国际开放仓储目录整合研究与实践*

张云玲[1] 罗婷婷[1,2] 赵瑞雪[1,2] 鲜国建[1,3]
(1. 中国农业科学院农业信息研究所，北京 100081；2. 国家新闻出版署农业融合出版知识挖掘与知识服务重点实验室，北京 100081；3. 农业农村部农业大数据重点实验室，北京 100081)

摘要：开放仓储目录是对开放仓储的描述说明和索引，是开放学术资源利用、发现、共享的基础。本文首先通过对OpenDOAR、ROAR、BASE等5个国际主流开放仓储目录的建设现状进行调研分析，发现在国际开放仓储目录建设方面，还存在仓储目录收录不够完整、目录元数据项不够丰富、目录更新时效性有待提高、揭示系统功能相对单一等不足。在此基础上，本文提出开放仓储目录元数据整合研究，包括元数据描述规范设计、基于OAI协议和ETL工具收割元数据，使用数据清洗工具OpenRefine对元数据进行“形式去重”和OAI-Identify获取结果的“内容去重”，并建立对多源异构仓储目录进行匹配融合的方法路径，形成数据内容更丰富、数量更加全面的全球开放仓储目录GOAR核心集和扩展集。最后从建立动态更新融合机制、常态化监控机制和目录发布系统三方面提出下一步研究方向。

关键词：开放仓储；目录整合；OAI-PMH；元数据融合；开放获取
中图分类号：G250　DOI：10.3772/j.issn.1673-2286.2022.01.004
引文格式：张云玲，罗婷婷，赵瑞雪，等. 国际开放仓储目录整合研究与实践[J]. 数字图书馆论坛，2022 (1)：26-36.

20世纪90年代末，自由科学运动使越来越多的人意识到开放获取将在科学成果传播中发挥重要作用[1]。2002年布达佩斯开放获取倡议（Budapest Open Access Initiative）提出实现开放获取的两种途径：一是创办开放获取期刊（Open Access Journals，OAJ）；二是实行自我存档（Self-Archiving），即建立开放获取仓储[2]。2016年德国马普学会发起了“OA2020倡议”（OA2020 Initiative）旨在加速推动大规模学术期刊的开放获取[3]。2018年欧洲科研资助机构联盟发布开放获取S计划，目标是使学术出版物全面和即时的开放获取成为现实[4]。以上一系列的开放获取运动为开放科学的到来创造了条件。

当前，开放科学已成为世界各国一项重大科学战略和科学政策，被多个国家/地区或组织以路线图方式推进实施[5]。2012年全欧科学院ALLEA（All European Academies）和欧盟发布《关于21世纪开放科学的联合宣言》呼吁科学界采取果断措施，将开放科学和创新作为一种手段，加速发现应对重大社会挑战的解决方案[6]。2018年欧盟发布《欧洲开放科学云实施路线图》，旨在通过数据基础设施、科研数据管理、开放科学参与规则和治理框架等6条行动路线的实施推动开放科学运动并确保欧洲在数据驱动型科学领域处于领先地位[7]。2021年11月联合国教科文组织（UNESCO）发布开放科学建议书，首次从开放科学的定义、核心价值、行动领域、政策机制监测等方面为开放科学政策和实践提供了一个系统、权威的国际框架[8]。同时我国也在积极推动开放科学运动。2016年中国科学院启动“科技论文预发布平台China Xiv”项目，期望通过推动科研成果的开放获取构建新型科研成果交流和共享平台[9]；2018年国务院办公厅发布《科学数据管理办法》。以上一系列活动、宣言都揭示出科学研究发展过程中的必然趋势，即“走向开放化”[10]。

开放学术资源是开放科学的物质基础。在以大数据为基础的数据密集型科学研究范式下，越来越多的

* 本研究得到中国农业科学院科技创新工程项目（编号：CAAS-ASTIP-2016-AII）资助。

科研人员关注开放学术资源的共享和再利用，开放仓储是开放学术资源有效管理与发现利用的重要途径。在过去十多年中，开放仓储数量迅速增长，面对海量的开放仓储，科研人员检索到所需开放仓储的难度越来越大，为了更加有效地查找和利用开放仓储，开放仓储的注册目录系统应运而生。目前国际上已出现较多覆盖面广的开放仓储目录如OpenDOAR（Open Access Directory）、ROAR（Open Access Directory），因其数据来源的不同各目录系统在平台功能、学科范围等方面都呈现出各自的特点，与当前用户需求相比较，还存在收录不够完整、目录元数据项不够丰富、目录更新时效性有待提高、揭示系统功能相对单一等不足。为此，本文拟开展全球开放仓储目录整合研究与实践，力争为开放学术仓储资源的进一步高效利用、发现、深度整合，以及促进面向开放科学的数字资源基础设施建设提供参考。

