关联数据的支撑平台。

4.2 关联数据消费

在批评 MARC 格式封闭时,曾有人把 MARC 数据库比喻为与互联网隔离的"谷仓"。随着越来越多的机构尝试以关联数据发布信息,这些采用词表各异的书目数据,是否形成了新的相互隔离的谷仓再度成为话题。不同词表间映射还相对简单,只需做一次,问题更大的是数据本身。书目信息如 MARC 记录中的数据,如果只是转换为字符串,仍然是供人阅读、至多供人检索,只有作为对象、有 URI,才能供机器利用、具有关联性、可重用。如果只是单纯地发布为关联数据,从让图书馆目录(OPAC)更容易为用户获取角度看,其效果未必好于做过搜索引擎优化(SEO)的基于 MARC 记录的 OPAC。无论从用户体验还是数据关联角度,消费已有数据是关联数据实践中相当重要的工作,这就需要进行相同实体判定,找到匹配的已有关联数据实体,采用或关联其 URI,即所谓的发布对象的调和与解析(Reconciliation & Resolution)。

实际上,这也是实用的关联数据项目的基础,前述各种关联数据应用及研究均关注并致力于处理这个问题。如 Casalini 介绍其 BIBFRAME 项目,以 MARC 数据转换为 BIBFRAME 发布,其第一个步骤即用 URI 强化 MARC 记录,作为项目的中心任务,需要做实体识别、调和和数据强化[35]。

LC 在 BIBFRAME 第 2 阶段试验中关注到 MARC 记录中出版社缺少规范控制问题,重复出现率颇高的出版社(在 BIBFRAME 中属于"提供者"),在转换为 BIBFRAME 后造成大量匿名资源、空节点,既无法重用,也影响未来系统运行效率。LC 因此试验建立"提供者词表"为其提供 URI,从而改善转换后数据的可重用性,提高系统性能[36]。这是在缺少 URI 资源的情况下,自建关联数据基础设施并加以利用的举措。

基于安德鲁·梅隆基金系列课题的 LD4 社群,在 2017 年 5 月成立了一个开放参与的调和工作组(LD4 Community Working Group on Reconciliation),致力于解决这个问题。工作组目标是为文化遗产机构(GLAM 即美术馆、图书馆、档案馆、博物馆)的资源元数据的调和与解析,总结匹配算法、工作流程、工具和功能需求[37]。工作组认为调和与解析相关工作,在实际应用中可能不只是匹配一种情况,因此公开征集用例、功能需求、当前服务、工作流程等,截至 2017 年底共征集并评估了 51 个用例,比如:#76跨语言匹配,#70 运行优化,#68 使用本体,#39断言两个实体不同。

为配合 URI 强化 MARC 记录的工作,MARC21 格式进行了数据化改造。除 2.3 节介绍的更新和增加控制子字段 $0 和 $1,还有一个变化是新增 758 字段资源标识符,用于书目记录描述的资源(即载体表现对应的作品或内容表达的 URI)及相关资源(包括相关作品、内容表达和载体表现等)。"MARC 修订建议 2017-09"称,扩大 MARC 支持关联数据标识符,是图书馆将关联数据构建到元数据生产过程的迁移路径的一部分;将书目描述与适当的作品或其他资源实体相关联,是将 MARC 书目数据转换为关联数据的主要挑战;记录 MARC 记录中相关资源的 URI,使得当 MARC 记录转换为关联数据表示时,URI 可以作为对象包含在 RDF 语句中;在无

URI 可用的情况下,非 URI 标识符对于正确的数据调和(实体解析)仍然是有价值的[38]。

4.3 数据基础设施建设

新技术的应用推广说到底是一个经济问题,而不是"与时俱进"的理念问题。一项技术,如果不能实质性地改善图书馆的运营,如提高工作效率或拓展新的服务,是无法得到广泛应用的。当年 MARC 的成功,就基于细致的成本分析[39]。

如 4.2 节所述,消费已有关联数据资源是应用中的重要工作,也是充分体现关联数据应用价值、费用效益的良好途径。这些促进数据共享和消费的数字基础设施,就是数据基础设施[40]。如同水电煤等公共基础设施,数据基础设施是数据经济运行所必需的服务和设施。图书馆界需要借助已经发布的资源,消费其他领域提供的关联开放数据(Linked Open Data,LOD),同时本身应该并且也可以有所贡献,参与到建设关联数据基础设施中去。双赢才能有持续性。

