

高校图书馆持久标识符应用研究

□ 罗鹏程* 崔海媛 聂华 朱玲 韦成府

摘要 为了实现高校图书馆数字资源的持久标识,促进系统间互操作性的增强,本文对持久标识符进行了广泛深入调研,对当前应用最为广泛的4种持久标识符的原理和国内外应用情况进行了梳理,从费用、使用量、互操作、标识符与元数据四个方面对其做了全面地比较分析。结合北京大学图书馆的实际情况,介绍了持久标识符在开放研究数据平台、机构知识库、系统间互操作上的应用。

关键词 持久标识符 数字对象标识符 句柄 图书馆资源 唯一标识符

分类号 G250.7

DOI 10.16603/j.issn1002-1027.2017.05.015

1 引言

在网络环境下,统一资源定位符(Uniform Resource Locator, URL)容易改变,不能很好地用于资源的持久标识和引用。同一资源可能有多个副本,多种格式,从而拥有多个URL地址,通过URL无法将其识别为同一资源。持久标识符能够对资源进行持久地、唯一地标识,有助于资源的引用、识别、定位和长期保存,在数字出版、数字资源长期保存等领域得到了广泛研究和应用,已有多种持久标识符方案被提出并应用于实践。

在数字出版领域,持久标识符得到了非常广泛地应用。国外期刊论文数字对象标识符(Digital Object Identifier, DOI)注册机构交叉引用公司(CrossRef)成立于2000年,已注册DOI数量超过8000万个;国内期刊论文DOI注册机构中文DOI成立于2007年,已注册DOI数量超过了2500万个。在科学引文索引(Web of Science)网络版、爱思唯尔引文数据库(Scopus)、谷歌学术等数据库中,均支持对DOI的检索;许多期刊的官网(如《自然》、《科学》)中也包含了DOI等持久标识符信息。

图书馆作为数字资源长期保存的重镇,也特别重视持久标识符的应用和推广。早在1996年,美国国会图书馆便在美国国家数字图书馆项目中使用句柄(Handle)为数字化资源赋予持久标识符^[1]。德国

科技图书馆于2004年开始使用DOI为科学数据赋予持久标识符,并于2009年领导成立了数据引用组织(DataCite),专注于研究数据和科研产出的持久标识^[2]。法国国家图书馆于2006年实现了档案资源键(Archival Resource Key, ARK)持久标识符服务,用于馆藏数字文档和书目记录标识^[3]。大英图书馆作为DataCite创建者之一,面向英国高等教育机构,为数据提供DOI注册服务^[4]。美国加州数字图书馆于2009年实现了DOI和ARK持久标识符注册服务EZID(Easy-eye-dee),用于标识本馆数字资源,同时对外提供付费注册服务。中国国家图书馆参考Handle系统,保持与其兼容的基础上(在Handle中保留了108前缀)自行研发了中文数字对象标识符(Chinese Digital Object Identifier, CDOI)系统,面向国内公共图书馆提供资源的注册服务,并于2012年投入试运行^[5]。

当前,国内外高校图书馆对持久标识符的应用主要集中在机构知识库和研究数据管理中。如麻省理工学院机构知识库(Dspace@MIT),康奈尔大学机构知识库(eCommons),复旦大学社会科学数据平台使用Handle对资源持久标识,哈佛大学研究数据平台(Dataverse),密歇根大学政治与社会学研究所校际联盟数据平台(Inter-university Consortium for Political and Social Research, ICPSR)使用DOI对

* 通讯作者:罗鹏程,ORCID:0000-0001-9598-0715,luopc@lib.pku.edu.cn。

资源持久标识。与各国国家图书馆相比,高校图书馆持久标识符的研究还比较匮乏,持久标识符的应用缺乏整体规划。随着高校图书馆数字化资源的不断积累,资源的长期保存变得越发重要,持久标识符系统作为长期保存的重要组件急需建立并完善。同时,高校科研产出管理的信息化程度在不断提高,其规范化程度也需要进一步加强。持久标识符作为资源的身份号码,有利于资源的持久引用、查找和定位,各异构的系统间可通过标识符识别、关联资源,提高系统间互操作性。文章将对比分析各持久标识符系统的发展、原理、应用、资费、互操作等,并结合北京大学图书馆的实际情况,介绍持久标识符的相关应用。

2 发展概况

万维网出现之初,人们便意识到 URL 不具有稳定性,随着时间的推移常常成为死链。持久标识符基于中间解析系统实现资源的持久标识,当资源对应的 URL 改变时,仅需更新解析系统中的映射关系,用户仍可通过相同的标识符访问资源。从上世纪 90 年代中期开始,一些持久标识符方案陆续被设计、实现,主要包括:Handle(1994 年)、PURL(1995 年)、DOI(1998 年)、ARK(2001 年)。