1　国内外研究现状

目前，国内外对于仓储目录系统的研究主要集中在两个方向，开放获取仓储目录和科研数据仓储目录，前者的研究对象是开放获取仓储OAR（Open Access Repository），即实现开放获取的绿色道路（Green Road），后者的研究对象是科研数据仓储RDR（Research Data Repository），二者的仓储覆盖范围互相交叉。本文重点研究开放获取仓储，即开放仓储，同时参考科研数据仓储目录系统的相关研究，为后续的目录整合工作提供更加开阔的视野。

在国内，有关仓储目录系统的研究主要集中于对国外各领域具有代表性的仓储目录案例进行介绍。王翠萍等[11]选取5个仓储目录系统（re3data、OAD、OpenAIRE、OpenDOAR、ROAR），从资源收录量、检索功能、软件应用情况等方面介绍仓储目录建设现状；张莎莎等[12]以re3data为数据源从责任主体、平台功能、数据资源、数据传输4个方面分析总结英国科研数据发布平台的特点及建设经验；刘峰等[13]从建立时间、国别、学科领域、开放程度等方面对科研数据仓储目录Databib进行统计分析；夏姚璜[14]对re3data的标签检索和图标符号体系进行介绍并对比分析中美两国仓储建设特点，提出我国应该更加重视科学数据仓储注册目录建设，建立本国的目录体系；管凤贞等[15]对OpenDOAR注册的中国机构仓储建设现状进行梳理，指出我国机构仓储存在的全球可见性、可访问性、政策保障、知识服务方面的问题；杨丽娜等[16]对OpenDOAR中资源环境领域开放仓储的基本情况和资源特征进行分析总结；郑一波等[17]通过元数据获取采集、转换映射、集成融合与质量控制等构建支持资源发现、定位和获取的新型联合目录体系，发挥资源集成优势，促进文献资源的共享利用。除此之外，近年来国内对于开放学术资源建设的实践研究也取得了很多成果，包括中国科学院建设的GoOA开放期刊集成服务系统和OAinONE自然科学领域开放学术资源一站式检索发现平台[18]，以及中国医学科学院信息研究所开发建设的基于世界卫生组织西太平洋地区医学索引[19]实现的多源期刊元数据汇聚探索实践。

在国外，有关仓储目录系统的研究主要围绕对全球、各地区仓储发展现状进行介绍。Pinfield等[20]以OpenDOAR为数据源，从国家/地区、仓储类型、开发软件、开放协议类型等方面介绍2005—2012年全球开放仓储的发展情况，对比分析OpenDOAR和ROAR在数据管理方式上的区别；Summann等[21]从数据收集与预处理、数据存储、可视化工具等方面对全球仓储监测平台BASE（Bielefeld Academic Search Engine）进行介绍；Hitchcock等[22]从数字资源长期保存的角度提出基于OAI的机构仓储服务提供者模型，并强调相关保存政策的重要性；Abdullah[23]同样以OpenDOAR为数据源，介绍亚洲高校的开放获取仓储建设现状，分析各国家、机构的仓储在全球的可见度和影响力；Bhardwaj[24]以re3data为数据源对全球开放科研数据仓储的国家分布、元数据标准、数据开放协议等内容进行分析总结。

综上所述，国内外对于仓储目录系统的研究大多以某个仓储目录为数据源探讨某一国家、某一学科的开放仓储建设现状和存在问题，抑或是对学术期刊或文献资源融合汇聚的研究实践，缺少较全面系统的关于开放仓储目录的对比研究和进一步整合实践，这对于科研用户选择满足特定科研实践需求的仓储和开放学术资源的一体化集成发现与深度利用共享都造成了不便。