图书馆界的早期实践已经发布了一些具有 URI 的取值词表。本领域探索关联数据的先行者之一 LC,没有首先选择把馆藏目录发布为关联数据,而是提供关联数据服务(id.loc.gov),发布可供更大范围利用的名称规范库、主题词表、小型专用叙词表、各种代码表等。这些取值词表和 BIBFRAME 词表一起,成为图书馆关联数据应用的数据基础设施。RDA 的数十个参考取值词表(http://www.rdaregistry.info/termList/)也具有同样作用。PCC 的 MARC 中 URI 工作组编制了《创制和获取 URI:常用词表和参考源指南》[41],收录本领域及互联网上其他可用数据集,帮助图书馆目录消费关联数据。

为争取所在机构的支持,全面转向关联数据,LD4P 和 LD4P-Lab 项目选择可以通过外部数据强化内容的特藏资源入手开展实验,以突显关联数据的优势。同时他们利用项目资助开发了 5 个词表,大多为 BIBFRAME 的扩展,包括 bibliotek-o、艺术与珍稀资料扩展(Art and Rare Materials,ARM)、地理空间和地图资源本体(Geospatial and Cartographic Resources Ontology,GCRO)、LD4L 电影本体和演奏音乐本体(Performed Music Ontology,PMO),为各类特藏资源应用关联数据提供可选词表[42],加入数据基础设施中。

欧洲国家也积极加入。如丹麦通过应用 RDA 走向关联数据的 4 个步骤中[43],国家规范档建设是关键一步。国家规范档创建、维护和共享跨图书馆界的规范记录和标识符,包含作品和内容表达的书目规范记录,包含个人和团体的非书目规范记录,也可扩展包含其他实体类,并集成到国家数据基础设施和 VIAF。

5 结语

综上所述,文献编目从数字化到数据化的观念更新,在最近十年中已经从理论跨进到实践。通过关联数据在图书馆书目数据相关领域的应用,欧美各国正进入数据化的实际应用阶段。如 PCC 在《2018—2021 年战略方向》中所言,现在是时候超越关联数据相关理论层面的知识和技能,进入实施阶段[11]。

国内本领域相关研究并不缺乏,也有一些实际应用发布,最突出的是上海图书馆开放数据平台(data.library.sh.cn),公开发布上图数字人文项目所用的基础知识库、文献知识库和本体词表等。尽管如此,与中文资源相关的数据基础设施仍然严重不足。关联数据应用强调数据复用,因此并不需要完全重做一套,尤其是 RDF 词表,可以通过应用配置文件复用已有词表;取值词表中也有很多代码值不受语言影响,宜直接采用现有词表,如广受青睐的 RDA 参考取值词表。但主题词、分类号,词间差异大、体系不同,更适合自己建立。名称和题名规范档如不能加入国际合作,也宜通过调和与解析,建立与 VIAF、国际标准名称标识符(ISNI)和 Wikidata 等通用资源之间的关联。建立自己的关联数据基础设施,是我国加入编目数据化潮流的必经之路。

参考文献

[1] Marcum D B. The future of cataloging [J].Library Resources & Technical Services.2006,50(1):5-9.

[2] 胡小菁.编目的未来[J].大学图书馆学报,2008,26(3):18-22.(Hu Xiaojing. The future of cataloging in the Internet era[J]. Journal of Academic Libraries, 2008,26(3):18-22.)

[3] Tennant R. MARC must die [J]. Library Journal, 2002,127(17):26-27.

[4] Librisbloggen. LIBRIS available as linked data [EB/OL].[2019-02-20].http://librisbloggen.kb.se/2008/12/03/libris-available-as-linked-data/.

[5] Wikipedia.Digitization [EB/OL]. [2019-02-20].https://en.wikipedia.org/wiki/Digitization.

[6] 迈尔-舍恩伯格,库克耶著.大数据时代:生活、工作与思维的大变革[M].盛杨燕,周涛,译.杭州:浙江人民出版社,2013:104.(Mayer-Schonberger V, Cukier K. Big data: a revolution that will transform how we live, work, and think [M]. Sheng Yangyan, Zhou Tao, trans. Hangzhou : Zhejiang People's Publishing House, 2013:104.)

[7] Wikipedia.Linked data[EB/OL]. [2019-02-20].https://en.wikipedia.org/wiki/Linked_data.

[8] Galeffi A,Bertolini M V,Bothmann R L,et al.国际编目原则声明(ICP)[EB/OL].陈琦,译.[2019-02-20].https://www.ifla.org/files/assets/cataloguing/icp/icp_2016-zh.pdf.(Galeffi A,Bertolini M V,Bothmann R L,et al. Statement of International Cataloguing Principles (ICP)[EB/OL].Chen Qi,trans.[2019-02-20]https://www.ifla.org/files/assets/cataloguing/icp/icp_2016-zh.pdf.)