(1) Handle。

Handle 由美国国家研究创新机构(Corporation for National Research Initiatives, CNRI)设计实现,最早用作 CNRI 数字对象架构中的一个组件,于 1994 年作为计算机科学技术报告项目的一部分提出并实现。2014 年 1 月, CNRI 和国际电信联盟共同发起,在瑞士日内瓦注册成立了新的 Handle 全球运营与管理组织——数字对象编号管理基金会(Digital Object Numbering Authority Foundation, DONA)。DONA 由 4 个初始的多主管员(Multiple Primary Administrators, MPA)组成:德国哥廷根科学数据分析与处理公司(Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, GWDG)、工信部电子科技情报研究所—中数创新公司—西恩多纳公司(ETIRI-CDI-CHC)联盟、美国 CNRI、国际电信联盟^[6]。CNRI 主要关注数字图书馆、政府和老用户的标识符应用, GWDG 主要关注科学数据标识符分配, ETIR-CDI-CHC 主要关注中国物联网和商业方面的应用^[7]。

(2) PURL。

持久统一资源定位符(Persistent Uniform Resource Locator, PURL)由联机计算机图书馆中心(Online Computer Library Center, OCLC)设计,于 1995 年在 Apache HTTP 服务器的基础上实现 PURL 功能,并于 1996 年应用于因特网编目项目^[8]。2007 年, OCLC 与 Zepheira 公司合作对 PURL 服务做了重新设计^[9],发布了 PURLZ-Server-1. X 版本。2011 年三圆石公司(3 Round Stones)接管了 PURL 的开发工作,于 2013 年发布 PURLZ-Server-1. 6. 4,此后该软件不再更新,而在三圆石公司开发的关联数据管理系统 Callimachus 中包含 PURL 的所有功能。PURL 认为持久性并不是一个技术问题,而是一个机构的服务功能,因此 PURL 采用了 URL 的形式,但它比普通 URL 能得到更多持久性保障^[10]。

(3) DOI。

数字对象标识符(Digital Object Identifier, DOI)源于国际出版商协会、国际科学技术和医学出版商协会、美国出版商协会的联合倡议,由国际数字对象标识符基金会(International DOI Foundation, IDF)管理。1997 年, DOI 在法兰克福书展上发布,同年 IDF 正式注册成立。2000 年, DOI 首个注册代理机构 CrossRef 成立,致力于为学术出版物提供 DOI 注册服务^[11]。2010 年 DOI 系统正式通过成为国际标准,并于 2012 年由 ISO 正式出版国际标准《ISO 26324 信息和文档—数字对象标识符系统》(ISO 26324 Information and Documentation—Digital Object Identifier System)。DOI 认为没有元数据,标识符的意义不大,为此 DOI 开发了相关数据模型存储元数据,并基于 Handle 系统实现了标识符解析功能。

(4) ARK。

档案资源键(Archival Resource Key, ARK)由美国国家医学图书馆约翰·昆茨(John Kunze)和罗杰斯(R. P. C. Rogers)设计,于 2001 年发布了互联网草案 draft-kunze-ark-00. txt。目前,加州数字图书馆负责 ARK 的维护和注册,最新草案为 2013 年发布的 draft-kunze-ark-18. txt。用户可以根据 ARK 方案构建自己的标识符服务,也可以使用加州数字图书馆提供的 EZID 为资源分配标识符。ARK 认为标识符的主要功能是将用户引入服务所在地,资源 URL 的失

效是由服务提供者造成的,与 URL 本身无关^[12]。因此,ARK 采用了可执行(Actionable)形式的标识符,它具有 URL 的形式,并可直接访问。

3 国内外应用

Handle、PURL、DOI、ARK 均起源于 2000 年前后,有多年的发展历史,在国内外有广泛的应用,有的甚至已经突破了数字资源标识的范围,并推广到物联网领域。

(1) Handle 的应用。

Handle 最初使用者主要包括美国国会图书馆、DSpace 和 DOI,2000 年以后用户群体进一步扩大到科学数据、电影电视行业、物联网等领域。目前,Handle 在数字图书馆与内容仓储、数字出版、非正式出版物、数字博物馆、远程教育、科学研究、数字权益管理、信息安全管理与隐私保护等领域得到大量应用^[13],下面介绍近年来几个重要的 Handle 应用。

澳大利亚国家数据服务(Australian National Data Service,ANDS)于 2009 年启动,由澳大利亚政府资助,基于 Handle 系统构建了持久标识符服务(Persistent Identifier Service,PIDS)。PIDS 面向澳大利亚机构和个人免费提供标识符注册。包括两种服务模式:对于研究者个人,可以登录澳大利亚研究数据系统使用 PIDS 自服务工具创建持久标识符;对于机构,需要先注册,然后使用客户端实现自动注册。PIDS 根据 ANDS 的需求对 Handle 客户端做了包装,提供了基于 Tomcat 的 Web 应用,并且开放源代码^[14]。

欧洲持久标识符联盟(European Persistent Identifier Consortium, ePIC)成立于 2009 年,基于 Handle 构建了持久标识符服务,面向欧洲研究社区,为成员机构提供注册服务。ePIC 具有高可靠性,在欧洲拥有 5 个数据中心,相互备份对方持久标识符。当前,ePIC 使用了 30 个 Handle 前缀,并已创建了 3 千万个标识符^[15]。此外,ePIC 系统源代码开放,服务提供者可利用 ePIC 系统构建自己的持久标识符系统。

