基于 D2R 发布学者关联数据集探究*
——以图书情报领域为例

牛永骎[1] 常娥[1,2]
[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] 探讨学者关联数据集的定位及其构建方法,以期为学科发展、学者评价与信息共享利用提供便利。[方法/过程] 在阐释现有机构知识库内涵基础上,分析学者关联数据集之功能特点,并以我国图书情报领域为例,通过开源软件 D2R 发布该领域学者的关联数据集。[结果/结论] 学者关联数据集不同于机构知识库,它以所属学科领域的学者为数据起点,以网罗一切相关信息资源,并以完全开放、关联与共享的方式提供知识。在学者关联数据集的构建与发布过程中,重点需克服实体 URI 定义、作者重名、专著与网络学术记录难以采全等问题。

关键词: 学者关联数据集 机构知识库 D2R
分类号: G250
DOI: 10.13266/j.issn.0252-3116.2017.19.002

1 引言

学者作为科研活动的基础与中坚力量,在科研活动中积累了大量宝贵的思想和学术成果。但目前尚缺乏对学者信息进行有效集中管理的方式,在学术成果采集过程中往往出现采全率低、学者重名等多种问题,使得大量的学者信息处于“孤岛”状态,导致学者学术成果零散,难以就学者整体学术成果进行挖掘与分析等,这为学者评价、学术信息传播与共享带来了极大的阻碍。

在大数据时代背景下,已有研究者与机构开始关注如何将零散的学者学术成果进行组织与管理,并不断引入新的思想和技术手段使其更加切实可行,但类似探索尚处于起步阶段,相关实践也仅仅是将各类信息资源进行简单汇总,缺乏深入组织,且信息资源的开放程度较低,导致建成的知识库利用率不高。实质上,对于每一位学者而言,无论是其个人基本信息、所发表的论文或学术专著,还是散落在网络上的各种学术记录等,它们都密切相关,各类不同信息相互关联才得以构成一个相对完整的“学者”。

关联数据作为一种新兴技术,其提出旨在利用统一资源标识符(Uniform Resource Identifier,简称 URI)和资源描述框架(Resource Description Framework,简称 RDF)发布、共享、连接各类数据、信息和知识,将文件网络转变成数据关联网络,从而进一步推动语义网的发展。将关联数据思想引入学者相关信息资源的组织中,并在网络中发布学者关联数据集,将有效打破学者信息“孤岛”状态,增强学者信息的共享与传递能力。从创建与发布关联数据集的平台来看,可以有多种选择,包括 Pubby、Drpual 等。而目前广为使用的开源软件 D2R 作为一种将关系型数据库发布为关联数据的工具,不仅可以帮助 RDF 或超文本标记语言(Hyper-Text Markup Language,简称 HTML)浏览器去定位数据库内容,而且允许各种应用通过 SPARQL 终端对数据库进行查询,而它最显著的优势来自其独立而灵活的 RDF 映射语言和表达规则,从而完成复杂关系结构的灵活映射[1]。有鉴于此,本文以图书情报学科为例,基于 D2R 尝试构建并发布了学者关联数据集,并重点探讨了在学者关联数据集构建过程中,各实体 URI 的定义、学者重名、专著以及网络学术记录难以采全这几大难题及其相应解决方案。

* 本文系国家社会科学基金项目“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: 牛永骎(ORCID: 0000-0002-8790-7072),硕士研究生; 常娥(ORCID: 0000-0001-6865-2118),副研究馆员,博士,硕士生导师,通讯作者,E-mail: chang_e@seu.edu.cn。
收稿日期: 2017-06-20 修回日期: 2017-08-18 本文起止页码: 13-21 本文责任编辑: 刘远颖

2 学者关联数据集与机构知识库之对比分析

数据资源的整合与共享已受到来自全世界的广泛关注并取得一定成果，单就关注研究人员信息资源整合而言，目前较为著名的是美国康奈尔大学的 VIVO 网，它通过关联数据和本体技术实现科学家及其研究成果的语义互联，并以提供专业社交系统方式为研究者之间的交流与合作提供便利[2]。VIVO 网信息资源主要由各机构、学术期刊和研究人员本人添加，资源类型主要包括学术、教育和服务三方面，暂未涉及学者科研数据、网络学术记录等信息资源[3]。VIVO 网在信息组织方面的新尝试具有开拓性和先导性，但由于 VIVO 本体涉及科学家、出版物、资源、资助者、地点、服务等诸多方面，建设难度较大，目前其涉及的领域仍然集中在生物医学领域，其使用范围与影响力远未达到预期效果。机构知识库作为一种较为传统的信息组织方式，因其构建难度较小、可操作性强等原因长期以来受到各机构与学者的广泛关注，建设成果丰硕。相比于 VIVO 网和机构知识库，学者关联数据集所采集的资源类型更接近于后者，且非常容易混淆，故本文将机构知识库作为一个重要参照进行对比分析。