2　开放仓储目录调研分析

当前，开放科学实践在全球得到了前所未有的快速发展，开放获取资源日益膨胀，全球科学领域兴起了大量科学社交网络平台，形成了不同类型的平台模式[25]。例如，专业性科学数据仓储目录平台re3data、Dataportals、Databib，政府或研究机构门户站点建立

的注册目录平台美国政府开放数据目录、美国能源部开放数据目录，综合性开放仓储目录平台OpenDOAR、OAD、ROAR等。本文选取的开放仓储目录均支持OAI-PMH互操作协议，其中Illinois大学图书馆OAI-PMH Data Provider Registry和OAI官方OAI-PMH Registered Data Providers是创建历史较长的两个开放仓储目录，目录收集建设方式和系统功能设置都相对传统。同样位于英国的两个开放仓储目录OpenDOAR和ROAR，是全球较为领先的目录检索系统，学科领域覆盖面广、资源类型多样性强。BASE（Bielefeld Academic Search Engine）在开放仓储目录索引基础上对学术文档进行整合，是世界知名学术搜索引擎。以上5个开放仓储目录呈现了开放仓储目录系统的不同发展形态，十分具有代表性，下文将从基本概况、资源收录情况、目录系统功能、目录技术选型等方面对它们进行对比分析。

2.1 基本概况

开放仓储目录OpenDOAR是由英国诺丁汉大学（the University of Nottingham）和瑞典兰德大学（Lund University）图书馆于2005年2月共同创建的开放获取仓储、学科资源库目录检索系统，是全球OA仓储权威目录网站，与姐妹工程DOAJ形成有效分工，OpenDOAR以OAR资源为对象，DOAJ则针对OA期刊，两者覆盖全部OA资源。截至2021年12月，在OpenDOAR登记的仓储已经有5 794个。

开放仓储注册平台ROAR是英国南安普顿大学（University of Southampton）主办的开放搜索国际数据库，是开源数字仓储平台Eprints的一部分。它对开放仓储及其内容的创建、位置和增长进行索引，通过及时提供有关世界各地仓储的增长和状态信息来促进开放获取的发展。从2003年建立至今，已经有5 386个仓储在ROAR注册。

OAI-PMH Data Provider Registry由Illinois大学图书馆的托马斯哈宾教授负责维护，收集不同来源中OAI兼容仓储中的Identify、ListSets、ListMetadataFormats和示例记录，将数据添加到数据库中，编制索引使其可浏览和搜索，同时会定期从OAI官方列表中更新数据，将新发现的开放仓储加入。截至2021年12月收录了5 247个开放仓储。

OAI官方OAI-PMH数据提供者注册表Registered Data Providers是由数据提供者自行注册的开放仓储目录系统，是很多开放仓储目录系统的数据源，截至2021年12月OAI官方收录的仓储有5 395个。每个仓储都提供仓储名称（Repository Name）、base URL和OAI标识符命名空间等注册记录，同时可以发出Identify请求以XML格式返回开放仓储相关描述信息。

BASE由德国比菲尔德大学图书馆于2004年创建并负责营运，专注于开放学术网络资源，BASE基于大量的OAI开放接口，实现了海量元数据的采集获取、标准化处理和索引发布服务。截至2021年12月底，BASE共有9 300多个资源目录来源。

2.2 仓储目录对比分析

2.2.1 资源收录情况

资源收录量和资源收录范围是开放仓储目录建设水平的重要考量标准。OpenDOAR将开放仓储分为机构仓储、学科仓储、政府仓储和聚合仓储4种仓储类型和12种资源类型，一个开放仓储可以涵盖多个资源类型但只能属于一种仓储类型，68%的开放仓储资源类型中均包含期刊论文[24]，89%的开放仓储是机构仓储。ROAR中对开放仓储类别的关注角度与OpenDOAR不同，未对开放仓储的资源类型和仓储类型进行严格区分，将开放仓储分为机构仓储、研究数据、开放和关联数据等9种仓储类型，一个开放仓储只属于一种仓储类型，其中机构仓储占79%以上。BASE整合来自Datacite、CiteSeerX、PubMed Central等多个来源的2.7亿多份学术文档，按照文件类型分为期刊论文、专利、数据集等20余类学术资源，其中60%的资源状态为开放获取，为了解决不同仓储中文件类型不统一的问题，BASE通过将文件类型映射到由数字代码标识的一致类别中来对其进行统一。