基于 Handle 的中国国家物联网标识公共服务是“国家物联网标识管理公共服务平台”项目的组成部分。该项目由国家发改委于 2013 年批复,由中科院计算机网络信息中心牵头,联合工信部电子技术情报研究所(ETIRI)、工信部电信研究院、中国

物品编码中心建立物联网标识统一管理和公共服务平台。ETIRI 联合北京中数创新技术有限公司(CDI)、北京西恩多纳信息技术有限公司(CHC)共同组成了“ETIRI-CDI-CHC”联合体,并成为 DONA 的多主管员(Multiple Primary Administrators, MPA)。基于 Handle 构建了国家物联网标识公共服务,并将其应用于婴幼儿配方乳粉领域,现已经注册超过 2.2 亿个标识符^[16]。

(2) PURL 的应用。

PURL 的应用相对较少,目前主要包括联机计算机图书馆中心、美国国家生物医学本体中心、佛罗里达图书馆自动化中心、美国政府出版署等。澳大利亚国家图书馆曾经使用 PURL 提供持久标识符服务,不过现在该服务已停止运行,联机计算机图书馆中心的 PURL 服务目前也不能登录使用。与其他标识符相比,PURL 软件开发不够活跃,2010 年曾发布 2.0 Beta 版,但是之后一直没有发布正式版。目前 PURL 软件的开发已经融入到了 Callimachus 系统,为关联数据提供 URI 标识。

(3) DOI 的应用。

DOI 主要用于数字出版领域,目前有 10 个注册代理机构负责标识符的分配和应用服务,包括 CrossRef、DataCite、中国科学技术信息研究所等。在 CrossRef 中注册期刊超过 4.7 万种,注册 DOI 数量超过 8000 万个^[17],成为 DOI 最大的注册代理机构;在中国科学技术信息研究所中文 DOI 中,注册期刊超过 6200 种,注册 DOI 数量超过了 2500 万个^[18],仅次于 CrossRef。学术出版物一直以来是 DOI 最大的应用领域,然而随着开放获取运动的发展和开放数据理念的推广,数据集也逐渐成为 DOI 的重要应用领域。

DataCite 面向数据进行 DOI 注册,成立于 2009 年,目前已有 20 多个正式会员(如德国科技图书馆、加州数字图书馆等)、约 900 个数据中心,注册 DOI 数量超过 600 万个、每月解析量超过 700 万^[19],在数据集持久标识符领域有着巨大的影响力。DataCite 信息基础设施为元数据仓储(Metadata Store)系统,用于标识符的分配和元数据的管理。除数据集注册外,DataCite 还拥有数据仓储注册服务 re3data.org,其中含有来自 60 多个国家的 1700 多个数据仓储。

(4) ARK 的应用。

ARK 已经在许多机构得到应用,加州数字图书馆 ARK 注册列表中包含了来自全球的 400 多家机构,如法国国家图书馆、美国西北数字档案馆、德国柏林犹太博物馆、哈佛大学、谷歌等。法国国家图书馆是 ARK 最重要的应用机构之一,下面以其为例,介绍 ARK 在法国国家图书馆的应用。

法国国家图书馆基于以下两方面的因素采用 ARK:一是财务独立,ARK 源自公共文化领域,其注册、应用均无需付费,而 DOI 源自商业领域,费用相对较高;二是技术独立,ARK 标准没有强制使用特定厂商的服务,也不依赖于特定意图的全球软件基础设施,机构能以最简单的方式在本地实现,而 Handle、PURL 则依赖于特定的软件基础或外部服务。基于此,法国国家图书馆于 2006 年实施了 ARK 持久标识符方案,初始阶段将其用于数字化文档和编目记录的标识。之后领域得到进一步扩大,应用于 data.bnf.fr 的链接数据服务和可伸缩保存与归档仓储。此外,法国国家图书馆还积极参与 ARK 国际合作,与美国国家医学图书馆一起,为加州数字图书馆 ARK 注册列表提供备份服务^[3]。

4 比较分析

表 1 对 4 种主要的持久标识符进行了对比。主要从费用、使用量、数据安全、互操作、标识符与元数据 5 个角度进行了对比。

(1) 费用。

Handle 的费用依据不同 MPA 而不同。CNRI 每个 Handle 前缀一次性注册费用为 50 美元,并且每个前缀需要缴纳每年 50 美元的年费,提供免费软件运行本地解析服务,但不提供技术支持。ETIRI-CDI-CHC 目前针对图书馆的收费标准大致为每年 1.2 万元人民币,并提供软件和技术服务。GWDG 费用标准不明。PURL 完全免费,可以使用已有

PURL 服务注册持久标识符,也可以使用 PURL 开源软件在本地运行 PURL 服务。ARK 的注册不收取费用,可以使用加州数字图书馆的 EZID 服务,对于研究机构和非赢利性组织的年费为 1500 美元,每年最多可以创建维护 100 万个 ARK。也可以依据 ARK 标准自行实现 ARK 服务,无需向任何机构缴纳费用,如法国国家图书馆。