2.1 机构知识库的概念及其发展概况

2002 年，美国麻省理工学院建立的首个机构知识库 Dspace@ MIT 在国内外掀起了一阵研究和建立机构知识库的热潮。除各类机构知识库层出不穷外，相关研究也是方兴未艾，国内外对于机构知识库的定义表述很多，但内涵基本一致，即以某一机构为中心，收集、整理和保存该机构所有学术成果、视频文档以及各种实验数据、手稿记录等相关信息资源的数据库[4-7]。

自 2002 年发展至今，机构知识库已有了显著的成就。截至 2017 年 1 月 17 日，OpenDOAR 显示收录了 3 291所机构知识库[8]，总数相较于 2014 年的2 624 所[9] 增长了 667 所，增速较快，而中国则维持与 2014 年数量相同的 39 所。近年来，有研究者开始关注面向关联数据的机构知识库构建以及探究学者知识库构建问题。例如，中国科学院机构知识库平台利用关联数据来实现语义扩展[10]，清华大学建成“清华学者库”，上海交通大学的“交大学者库”等。我国较早提出学者知识库概念的是何继红，她认为当前机构知识库因得不到学者支持而存在建设困难的问题，主要是源于机构知识库没能将其实质即学者作为建设的核心，因此提出建立机构学者库[11]。2010 年，何继红又进一步明确提出学者知识库的建设方法和构建模型[12]。除此之外，张首红于 2009 年提倡建立学者观点知识库[13]；周小萍在 2016 年探讨了以辽宁大学为例的高校学者知识库的构建方法[14]；吉宽宇则探讨了学者库建设与服务中著作权侵权控制问题[15]。而在相关研究中将学者知识库与关联数据相联系的，目前只有杨萌，她于 2015 年提出采用 Drupal 发布学者关联数据集的研究[16]，但她所定义的学者知识库是在院系背景下把学者的研究成果按照不同资源类型分类保存。

对于学者知识库这一概念，各个研究者在秉承以学者为主体的前提下，又加入了各自不同的理解。但无论是实践还是理论研究，基本未能跳出机构知识库框架的束缚，即重点讨论的是如何构建某一机构的学者知识库。诸如“清华学者库”“交大学者库”这种机构类的学者知识库，仍属于机构知识库研究范畴，其未来走向如何尚未可知。

2.2 学者关联数据集的概念及其定位

有别于上文所述机构知识库和学者知识库，本文提出以学者关联数据集方式采集、组织、整理与发布学者全景式学术成果数据。所谓学者关联数据集是指完全以学者为中心，对其基本个人信息、学术论文、专著、专利、网络学术记录（如博客、微博、个人网站等）以及实验记录、手稿、照片等与学术活动有关的信息进行完整采集、整理与保存，并以 RDF 和 URI 方式在网络上发布与共享的信息资源数据库。由此可见，学者关联数据集围绕具体学者，收集一切与其学术活动相关的信息，对于学者生活及其他方面资料不予组织。

学者关联数据集在某种程度上可以被视作机构学者知识库的泛化，它打破了机构的束缚，一般以学科领域为单位采集学者的基本信息和学术成果，并以 RDF 数据模型将其进行关联化组织，形成开放的数据关联网络。对于学者关联数据集，可以从不同角度对其进行分析与利用。首先，从学科角度来看，学者关联数据集可以对整个学科中的相关研究人员及其所有成果做关联汇总，以明确学科领域总体发展现状与水平，便于进行学科发展态势分析，获知热门研究方向以及领军人物等信息。其次，从学者自身角度来看，学者关联数据集通过完整收集某一学者的各种信息和科研产出，不仅反映该学者的科研状况与学术地位，也因为集成、关联与开放的组织形式而带来相对更多的学术关注，由原本用户仅关注某一学者的一两篇文章，转变为可以全面了解该学者的学术成长轨迹、速度以及研究方向等各种信息，这不仅大大提升了学者的学术影响力，

而且对于学者评价、学术人才引进等亦具有重要参考价值。最后，从用户角度来看，学者关联数据集不仅提供了学者的各类学术成果信息，而且提供了实验记录、手稿、网络学术记录等各种参考资料，用户通过对学者各类数据信息的思考、评判、总结与挖掘等处理，从而产生更有利用价值的学术信息。目前来看，这些功能是普通文献数据库所无法替代的。

2.3 两者对比分析

机构知识库(含学者知识库)与学者关联数据集既有区别又有联系。机构知识库与学者关联数据集的共同点在于：它们同属资源数据库范畴，其目的都是对某一范围内的相关信息进行汇总与整理，从而便于检索使用和知识共享。虽然机构知识库与学者关联数据集使用了不同的知识组织方式，所涉及信息范围亦不相同，但它们最终都会落实到以具体学者为出发点进行信息资源采集，在采集过程中都会涉及学者甄别、筛选以及将数据进行汇总的工作。