此外，不同开放仓储目录系统在对开放获取仓储进行学科分类时采用的标准也不同。OpenDOAR的学科分类来源于英国高等教育资助委员会HEFCE研究评估系统RAE system的UOA分类，将资源分为29个学科，包括心理学、教育学、农业兽医及食品科学等，由于每个仓储收录的资源内容往往涉及多个学科，所以一个开放仓储的学科分属于29类中的若干类别[26]。ROAR的学科分类按照美国国会图书馆分类法，BASE支持主题分类的杜威十进制分类法（DDC）。Illinois大学图书馆OAI-PMH Data Provider Registry和OAI官方OAI-

PMH Registered Data Providers的开放仓储目录没有对仓储进行分类，仅提供一般信息。

2.2.2 目录系统功能

OpenDOAR提供基本搜索和高级搜索功能，并且无须键入便可根据国别进行浏览。在基本搜索中用户可通过键入仓储名称跳转到详细信息页面，高级搜索功能中，用户可通过仓储类型、主题领域、开发软件、国别等8种途径交叉查询，搜索结果可按国别或字母顺序进行排序，下拉框中还可以选择满足任一条件或满足所有条件的模糊或精确搜索。ROAR除了提供与OpenDOAR相同的8种交叉查询途径外，还提供了ROAR ID、创建时间等5种过滤方式，并且ROAR的浏览方式在OpenDOAR基础上增加了仓储类型、年份、软件3种。OpenDOAR对开放仓储详细信息的介绍从开放仓储基本信息、机构信息、开放获取相关政策3个模块分类展示。ROAR中除了对开放仓储基本信息进行描述外，还增加了可视化展示，包括开放仓储网站首页缩微图和开放仓储活跃度曲线。

BASE系统提供丰富的浏览功能，可从杜威十进分类法DDC、文献类型、重用许可类型、获取方式和数据来源等维度进行海量学术资源的快速浏览与定位。在BASE中可使用两种不同的搜索界面，提供单个搜索字段的基本搜索（默认情况下在文档的所有部分中搜索）以及具有多个搜索字段和更复杂搜索选项的高级搜索。Illinois大学图书馆OAI-PMH Data Provider Registry界面提供一个简单搜索框，用户键入任何字段将与仓储名称模糊匹配。OAI官方OAI-PMH Registered Data Providers无搜索功能。

2.2.3 目录技术选型

采用标准的数据规范和数据政策有利于更加科学地管理开放仓储目录系统，本研究从数据规范、仓储开发软件角度分析开放仓储目录系统的技术选型。目前 OpenDOAR列出开放仓储的资源提交政策、资源内容政策、长期保存政策、元数据再利用政策以及全文再利用政策等5个方面的政策[27]，对于所有的政策，分别给予“未知”“未陈述”“未定义”“未明确”和“已定义”5 个等级；对于元数据再利用政策和全文再利用政策，则还有“禁止再利用”“不允许自动获取”“不稳定”“允许非商业用途”和“允许商业用途”5个等级[28]，并且在OpenDOAR中仓储最常用的协议是OAI-PMH、RSS、ATOM、SWORD[20]。

在OpenDOAR和ROAR中40%以上的开放仓储使用Eprints和DSpace，二者也是最早支持OAI协议的开发系统平台。Erpints是南安普顿大学开发的开源软件，也是在学科仓储建设中使用最多的平台；DSpace是由美国麻省理工学院和惠普公司合作开发的面向机构仓储的系统软件，也是目前知名度最高的自存档平台[28]。OpenDOAR和ROAR使用率前三的软件分别还有WEKO和Bepress，WEKO是一款日文开源仓储软件[29]，日本的机构仓储除了个别使用DSpace外其他大多都使用WEKO，这反映了日本机构仓储建设在国家开放获取政策和日本国立情报研究所的全方位技术支持下国际影响力逐渐增强。ROAR中排名第三的Bepress是商业机构仓储系统[30]，用户分布主要集中在美国。其他开发软件因语种限制仅在某些国家使用，如dLibra仅在波兰使用，OPUS仅在德国使用。