DOI 的费用依据其获取途径的不同而不同。作为 IDF 的会员,其收费标准为:DOI 创立会员会费 7 万美元;一般会员会费 3.5 万美元;注册代理会员会费包含在运营费之中,按照费用分担模型计费;附属会员会费 2000 美元。从各个注册代理获取 DOI 的费用也有所不同,中文 DOI 以科研单位数据集注册为例,每年年费为 500 元人民币;数据集 1~5000 条时,每个 DOI 为 0.5 元人民币,多于 5000 则每个 0.3 元人民币。DataCite 会员年费为 8500 欧元,可注册 DOI 数量没有限制。加州数字图书馆 EZID 对于研究机构和非赢利性组织的 DOI 注册年费为 2500 美元,每年最多可以创建维护 100 万个 DOI。德国科技图书馆 DataCite DOI 年费为 150 欧元,可注册 0~500 个 DOI,超出 500 的 DOI 按量收费,501~5000 每个 0.20 欧元,5001~5 万每个 0.15 欧元,5 万~10 万每个 0.10 欧元,10 万~100 万每个 0.06 欧元,超出 1 百万每个 0.03 欧元。

从以上数据可以看出,DOI 的费用相对较高,但一般情况下不需要维护本地服务系统;PURL 完全免费,既可以使用第三方提供的服务,也可以维护本地服务系统;ARK 既可以使用第三方提供的收费注册服务,也可以自行开发提供服务;Handle 费用较为低廉,需要本地维护服务系统。

(2) 使用量。

根据欧盟第 7 框架项目“科学网络记录持久获取联盟”(Alliance Permanent Access to the Records of Science in Europe Network,APARSEN)调查显

表 1 各持久标识符比较

特性 标识符	费用	使用量	数据安全	互操作	标识符与元数据
Handle	低	大	不提交元数据,信息较隐蔽	好	只提供标识符注册
DOI	高	大	需向注册代理机构提交元数据	好	同时提供标识符和元数据注册
ARK	自建免费,第三方服务需付费	一般	自建信息较隐蔽,使用第三方服务需提交元数据	好	同时提供标识符和元数据注册
PURL	免费	少	信息较隐蔽	一般	只提供标识符注册

示,DOI 使用率占 32%,Handle 占 28%,URN 占 25%,PURL 占 6%,ARK 占 4%^[20]。截至 2016 年 12 月,在 re3data 中注册的研究数据仓储中,DOI 使用占 21.9%,Handle 占 6.7%,URN 占 1.3%,PURL 占 1.0%,ARK 占 0.8%。

从这些数据可以看出,当前 Handle 及 DOI 使用量占据了优势地位。目前,DOI 注册个数已经超过 1.2 亿,分配前缀 1.6 万个,每年的解析量超过 50 亿。截至 2015 年末,Handle 分配的前缀数量达到 1.9 万个,每月代理解析量为 2.47 亿。

(3) 数据安全。

DOI 主要应用于数字出版领域,信息公开程度高,将元数据和标识符存放在注册代理机构不存在过多安全性考虑。然而,有些机构却包含了大量非公开出版物,或者不愿意将元数据信息公开,DOI 无法满足他们的需求。此外,DOI 管理机构 IDF 主要由欧美控制,出于国家层面数据安全性的考虑,一些机构也不愿意使用 DOI,如中国国家图书馆。

Handle 系统具有分布式管理的特性,数据隐蔽性好。美国洛斯阿拉莫斯国家实验室(Los Alamos National Laboratory)使用 Handle 系统管理了超过 6 亿件非公开内部资料。南加州大学利用 Handle 系统对病例数据及用户进行了标识,将数据访问权限、用户权限等作为 Handle 属性,基于 Handle 系统内置的安全协议完成用户认证、授权等过程^[13]。

ARK、PURL 可以使用第三方提供的服务,这时数据将会暴露给第三方。机构也可以利用 PURL 开源软件在本地运行服务或者自己实现 ARK 协议,因此也可以实现敏感数据的保护。

从上面分析可以看出,DOI 适用于信息公开程度高的资源标识,Handle 系统以及本地运行的 ARK、PURL 服务可用于需要一定数据安全保护的资源标识。

(4) 互操作。

DOI 的互操作性最好,各注册代理机构之间、注册代理机构与其他系统之间已经展开了许多互操作工作。ORCID 与 DataCite 互操作网络(ORCID and DataCite Interoperability Network,ODIN)为 DataCite 与 ORCID 之间的合作项目,它可以将 DataCite 中的研究数据集关联到 ORCID 的作者信息,关联的数据集可同步到 ORCID 系统,并显示在作者简历页面中^[21]。在 CrossRef 中,已有 230 多个出版社将

ORCID 包含在元数据中,CrossRef 已经支持对 ORCID 中作者出版信息的自动更新^[22]。CrossRef 与 DataCite 也表示双方将展开合作,增强两个系统之间的互操作性,支持文章与数据集之间的相互关联,开发开放 API 和开源工具支持文章与数据集之间关联关系的展示^[23]。与之相对应,Handle、PURL、ARK 的互操作项目却很少。