机构知识库与学者关联数据集虽然存在相似之处，但两者之间的差异性更加明显，主要表现在以下4个方面：

(1)数据边界范围不同。机构知识库在收集组织信息时主要是以某一机构为边界画圆，落在该机构圆圈内的人员才属信息采集对象。值得注意的是，机构内知识资产拥有者除了是以教师为主的学术研究者外，还包含有学生、实验员、教学辅助人员等。学者关联数据集以具体学者为中心，采集该学者的所有学术成果，包含其基本信息、论文、专著、网络学术记录等。为了便于构建学者关联数据集，一般以学科领域为边界画圆，落在该领域圆圈内的人员即属于信息采集对象。学者关联数据集内的学者同样包含教师、学生以及实验员等。

(2)功能与用途不同。机构知识库的建设虽在一定程度上支持了学术研究，具有信息资源保存与利用的功能特点，但它同时还具有为组织机构服务的功能，主要涉及到大学评价、机构评价和机构内的教师职称评审等，以彰显组织文化，提升社会影响力。学者关联数据集主要着眼于学科知识信息组织与利用，为学科发展和学术研究服务，通过该数据集可以了解学科领域的分化、交融、形成与发展等，同时还可以用来评价具体学者，明确学者的学术成长路径、主要科研方向、科研水平和学科地位等。

(3)数据采集重点不同。机构知识库在采集学者相关信息时，将学者是否属于该机构作为首要准则。即学者进入这个机构工作时，其所产生的相关信息才进行采集，进入该机构前，或者换单位后，该学者的学术成果都不是某个机构知识库的信息采集对象。换言之，对于某一机构知识库而言，学者的学术成果信息并不是该学者的全部学术成果。学者关联数据集以具体学者为中心，对其所有相关信息进行完整采集，即无论学者的工作单位是否发生变化，凡是与该学者有关的一切信息资源均进行采集。因此，对于学者关联数据集而言，学者的学术成果是否能够采集完整是衡量数据集质量的一项重要指标。

(4)知识组织方式不同。目前机构知识库主要以一维线性的元数据方式进行资源组织，且大多数以封闭数据形式存在。因此，机构知识库方式无法支持相关资源关联检索和知识推理。然而，学者关联数据集则以RDF数据模型和URI标识作为核心技术，以关联数据方式进行资源的存储与共享。因此，开放性、关联性与多维网络化获取是学者关联数据集的重要特点。

3 学者关联数据集的构建与发布

构建学者关联数据集是一项庞大而又复杂的工作，涉及到人员的选定、学者的甄别、学术成果及网络学术记录的采集等各种问题。由于学者关联数据集总体牵涉到的学者和成果数量相当庞大，导致建设过程很难一步到位，需要分学科、分层次、分批进行收录。首先，是对学科领域学者对象进行筛选，以此明确数据采集的核心范围；其次，是对实体类进行选择与抽取，除需确定关联数据集包含哪些资源类外，还需明确各资源类的属性描述，以此生成各数据表；再次，是对注册词表进行选择，并对实体类进行RDF化属性描述。优先选择已注册各种词汇表进行资源属性与关系描述，可增加学者数据集与其他数据集的融合性；再次，在实体类之间建立各种连接，以此构建真正的数据关联网络；最后，选择平台发布关联数据集，便于用户进行访问和利用。本文将以图书情报学科领域为例，探索构建学者关联数据集各步骤的具体实现方法。

3.1 领域学者对象筛选

在学者关联数据集构建前期，按照学科领域筛选出领域学者对象是重点。原则上，需要对某一学科的所有学者及其学术成果进行完整采集。然而由于学科规模十分庞大，例如数学、化学、生物学等，再加上学科领域学者群及其学术成果是动态更新的，所以几乎无法完整采集所有学者信息，因此可以学科作为界限，分层次、分批对领域学者及其学术成果进行收录。针对

领域学者对象筛选问题，本文认为主要有两条路径可以实现：

一是将学科作为总体界线，根据学者发文量、h 指数及类 h 指数等标准对领域学者进行排序并根据排名先后分批次录入。该方法可以反映学科概况，体现学者科研水平，且操作性强，但是在实际运用中缺乏针对性，且 h 指数受限于具体数据库，因此容易遗漏其他学者信息，用户难以快速寻找自己感兴趣的相关内容。

二是对学科进行细分，在一级学科基础上细分出二级学科，然后再细分研究方向。根据二级学科或研究方向筛选出相关学者。该方法较好地弥补了路径一缺乏针对性的缺陷，并且提供了明确的学者研究方向，用户可以根据这一属性快速定位和筛选自己感兴趣的内容。然而，由于一个学科中不同研究方向必然存在发展水平差异，使得学者与成果数量悬殊较大，不同方向之间难以平衡。此外，在研究方向的划分上可能存在争议，实际发展过程中并不是所有学科都有明确的方向划分，可能存在众多零散分支，同时一个学者拥有两个及以上研究方向的情况比比皆是，从而造成数据冗余，这些都将给学者信息资源组织带来挑战。