2.3 总结分析

通过调研发现和对比分析发现（见表1），从开放仓储目录收录的国别分布看，各目录收录的美国、英国、德国等发达国家的开放仓储数量差别不大，差别主要体现在开放科学浪潮下亚洲、南美洲地区新建的开放仓储数量大幅增加。从开放仓储对元数据项的分类维度看，OpenDOAR的分类相较于其他仓储更加规范化和多元化，从不同角度对开放仓储进行揭示，而ROAR分类更加关注科研数据、开放数据集、关联数据等数据层面的资源类型。各开放仓储目录的系统功能因目录系统建设成熟度和重点关注的元数据项不同而呈现出不同的特色，OpenDOAR、ROAR和BASE的系统功能非常相似，但OpenDOAR更关注的是全球开放仓储在不同学科不同区域的分布情况和发展程度，所以在系统中通过可视化图表进行不同维度的展示。ROAR则更加关注对全球开放仓储的活跃度监测，在每个仓储的描述中都添加了活跃度曲线。BASE已经具备成熟的目录发布体系，每年都会发布一个全球开放仓储发展情况的统计报告，用户可根据需要下载PDF。

在开放仓储目录更新维护方面，OpenDOAR的记录大部分是由人工创建和维护的，凡是在仓储目录平台列出的仓储都会经过OpenDOAR团队审核[31]，所以

OpenDOAR质量更高、更权威。ROAR的记录基于自动收割获取，资源体量更大、及时性更强，缺点是会收割到数量相当的无效站点，并且在2019年底ROAR由于控制存储器发生重大故障导致无法正常对内容增长进行跟踪，至今尚未恢复。OAI官方的OAI-PMH Registered Data Providers由于缺少专业团队维护，该目录中很多仓储网站已经无法访问或直接链接到外部网站。综上所述，各开放仓储目录存在的不足主要体现：在开放仓储收录范围与数量上存在互补性，如日本、中国和法国的开放仓储收录数量在仓储目录OpenDOAR、ROAR和ROAR中差距明显；各开放仓储目录平台关注的元数据项不尽相同，同一个开放仓储在不同目录系统中元数据项丰富程度不同，在OpenDOAR中有对开放仓储的仓储类型和资源类型的描述，在ROAR中没有资源类型的描述但会关注订阅方式；各开放仓储目录平台具备的功能参差不齐，如表1系统功能中除基本检索和高级检索外的各目录系统独有的功能；部分开放仓储目录平台的数据更新缺乏及时性，如Illinois大学图书馆OAI-PMH Data Provider Registry缺乏维护更新，大部分地址已经失效或发生变更。

表1 各开放仓储目录对比分析

仓储目录	收录规模	元数据项	系统功能	更新维护
OpenDOAR	日本681个 中国64个 法国160个	仓储类型RepositoryType 资源类型ContentType	基本检索 高级检索 可视化图表	记录大部分是由人工创建和维护的，而不是通过自动采集，以确保高水平的质量
ROAR	日本198个 中国96个 法国96个	仓储类型RepositoryType 订阅方式RssFeed	基本检索 高级检索 活跃度曲线	记录的创建是基于自动收割程序，具有即时性，但缺点是它往往会收集到相当数量的无效站点
Illinois大学图书馆OAI-PMH Data Provider Registry		仅提供仓储名称RepositoryTttle和OAI地址	提供开放仓储元数据格式MetadataFormats统计	大部分地址已经失效或发生跳转
OAI官方OAI-PMH Registered Data Providers		仅提供仓储名称RepositoryTttle和OAI地址	无系统功能设置	数据提供者自行注册添加
BASE	日本615个 中国86个 法国219个	仓储类型RepositoryType	基本检索 高级检索 年度报告Statistics for 2019	专业人员审核维护