(5) 标识符与元数据。

Handle、DOI、PURL 协议对标识符本身限制较少,标识符中并无太多元数据信息。ARK 中 Qualifier 的使用,将资源的格式、版本、结构等元数据信息引入到标识符中,用户可以根据标识符的形式获得这些信息。然而资源本身可能发生变化,标识符的持久又使得不能对标识符做修改,这导致在标识符中引入过多元数据信息可能会引起不一致。标识符作为资源的唯一标识,仅需确保唯一性即可。标识符仅由少量字符串组成,过多引入元数据信息,会让用户难以理解。对于标识符所对应资源的描述,可以采用相应的元数据方案。元数据常常采用 XML 或者 JSON 格式展示,易于用户理解。更重要的是,当资源发生变化时,仅需更新存储的元数据信息,无需对标识符做修改。

从以上对各个持久标识符系统的比较分析,结合高校图书馆的实际需求,笔者认为 Handle 更加适合于高校图书馆持久标识应用。首先,Handle 以及基于 Handle 的 DOI 使用量最大,它得到持久维护的可能性高,从而持久性将会更好。其次图书馆属于非营利性机构,有着大量数字资源,Handle 相对开放、廉价的特性,更加符合图书馆的机构性质。再次,高校图书馆资源众多,其中不乏需要安全保护的资源,Handle 较好的数据隐蔽能够提供更好的信息安全保护。最后,DOI 基于 Handle,DOI 良好的互操作性也使 Handle 具有较好的互操作潜能,同时 Handle 对标识符本身限制较少,可以支持更加灵活的标识符构造。以前对 DOI 的担心主要来自国外技术垄断,目前句柄系统(Handle System)已经被交给 DONA 管理,中国 ETIRI—CDI—CHC 也已成 MPA,共同负责 Handle 的运营。此外,虽然 DOI 基于 Handle,但是它具有一定的商业性,资源标识的成本较高,适合于图书馆中具有首次公开发布的资源标识。

5 北京大学图书馆的应用需求

数字化资源的不断积累以及信息化建设程度的提高,让资源长期保存与系统间互操作变得越来越重要。北京大学图书馆拥有大量数字化的资源(如古籍、扫描加工的数字特藏等),目前图书馆信息化与数据中心正在对长期保存进行调研和试验,持久标识作为长期保存的重要组成部分也需建立和完善。图书馆正在积极建设“北京大学学术成果生态系统”(由机构知识库、期刊网、开放研究数据平台、学者主页四个部分组成)^[24],在系统建设的过程中,生态系统内部以及与外部系统(如科研管理系统)互操作的需求越来越迫切,如图1所示。在系统互操作的过程中,需要对各种实体(项目、学者、出版物、数据集)进行关联,这就要求实体能够被标识。各应用系统内部ID易于变化且具有局部唯一性,不能够胜任资源长期、唯一标识的要求,采用国际通用的持久标识符能够有效解决这一问题。

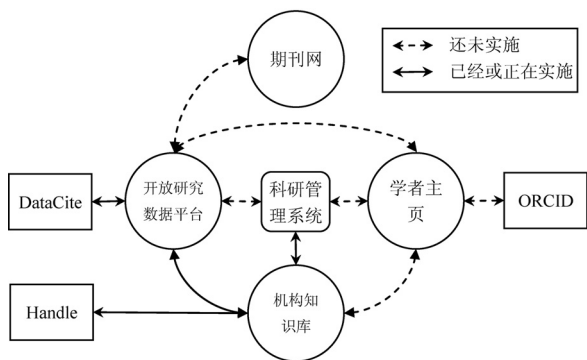


图1 北京大学学术成果生态系统及科研管理系统间互操作

6 北京大学图书馆的应用实施

基于前期开展的持久标识符应用调研,结合北京大学图书馆信息化建设需求,笔者所在团队先后对开放研究数据平台、机构知识库实施了资源的持久标识,并且基于资源的持久标识,实施了开放研究数据平台和机构知识库之间、机构知识库和科研管理系统之间的关联与互操作。

6.1 开放研究数据平台 DOI 标识

北京大学开放研究数据平台主要收录了北京大学产出的高质量研究数据,具有首次出版发布的特点,因此选择了在数字出版领域中的应用最广泛的DOI持久标识符。DOI有10个注册代理机构,分别专注于不同领域,虽然中文DOI也提供数据注册服务,但是与专注于数据集注册服务的DataCite相比,无论是在数据集注册的规模上,还是在相应的服务上都存在较大的差距,因此开放研究数据平台选择了DataCite服务^[25]。

DataCite服务的核心为元数据仓储系统(Meta-data Store)(图2)。元数据仓储系统主要负责元数据的管理,分配的DOI标识由Handle系统管理。元数据仓储的管理机制分为三层结构:管理员(ADMIN)、分配员(Allocator)、数据中心(Data Centre)。管理员即DataCite管理者,分配员主要负责各个国家或地区数据中心的创建,数据中心主要负责数据集的创建并为其分配DOI标识。