[9] 国际图联规范记录的功能需求与编号(FRANAR)工作组.规范数据的功能需求[EB/OL].[2019-02-20].https://www.ifla.org/files/assets/cataloguing/frad/frad_2009-zh.pdf.(IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR). Functional requirements for authority data[EB/OL].[2019-02-20].https://www.ifla.org/files/assets/cataloguing/frad/frad_2009-zh.pdf.)

[10] Oury C. ISSN register as linked data: using Bibframe for serials and other continuing resources[EB/OL].(2017-09-27)[2019-02-20].https://wiki.dnb.de/download/attachments/125433008/BibframeWorkshop_ISSNPortal_OURY.pdf?version=1&modificationDate=1507209493000&api=v2.

[11] PCC (Program for Cooperative Cataloging) Strategic Directions: January 2018-December 2021 [EB/OL].(2018-02-23)[2019-02-20].https://www.loc.gov/aba/pcc/about/PCC-Strategic-Directions-2018-2021.pdf.

[12] 国家图书馆《中国文献编目规则》修订组.中国文献编目规则(第二版)[M].北京:国家图书馆出版社,2005:5-6.(National Library of China's Revision Group of *Chinese literature cataloging rules*.Chinese literature cataloging rules, 2nd edition[M].Beijing : National Library of China Publisher, 2005:5-6.)

[13] 国际图联书目记录的功能需求研究组.书目记录的功能需求:最终报告[EB/OL].王绍平,等,译.[2019-02-20].https://www.ifla.org/files/assets/cataloguing/frbr/frbr-zh.pdf.(IFLA Study Group on the Functional Requirements for Bibliographic Records. Functional requirements for bibliographic records [EB/OL]. Wang Saoping, et al, trans. [2019-02-20].https://www.ifla.org/files/assets/cataloguing/frbr/frbr-zh.pdf.)

[14] IFLA.ISBD: international standard bibliographic description, consolidated edition [M/OL].[2019-02-20]. http://www.ifla.org/files/assets/cataloguing/isbd/isbd-cons_20110321.pdf.

[15] Riva P, Le Bœuf P, Žumer M. IFLA library reference model: a conceptual model for bibliographic information [EB/OL]. [2019-02-20].https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_

rev201712.pdf.

[16] 编目精灵.IFLA 图书馆参考模型中的 Nomen 实体(附 LRM 成为 IFLA 标准)[EB/OL].(2017-08-25)[2019-02-20].http://catwizard.net/posts/20170825110949.html.(Catwizard.Nomenentity in IFLA-LRM[EB/OL].(2017-08-25)[2019-02-20].http://catwizard.net/posts/20170825110949.html.)

[17] Glennan K. Recording names and access points[C/OL]. Preconference to ALA Midwinter meeting in Denver, Colorado, RDA Toolkit Redesign Update and Preview(2018-02-09)[2019-02-20].http://www.rda-rsc.org/sites/all/files/Recording%20names%20and%20access%20points.pdf.

[18] 胡小菁.规范控制:从名称选择到实体管理[J].数字图书馆论坛,2018(1):2-7.(Hu Xiaojing.Authority control: from selection of a name to entity management[J]. Digital Library Forum, 2018(1):2-7.)

[19] MARC PROPOSAL .Use of subfields $0 and $1 to capture uniform resource identifiers (URIs) in the MARC 21 Formats [OL]. (2017-05-16)[2019-02-20]. http://www.loc.gov/marc/mac/2017/2017-08.html.

[20] Zeng M L, Qin J. Metadata [M]. Neal-Schuman Publishers, 2008:15.

[21] 编目精灵.Bibframe.org 的类与属性发布[EB/OL].(2014-01-27) [2019-02-20].http://catwizard.net/posts/20140127230619.html(Catwizard.Classes and properties in Bibframe.org issued[EB/OL].(2014-01-27)[2019-02-20]. http://catwizard.net/posts/20140127230619.html.)

[22] 编目精灵.BIBFRAME2.0类的变化[EB/OL].(2016-05-22)[2019-02-20].http://catwizard.net/posts/20160502220754.html.(Catwizard.Change of Classes in BIBFRAME 2.0.[EB/OL].(2016-05-22)[2019-02-20].http://catwizard.net/posts/20160502220754.html.)

[23] BIBFRAME 2.0 RDF Conventions [EB/OL]. [2019-02-20].http://www.loc.gov/bibframe/docs/bibframe2-rdf-conventions.html.