以图书情报学为例，我国最新的研究生招生信息网数据表明，2017 年共有 73 所院校招收该专业硕士，其中有 20 所高校并不划分具体研究方向；其他 53 所院校共划分出 199 个研究方向，其中有 139 个研究方向仅出现一次，其研究方向划分异常零散。有鉴于此，本文更倾向于选择第一种方案，即从学科整体出发根据学者影响力进行分层，该做法虽前期存在诸多缺漏，但随着数据集的不断完善弊端必会逐渐减少。

3.2 实体抽取与数据表生成

考虑到数据的可更新性，本文选择了可将关系数据库内容发布为关联数据的转换平台 D2R。囿于研究时间和精力，本文领域学者关联数据集所选取的实体主要包含 3 个：人员（person）、论文（article）和专著（book）。将论文和专著分别与人员相关联。人员实体主要介绍学者基本信息，一般包含人员 ID、姓名、邮箱、单位和职位等几个属性。而论文和专著作为学者科研成果的展现，一般包含论文 ID/专著 ID、人员 ID、题名、责任者、关键词、页码、摘要、出版者和出版时间等几个属性。数据表及其所含属性信息如图 1 所示：

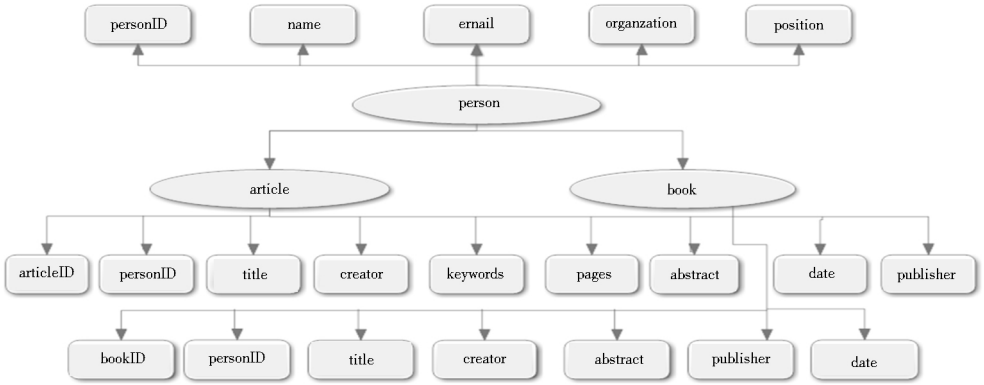

图 1 领域学者实体与属性关系

3.3 词表选择与实体 RDF 化描述

词表在关联数据的发布中具有重要作用，它可以将词汇标准化，为接下来实体的 RDF 化描述提供前提。在本文的案例中，笔者主要构建了 3 个实体以及与之相关的 15 个属性，涉及多个标准词表。本文主要选取了书目本体（Bibliographic Ontology，简称 BIBO）词表、都柏林核心元数据（Dublin Core，简称 DC）词表、描述个人基本信息的朋友的朋友词表（Friend-of-a-Friend，简称 FOAF）和人物档案词表（A vocabulary for biographical information about people，简称 Bio）进行词汇的 RDF 映射。具体如表 1 所示：

表 1 实体属性与词表对应情况

词表	实体属性
BIBO	article、book、abstract、email、pages
DC	publisher、title、creator、date、identifier
FOAF	person、name、organization
Bio	position、keywords

根据研究需求，需要将 3 组实体分别进行 RDF 化描述，即用“实体-属性-值”的数据模型来描述每一个实体特征。“学者（person）”实体主要包括学者 ID、姓名、邮箱、机构和职位 5 个属性，将其转换为 RDF 数据，如图 2 所示。“论文（article）”实体主要包括论文 ID、人员 ID、题名、作者、关键词、页码、摘要、出版者和出