目前中国还不存在分配员,并且成为分配员的费用较高,因此开放研究数据平台选择了国外分配员。在实际调研中,主要联系调查了德国科技图书馆

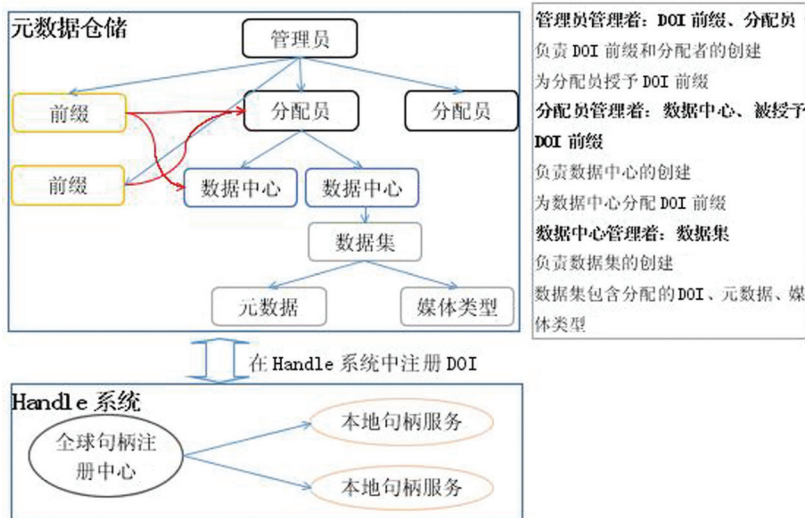


图2 DataCite元数据仓储系统结构

和加州数字图书馆。由于当前数据平台数据集的量不大,同时德国科技图书馆提供的服务能够在满足现有需求下花费最少,因此最终选择了德国科技图书馆作为分配员,北京大学图书馆成为其下的一个数据中心,分配的 DOI 前缀为“10.18170”。

在标识符注册实现方面,DataCite 元数据仓储提供网络界面的手动注册和表述性状态转移风格应用编程接口(Restful API)的自动注册方式。为了使得数据在正式发布时能够获得及时、便捷的 DOI 注册功能,开放研究数据平台采用了基于 Restful API 自动注册的功能,数据平台基于哈佛大学 Dataverse 进行二次开发构建而成,原先的 Dataverse 并不具有基于 DataCite 应用编程接口注册功能,为此进行了该模块的定制开发。自开放研究数据平台上线以来,得到了较大的关注,哈佛大学 Dataverse 团队进行了专门报道,并注意到北京大学开放研究数据平台采用了 DataCite 应用编程接口进行 DOI 注册,随即着手 DataCite DOI 注册功能的开发。

6.2 机构知识库 Handle 标识

机构知识库中收录的资源主要为期刊论文,考虑到一部分期刊论文已经具有 DOI 标识,并且绝大多数收录的资源均不具有首次出版发布的特点,同时机构知识库中资源数量众多,每年的增量较大,而单个 DOI 的注册费用较高,因此最终选择了 Handle 系统。

为了实施机构知识库中资源的持久标识符,对多种 Handle 应用软件进行了调研分析,主要包括 4 种方案。方案一:基于 CNRI 的 Handle. Net^[26];方案二:基于 CNRI 的 Handle. Net 和 ePIC API 系统^[27];方案三:基于 CNRI 的 Handle. Net 和 DataCite 元数据仓储^[28];方案四:基于 ETIRI-CDI-CHC^[29]系统。表 2 对各种 Handle 软件的特性进行了总结。

表 2 Handle 各应用软件比较

特性 软件	是否 开源	语言 类型	是否有元 数据管理	数据注册接口
方案一	是	JAVA	无	基于 HTTP 的 Handle 协议,有 Java 客户端库
方案二	是	JAVA、 RUBY	无	提供 RESTful API
方案三	是	JAVA	有	提供 RESTful API 和 Web 注册页面
方案四	否	未知	有	Web 注册页面

方案一:Handle. Net 基于 Java 实现了 Handle 协议,包含了 Handle 系统的核心功能。需要基于 Handle. Net 在本地搭建本地句柄服务(Local Handle Service, LHS),并使用 Handle 协议对 LHS 中的数据进行管理。DSpace 已经与 Handle. Net 进行了良好集成,其中的 Handle 插件(HandlePlugin)实现了 LHS 中 Handle 存储功能(HandleStorage),LHS 中数据的存储直接基于 DSpace 仓储。单独的 Handle. Net 虽然包含了持久标识符的核心功能,但并不包含便捷的 Web API,也没有元数据管理功能。

方案二:ePIC API 基于 Ruby 语言在 Handle. Net 基础上,包装了一层 RESTful API。该 API 使得持久标识符的注册管理可以基于 RESTful API 进行,相比直接使用 Handle 协议的标识符管理而言更加方便。ePIC API 为开源软件,提供了便捷的 Web API,但没有元数据管理功能。

方案三:DataCite 基于 Java 语言在 Handle. Net 基础上,构建了一套元数据仓储系统,使得持久标识符的注册管理可以基于 RESTful API 和 Web UI 进行,同时还可以对标识符所关联对象的元数据进行管理。DataCite 元数据仓储为开源软件。

方案四:ETIRI-CDI-CHC 提供付费标识符注册管理服务,包含了标识符注册和元数据管理功能,其费用约每年一万元左右。相比于从 CNRI 申请 Handle 前缀,其费用要高不少。