[24] Report to the PCC BIBFRAME task group: final report of the CONSER CSR to BIBFRAME mapping task group [EB/OL]. [2019-02-20]. http://www.loc.gov/aba/pcc/bibframe/TaskGroups/CSR-PDF/FinalReportCONSERToPCCBIBFRAMETaskGroup.pdf.

[25] The codex metadata model [EB/OL]. [2019-02-20].https://wiki.folio.org/pages/viewpage.action?pageId=1415393.

[26] Denenberg R. Re: BIBFRAME 2.0 question[EB/OL].(2017-01-27)[2019-02-20].https://listserv.loc.gov/cgi-bin/wa?A2=ind1701&L=BIBFRAME&P=45806.

[27] RDA Steering Comittee.Implementation of the LRM in RDA [EB/OL].(2017-02-03)[2019-02-20].http://rda-rsc.org/ImplementationLRMinRDA.

[28] Williamschen J. The Library of Congress BIBFRAME editor[EB/OL].(2018-09-17)[2019-02-20].https://www.casalini.it/EBW2018/web_content/2018/presentations/Williamschen_1.pdf.

[29] Bibliographic framework initiative [EB/OL].[2019-02-20]. http://www.loc.gov/bibframe/.

[30] McCallum S. Expanding the Pilot[EB/OL].(2019-01-27) [2019-02-20].https://www.loc.gov/bibframe/news/source/alamw2019-lc-mccallum.pptx.

[31] BIBFRAME in production: Libris XL: linked data: the Swedish union catalogue [EB/OL].(2018-09-18)[2019-02-20]. https://www.casalini.it/EBW2018/web_content/2018/presentations/Lindstrom_2.pdf.

[32] Smith M, Stahmer C G, Li Xiaoli, et al.BIBFLOW: a roadmap for library linked data transition (Prepared 14 March, 2017) [EB/OL]. [2019-03-10]. https://bibflow.library.ucdavis.edu/roadmap/.

[33] LD4P 2.Cohort Meeting #1, Oct 15, 2018 [EB/OL]. [2019-02-20]. https://wiki.duraspace.org/display/LD4P2/Cohort+Meeting+%231%2C+Oct+15%2C+2018.

[34] Possemato T. From MARC to BIBFRAME in the SHARE-VDE project[EB/OL]. [2019-02-20]. https://www.loc.gov/bibframe/news/pdf/share-vde-alaal2018.pdf.

[35] Possemato T.The mapping and the conversion workflow from MARC to BIBFRAME [EB/OL]. [2019-02-20]. https://www.casalini.it/EBW2018/web_content/2018/presentations/Possemato_1.pdf.

[36] Ford K. Anonymous resources, blank nodes, and providers, oh my! [EB/OL].[2019-02-20].https://www.loc.gov/bibframe/news/source/alamw2019-lc-ford.pptx.

[37] LD4 community working group on reconciliation, 2017-2018 work plan [EB/OL].[2019-02-20].https://github.com/LD4/ld4-community-recon/blob/master/WORKPLAN.md.

[38] MARC PROPOSAL NO.2017-09 [EB/OL].[2019-02-20]. http://www.loc.gov/marc/mac/2017/2017-09.html.

[39] 胡小菁, 李恺. MARC 四十年的发展及其未来[J]. 中国图书馆学报,2010,36(2):83-89.(Hu Xiaojing, Li Kai. MARC: forty years of development and its future[J]. Journal of Library Science in China,2010,36(2):83-89.)

[40] Wikipedia.Data infrastructure[EB/OL]. [2019-02-20].https://en.m.wikipedia.org/wiki/Data_infrastructure.

[41] Program for Cooperative Cataloging Task Group on URIs in MARC. Formulating and obtaining uris: a guide to commonly used vocabularies and reference sources [EB/OL]. [2019-02-20]. https://www.loc.gov/aba/pcc/bibframe/TaskGroups/formulate_obtain_URI_guide.pdf.

[42] LD4P outputs[EB/OL].[2019-02-20]. https://wiki.duraspace.org/display/LD4P/LD4P+Outputs.

[43] Lindhard L J. BIBFRAME considerations-from an RDA-implementation perspective[EB/OL]. [2019-02-20]. https://wiki.dnb.de/download/attachments/125433008/201709_Lightning_Talk_BIBFRAME_Workshop_Lindhard.pdf?version=1&modificationDate=1507209428000&api=v2.

胡小菁 华东师范大学图书馆研究馆员。上海 200062。

(收稿日期:2019-03-23)