版时间 9 个属性，将其转换为 RDF 数据，如图 3 所示。“专著”实体与“论文”实体的 RDF 数据模型相似。

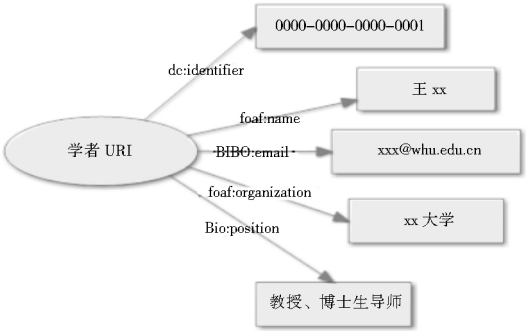

图 2 “学者”RDF 数据模型

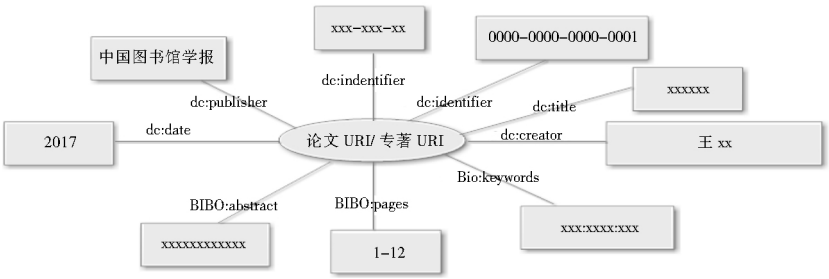

图 3 “论文/专著”RDF 数据模型

3.4 实体链接生成

生成实体之间的链接在发布关联数据过程中相当重要，直接影响到关联数据发布的质量以及用户使用的便利程度。实体链接是指在实体 RDF 化描述的基础上，在不同实体之间建立关联，从而构建关联数据网络。本文以华东师范大学许鑫副教授于 2015 年所作的关于图书情报领域学者的 h 指数与 ht 指数排名为基础[17]，选择排名靠前的 20 名学者进行领域关联数据集构建研究，并以邱均平教授为例进行展示。邱均平教授任职于武汉大学信息管理学院，他撰写了大量的论文和专著，本文选择其作品《国内外人文社会科学科研成果评价比较研究》(论文) 和《教育评价学：理论·实践·方法》(著作)，展示学者、论文与专著三实体间链接的构建过程，具体如图 4 所示：

3.5 关联数据发布与访问

关联数据发布平台 D2R 主要包括 D2R Server、D2R Engine 和 D2R Mapping 3 个组成部分。其中，利用 D2R Mapping 部件生成映射文件是发

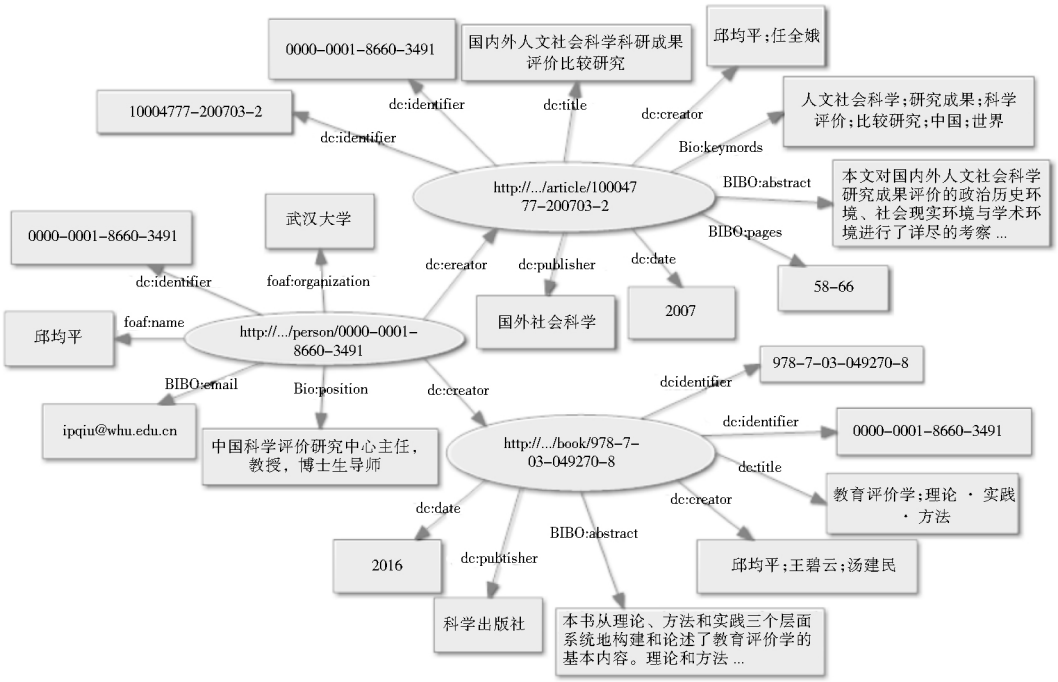

图 4 领域学者实体关联

布关联数据的关键所在，通过映射文件可以将关系数据表中的数据映射转换成 RDF 数据，并支持 URI 的全球性访问。针对本文的 RDF 数据模型和 Scholars 数据库内容，领域学者关联数据集的具体发布过程如下：