目前 Handle 由多个 MPA 管理,不同的 MPA 主要面向的领域不同。由于 ETIRI-CDI-CHC 主要面向物联网,而 CNRI 主要面向数字图书馆等领域。因此,北京大学机构知识库选择了 CNRI 进行 Handle 前缀注册,分配的 Handle 前缀为“20.500.11897”。考虑到当前机构知识库本身已经有元数据管理功能,并且机构知识库所基于的 DSpace 与 Handle. Net 软件有良好的集成。因此,北京大学机构知识库采用了方案一进行标识符的注册。需要说明的是,如果以后需要将更多的其他类型资源进行持久标识,并且需要进行元数据管理,则方案三能更好地实现该目标,并且机构知识库中已注册标识符也能够顺利迁移。

6.3 基于持久标识符的系统间互操作

目前,已经实现了开放研究数据平台与机构知识库间的互操作,机构知识库与科研管理系统间的

互操作正在积极实施。

在开放研究数据平台中,一些高质量数据集拥有多达数百的相关文献,数据发布者非常希望能够将这些相关文献进行展示,国外一些著名的研究数据管理服务已经提供该功能,例如密歇根大学的 ICPSR 便将数据集和出版物进行双向关联。为了实现这一目标,笔者所在团队采用的策略是:发挥机构知识库文献管理的优势,将相关文献存放在机构知识库中;机构知识库中所有文献、合集、社区均有全球唯一的 Handle 标识符,开放研究数据平台中的所有数据集均有全球唯一的 DOI 标识符,通过在数据集元数据中保存相关文献的 Handle,以及在相关文献的元数据中保存数据集的 DOI,实现了数据集与相关文献的双向关联;以 Handle 为标识,开放研究数据平台通过 Restful API 获取机构库中相应文献的元数据信息并进行展示。

机构知识库中存放了大量的科研产出,目前正在与科研管理系统进行融合,负责科研管理系统中学术成果的管理。在国外,科研管理系统特别重视对资源、学者等对象标识机制的支持。科研管理国际化标准“欧洲研究信息公共格式”(Common European Research Information Format, CERIF)从 1.5 版本开始便引入了标识符联盟(Federated Identifiers)实体,支持对 Handle、DOI、ORCID 等 20 余种标识符的识别和解析^[30]。一些国外大学也将科研管理系统与机构知识库进行了集成,如英国圣安德鲁斯大学。圣安德鲁斯大学使用爱思唯尔的 Pure 作为科研管理系统,在 Pure 中保存机构知识库中资源的 Handle,并对外进行展示,当用户需要下载全文时便可通过 Handle 链接进入到机构知识库中^[31]。北京大学科研管理系统主要管理着项目、人员、机构、成果等实体的信息,由于机构知识库已经拥有了成果管理功能,因此机构知识库将作为科研管理系统的成果子系统,并通过 API 实现系统间的互操作。机构知识库中成果采用了全球唯一的 Handle 进行标识,在与科研管理系统中的项目、人员等实体进行关联时也将以 Handle 为标识符,并基于此通过 API 获取成果信息。

6.4 应用效果

目前开放研究数据平台和机构知识库中所有资源均实现了持久标识。其中开放研究数据平台中注册 DOI 达到 110 个,机构知识库中注册 Handle 达

41 万多个,实现了以最小的成本满足现有资源持久标识的目的。开放研究数据平台与机构知识库间关联和互操作,满足了中国家庭追踪调查、中国健康与养老追踪调查等权威数据集相关文献发布的需求。机构知识库与科研管理系统的关联和互操作,使得科研管理系统更加与国际化标准接轨,提高了科研管理系统的可扩展性和适应性。

未来,希望将科研管理系统、开放研究数据平台、学者主页、期刊网、机构知识库进行更加完善的关联与互操作,最终的目标是实现用户能够从任意实体出发,找到所有与之相关实体的完整信息,同时确保各系统中实体信息的一致性。

参考文献

- 1 Arms C R. Historical collections for the National Digital Library: lessons and challenges at the Library of Congress[J]. D—Lib Magazine, 1996, 2(4).
- 2 Brase J, Lautenschlager M, Sens I. The tenth anniversary of assigning DOI names to scientific data and a five year history of DataCite[J]. D—Lib magazine, 2015, 21(1).
- 3 Sébastien P, Jean—Philippe T, John K. The ARK identifier scheme: lessons learnt at the BnF and questions yet unanswered [C]. Proceeding in the 1st Conference on Dublin Core and Metadata Application. Austin, 2014: 83—94.
- 4 British Library. Working with the British Library and DataCite: a guide for Higher Education Institutions in the UK [EB/OL]. [2016—12—28]. http://www.bl.uk/aboutus/stratpolprog/digi/datasets/WorkingWithDataCite_2013.pdf.
- 5 童忠勇,李志尧,孙秀萍. 国家数字图书馆数字资源唯一标识符系统的设计与实现[J]. 图书馆学研究, 2013(21): 53—58.
- 6 DONA. DONA foundation: summary of the minutes [EB/OL]. [2016—12—28]. https://www.dona.net/documents/public/8f990d1adf4d/Summary_of_Board_Minutes_July_3-4_2014.pdf.
- 7 Larry L. DONA foundation—administering the Global Handle Registry [EB/OL]. [2016—12—28]. <https://zenodo.org/record/31784/files/S1-Lannom.pdf>.
- 8 Hilse HW, Kothe J. Implementing persistent identifiers [M]. Consortium of European Research Libraries, 2006.
- 9 Zepheira. OCLC to work with Zepheira to redesign OCLC's PURL service [EB/OL]. [2016—12—28]. <http://zepheira.com/news/pr20070711/>.
- 10 Keith S, Stuart W, Erik J, et al. Introduction to persistent uniform resource locators [EB/OL]. [2016—12—28]. https://www.isoc.org/inet96/proceedings/a4/a4_1.htm.
- 11 IDF. International DOI foundation [EB/OL]. [2016—12—28]. http://www.doi.org/doi_handbook/7_IDF.html.

- 12 Kunze J. Towards electronic persistence using ARK identifiers [C]. Proceedings of the 3rd ECDL Workshop on Web Archives. Norway, 2003: 4-12.
- 13 郭晓峰, 孙洵. Handle 系统的发展及应用[J]. 数字图书馆论坛, 2013(8): 18-24.
- 14 ANDS. Identify my data[EB/OL]. [2016-12-28]. <http://www.ands.org.au/guides/identify-my-data>.
- 15 GWDG. ePIC - Persistent identifiers for eResearch[EB/OL]. [2016-12-28]. <https://zenodo.org/record/31785/files/S1-Schwardmann.pdf>.
- 16 周剑. Handle 互联网标识体系产业应用[EB/OL]. [2016-12-28]. <http://www.cspiii.com/sx/rdxw/2015-11-25-2384.html>.
- 17 Crossref. Annual report 15-16[EB/OL]. [2016-12-28]. http://www.crossref.org/Crossref_AnnualReport_2015-16.pdf.
- 18 中文 DOI. 中文 DOI 注册数量突破 2000 万了[EB/OL]. [2016-12-28]. <http://www.chinadoi.cn/portal/newsAction!detail.action?id=109>.
- 19 DataCite. DataCite metadata stats[EB/OL]. [2016-12-28]. <http://stats.datacite.org/>.
- 20 Lunghi M, Bellini E, Cirinna C, et al. Interoperability framework for persistent identifier systems[C]. Proceedings of the 9th international conference on preservation of digital objects. Toronto, 2012: 29-36.
- 21 ODIN. ORCID and DataCite interoperability network[EB/OL]. [2016-12-28]. <http://datacite.labs.orcid-eu.org/>.
- 22 Crossref. Crossref to auto-update ORCID records[EB/OL]. [2016-12-28]. <http://blog.crossref.org/2015/09/orcid-auto-update.html>.
- 23 CrossRef. CrossRef and DataCite announce new initiative to accelerate the adoption of DOIs for data publication and citation[EB/OL]. [2016-12-28]. <http://www.crossref.org/01company/pr/news111014.html>.
- 24 北京大学图书馆. 北京大学学术成果生态系统初步建成[EB/OL]. [2016-12-28]. <http://www.lib.pku.edu.cn/portal/news/0000001188>.
- 25 罗鹏程, 朱玲, 崔海媛等. 基于 Dataverse 的北京大学开放研究数据平台建设[J]. 图书情报工作, 2016(03): 52-58.
- 26 CNRI. Handle. Net software[EB/OL]. [2016-12-28]. http://handle.net/download_hnr.html.
- 27 ePIC. EPIC-API-v2[EB/OL]. [2016-12-28]. <https://github.com/pidconsortium/ePIC-API-v2>.
- 28 DataCite. Metadata store for DataCite central infrastructure[EB/OL]. [2016-12-28]. <https://github.com/datacite/mds>.
- 29 ETIRI-CDI-CHC. 国家物联网标识管理公共服务平台[EB/OL]. [2016-12-28]. <http://www.cndona.cn/joinupsys/instructionmanual/>.
- 30 Brigitte J. Bringing persistent identifiers into the CERIF data model[EB/OL]. [2016-12-28]. <http://dspacecris.eurocris.org/handle/11366/173>.
- 31 Anna C. CRIS-IR interoperability[EB/OL]. [2016-12-28]. <https://www.coar-repositories.org/files/Tutorial-CRIS-IR-Anna-Clements.pdf>.

作者单位: 北京大学图书馆, 北京, 100871

收稿日期: 2016 年 12 月 28 日

Exploring the Application of Persistent Identifiers in University Libraries

Luo Pengcheng Cui Haiyuan Nie hua Zhu Ling Wei Chengfu

Abstract: In order to identify the digital resources persistently in university libraries and to promote the interoperability between the systems, this paper has made an extensive research on persistent identifiers and its applications. Firstly, the paper introduced the four most widely used persistent identifiers and their applications at home and abroad, then made a comprehensive comparison and analysis from the aspects of cost, usage, interoperability, identifier and metadata. Finally, based on the practice of Peking University Library, this paper introduced the application of persistent identifier in the open research data platform, institutional repository and interoperability between the systems.

Keywords: Persistent Identifier; DOI; Handle; Library Resource; Unique Identifier