首先，通过 D2R 服务启动脚本 d2r-server.bat，并生成映射文件 mapping.ttl。

然后，启动关联数据发布服务 D2R Server，在浏览器中输入 http://localhost:2020，得到如图 5 所示的运行界面，即表示映射转换成功。

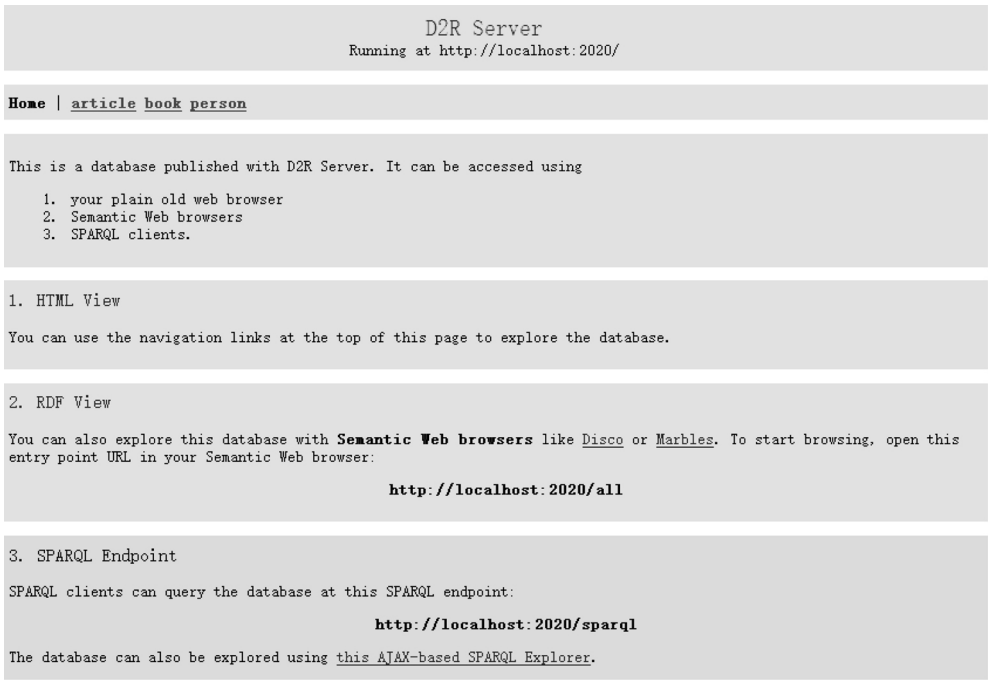

图 5 领域学者关联数据集 D2R 访问界面

用户可通过实体类进行关联数据的访问与查询。点击任何一实体的超链接都可得到其相关元数据属性信息，例如，在图 5 中点击“person”可得图 6 所示的内容。

图 6 “学者”实体类的访问界面

点击编号为“1”的“person”，可得到如图 7 所示的关联数据(部分)。图 7 具体给出了邱均平教授的基本个人信息以及论文与专著链接情况。

不同于学者姓名、邮箱、工作单位等实体属性信息，论文和专著都是实体类，因此都被赋予了实体 URI 标识，可直接点击访问。例如，点击 http://localhost:2020/resource/article/10004777-200703-2 则可获得编号为“272”的论文的详细信息，见图 8。

专著与论文的访问与查询方法完全一致。在完成相关信息查询或希望返回了解一些信息时，可以通过点击页面中的<http://localhost:2020/resource/person/0000-0001-8660-3491>，回到关于具体学者的浏览界

图 7 具体学者的查询界面

图 8　领域学者关联数据“论文”类数据界面

面或者可以直接点击“Home”返回到初始界面，完成数据访问过程。

4　重点问题探讨

4.1　实体 URI 定义问题

定义实体 URI 在构建关联数据集中是一项关键而又困难的工作，在确保唯一性的同时人们也希望 URI 包含语义功能。因此，本文在构建学者关联数据集过程中自定义了一套 URI 生成体系，即按照“http://域名/实体类型/实体 ID”模式创建。本研究域名采用 D2R 本地域名 localhost:2020，实体包括学者(person)、文章(article)和著作(book) 3 种类型，并分别构建 personID、articleID 和 bookID 进行实体唯一性识别。

首先，对于 personID，本文优先选择注册广泛的开放研究者与贡献者身份(Open Researcher and Contributor ID，简称 ORCID)进行识别，并以国际标准名称识别码(Register for the International Standard Name Identifier，简称 ISNI)作为补充，原因在于并不是每一位学者都注册了 ORCID，ISNI 采用了与 ORCID 一致的 16 位数字形式，且与之不重复。但值得注意的是，ORCID 和 ISNI 的 16 位编码数字为流水式，不具备语义功能，因此尚有改进的空间和必要，未来应考虑加入语义使之有规律且方便记忆。

其次，对于 articleID，本文将文章分为 4 种类型，即期刊、报纸、会议和学位论文，其中期刊 articleID 采用“国际标准连续出版物编号(International Standard Serial Number，简称 ISSN)-出版年-期号-序列号”形式进行编码，对于没有 ISSN 号码的期刊则采用“刊名首字母-出版年-期号-序列号”形式进行编码；报纸 articleID 采用“国内统一刊号-出版年月日-序列号”形式进行编码；会议 articleID 采用“主办方英文简称-届次(数字)-C-年份-序列号”形式进行编码；学位论文 articleID 采用“学校英文简称-D-年份-序列号”形式进行编码。

最后，专著实体的 ID 定义相对简单，鉴于每本书都有一个 10 位或 13 位的国际标准书号(International Standard Book Number，简称 ISBN)，因此 bookID 直接采用 ISBN 号码进行标识。

以上实体 URI 定义体系，尤其是实体 ID 定义，尚存在不完善之处，有些实体 ID 定义较为复杂，例如概念类实体 ID 该如何定义，本研究暂未涉及。

4.2　学者重名和别名问题

学者重名问题在中国乃至全球都是一项普遍存在而又难以解决的问题，它可能造成信息采集不全、计量偏差、研究成果归属混乱，甚至是盗用同名学者成果的学术不端行为。为解决这一问题，国内外学术界主要提出了人工辨识、数据库字段修正和基于机器学习的辨识等解决方案[18]。而为人熟知的 ORCID 也是区别学者的重要途径，但 ORCID 的诞生并不能完全解决重名问题，很多学者因为没有直接面对重名带来的困扰，以致注册 ORCID 的积极性不高。另外，对于已故研究者以及研究者在注册 ORCID 之前的学术成果的组织

与整理仍旧是一项棘手的工作。作为具有较高辨识度的 ORCID 方法在解决作者重名问题上还需要长时间的不断发展与积累。

考虑到同一机构中学者重名概率较低，本文提出“作者+机构”的组合识别模式以克服学者重名问题。在筛选出领域学者名单后，以“作者+机构”这一组合检索方式，在不同数据库中查找、收集和整理其学术成果，以克服学者重名而带来的数据噪音。为了获取学者机构信息，需要首先了解学者求学和工作经历，把某一学者所有相关的工作单位都收集起来，这样不仅可解决学者重名问题，而且可解决由于学者工作单位变动以及求学单位和工作单位不一致而带来的数据无法采全的问题。

学者别名问题指的是同一个学者拥有两个及两个以上姓名。在学术研究领域，学者一般以正式的姓氏名称进行署名，同时作为成年人，学者更改姓名的情形亦很少出现，因此对于正式出版的学术成果而言，学者别名问题可以忽略。然后，对于网络学术记录，学者别名问题确实是不容忽视的重要问题。如何正确识别学者的网名、笔名等各种别名问题有待进一步探索和研究。

4.3 专著采集问题

专著作为学者另一重要学术成果，在采集方面依旧存在相当大的问题。首先，著者或编者重名情况依旧存在，但因为专著数量相较于论文要少很多，并且同名者在类似领域能够同出专著的情况更为罕见，因此在做人工筛选的过程中难度相对降低，但仍然需要对作者进行前期了解。其次，没有专门的专著数据库，给专著信息的采集带来了困难。虽然在我国所有出版物都需上交一份给国家版本数据库，但目前版本数据库还没有提供开放的网络检索服务。当下，几乎没有任何一个网上资源可以查全一位学者的所有专著，因此在采集著作过程中需要结合多方力量进行补充与确认。如果学者本人博客或机构网站列出所著专著，需将其作为重要参照，但该类信息可能存在滞后或不全的问题。

以邱均平为例，本文分别选择学科实力最强的武汉大学图书馆、中国高等教育文献保障系统(China Academic Library & Information System，简称 CALIS)联合目录公共检索系统和国家图书馆进行其馆藏查询。经过实践调查，本文认为应将国家图书馆馆藏目录检索系统作为首要选择，它作为亚洲规模最大的图书馆、世界上最大的图书馆之一，馆藏量高达几千万，由此该系统的查全率远高于其他网站，但为了确保著作的全面性，还可将 CALIS 目录检索系统或者当当网等购书网站的信息作为保障和补充。

4.4 网络学术记录采集问题

网络学术记录包含众多未公开发表却又极有价值的信息，是学者成果的又一重要载体，对其进行开发和利用，具有重要的现实意义。考虑到网络学术记录多为免费、开放的原生数字资源，重点可从以下途径进行采集：①学者个人网站、网络博客，着重采集其中有关学术的文章、观点与评论等记录信息；②专业学术论坛或各学科领域论坛网站等，如小木虫、丁香园、阿果资源网及研学论坛等；③相关的开放资源网站，如分享学者公开课的爱课程、学堂在线、新浪微盘等；④通过学术搜索引擎进行补充，如著名的 Google 学术搜索、免费电子图书搜索引擎以及中国法网搜索引擎等。网络学术记录分布极其广泛，在采集时应针对各学科领域的研究内容和相关学者、机构等做全面的了解工作，才能使得网络学术记录资源尽可能采集完备。

由于学者网络 ID 灵活多变，网络学术记录极为分散，因此相较于论文和著作两种实体而言，网络学术记录的采集与处理难度更大，故本文在构建学者关联数据集过程中暂未涉及网络学术记录。

5 结语与展望

全景式学者关联数据集除了包含学者的论文和专著等学术成果外，还应该包括诸如博客、微博、个人网站等网络学术记录以及实验记录、手稿、照片等相关信息。本研究旨在提供一种构建学者关联数据集的方法和思路，但囿于研究时间和精力，笔者仅针对我国图书情报领域“学者-论文-专著”的连接情况进行了关联构建，还存在诸多缺漏之处。除学者、论文和专著这 3 个实体类外，未来还可以进一步拓展出更多的实体类，例如机构、个人网站、博客、照片、手稿、实验数据等，以实现整个学科领域所有学者、所有研究成果以及所有机构的大串联。如果将机构拓展为概念类，与学者类、学术成果类进行关联，那么学者关联数据集则将拥有并超越机构知识库之功能。由于该设想工程浩大，尚缺乏完整的体系结构，未来笔者将投入更多的时间和精力去实现并完善它，以构建出更加丰富多元的学者

关联数据集。

参考文献：

[1] 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.

[2] 张艳侠，齐飞，毕强.关联数据的语义互联应用研究-以 VIVO 为实例[J].图书情报工作，2013，57(17)：16-20.

[3] 科学家信息交流的语义模型：VIVO 本体系统[EB/OL].[2017-08-09].http://blog.sciencenet.cn/blog-1557-454594.html.

[4] WARE M. Institutional repositories and scholarly publishing [J]. Learned publishing，2004，17(2)：115-124.

[5] LYNCH C A. Institutional repositories：essential infrastructure for scholarship in the digital age [J]. Portal-libraries and the academy，2003，3(2)：327-336.

[6] 赵继海. 机构知识库：数字图书馆发展的新领域[J].中国图书馆学报，2006(2)：33-36.

[7] 张晓林，张冬荣，李麟，等.机构知识库内容保存与传播权利管理[J].中国图书馆学报，2012，38(4)：46-54.

[8] OpenDOAR. Directory of Open Access Repositories [EB/OL].[2017-08-17].http://www.opendoar.org/.

[9] 王锰，陈雅.2004-2013 年国内机构知识库演进与发展研究[J].图书馆理论与实践，2015(6)：42-45.

[10] 王思丽，祝忠明.利用关联数据实现机构知识库的语义扩展研究[J].现代图书情报技术，2011(11)：17-23.

[11] 何继红.关于机构库发展为机构学者库的探讨[J].情报资料工作，2008(1)：55-59.

[12] 何继红.以人为本 科学构建学者知识库[J].图书情报工作，2010，54(8)：131-135.

[13] 张首红.数字图书馆中自动创建知识库的研究——以自动创建“教育技术学学者观点知识库”为例[J].现代教育技术，2009，19(12)：95-98.

[14] 周小萍.高校学者知识库的构建研究——以辽宁大学学者知识库为例[J].农业图书情报学刊，2016，28(9)：17-20.

[15] 吉宽宇.学者库建设与服务中的著作权侵权控制研究[J].图书馆建设，2016(7)：16-21.

[16] 杨萌.基于 Drupal 发布学者知识库关联数据的研究[J].图书馆研究，2005，45(5)：22-26.

[17] 许鑫，王诺.图书情报领域学者 Ht 指数分析[J].西南民族大学学报，2015(1)：230-234.

[18] 袁军鹏，俞征鹿，苏成，等.作者重名辨识研究进展[J].数字图书馆论坛，2011(10)：60-65.

作者贡献说明：

牛永骎： 撰写论文，收集和分析数据；

常娥： 构建论文总体框架，并指导论文写作与修改。

Research on Publishing Scholar Repository Linked Data Based on D2R

Niu Yongqin[1]　Chang E[1,2]

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

[2] Southeast University Library，Nanjing 210096

Abstract: **[Purpose/significance]** This paper discusses the positioning and construction methods of the scholar repository linked data，with a view to facilitating the development of subjects，evaluation of scholars and information sharing and utilization. **[Method/process]** On the basis of explaining the knowledge of the existing institutions repository，analyzing the functional characteristics of the linked dataset of scholars，and taking the field of library and information as an example，this article will publish the linked dataset of scholars in this field by D2R. **[Result/conclusion]** The scholar linked dataset differs from the institution repository，starting from the scholars in the subject field，in order to recruit all the relevant information resources，and provides the knowledge in a completely open，correlated and shared way. In the process of building and releasing the data set，the emphasis needs to be overcome by the definition of entity's URI，the author's duplicate，monograph and the academic records of network is difficult to adopt.

Keywords: scholar linked dataset　institutional repository　D2R