3 国际开放仓储目录整合

3.1 整合目标与总体思路

开放仓储目录是对开放仓储的描述说明和索引，可以帮助科研人员定位、查找开放仓储。开放仓储目录的质量和覆盖范围决定了开放学术资源整合、利用的深度和广度，如前所述开放仓储目录在收录范围、资源分类、系统功能等方面各具特色，如果将同一开放仓储在不同开放仓储目录的元数据描述汇总，就可以从不同角度揭示开放仓储。本研究将基于五大开放仓储目录进行集成整合，并在此基础上进一步发现和增补其他仓储目录，构建一个收录范围更全面、描述内容更丰富的目录，为下一步实现对开放仓储的活跃度进行统计分析，并将静态的开放仓储目录列表上升为对开放仓储动态的监控，建立集用户检索、查找、多维度发现、利用的“一站式”开放学术资源服务奠定基础[17]。开放仓储目录的整合思路是开放仓储目录建设的重要依据，指导整个仓储目录建设工作。在实践探索过程中，主要围绕元数据描述规范设计、元数据采集与查重、元数据整合与映射等步骤确定整合流程，整合后的目录命名为全球开放仓储目录（GOAR，Global Open Academic Repository），如图1所示。

3.2 元数据描述规范设计

元数据是关于数据的数据，是描述一个具体的资源对象，并能对这个对象进行定位、管理且有助于它的

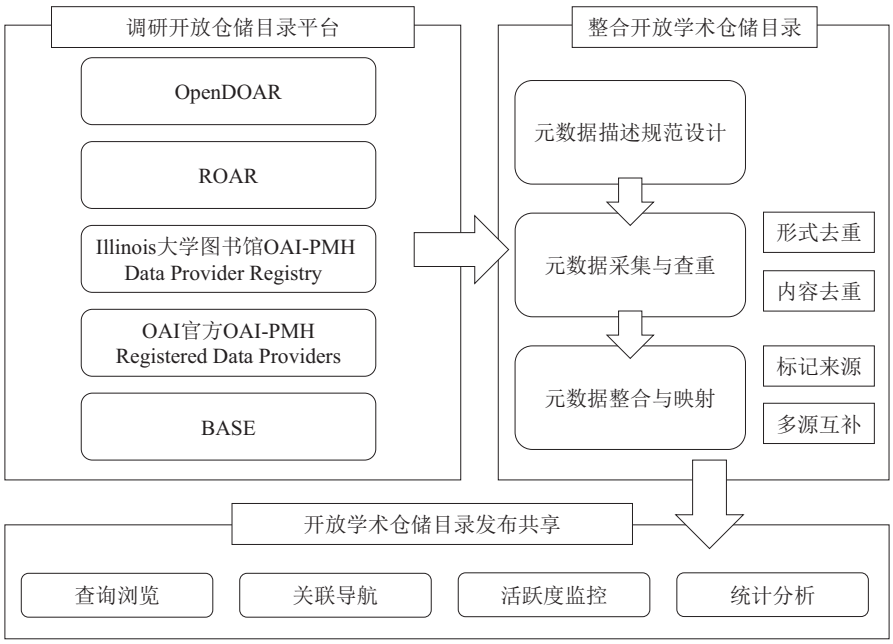

图1　开放仓储目录整合思路

发现与获取的数据[32]。开放学术仓储元数据是对开放学术仓储外部特征和内容特征的描述与揭示，其元数据质量在很大程度上影响着开放学术资源的可发现性、可利用性。元数据描述[33]，即采用一定的元数据规范来描述数据的组织结构和内容特征，从而实现对大量数据的高效管理。在前期调研过程中发现开放仓储目录OpenDOAR和ROAR的元数据描述较为规范、全面，因此本研究有选择地复用OpenDOAR和ROAR元数据集中的32个元数据描述元素。其中涉及的元数据元素可以大致划分为仓储信息、记录信息、机构信息、开放获取政策信息和溯源映射信息五大类。仓储信息包括仓储名称、仓储替代名称、仓储类型、OAI地址、资源类型、学科、开发软件、软件版本、记录数、全文数量、国家、地区、经度、纬度等，其中仓储替代名称主要面向小语种仓储名称的多语种表达，以保证用户查询的精准度[34]；记录信息包括记录创建时间、记录最后更新时间；机构信息包括机构名称和机构地址；开放获取政策信息包括是否开放获取、重用条件、是否允许机器收割、元数据政策地址等；溯源映射信息包括目录来源名称、目录源ID和GOAR中仓储ID，其中SourceName与SourceID一一对应，分别是开放仓储整合前的名称与ID，也是与源开放仓储目录系统进行映射溯源的标识，可以通过“https://v2.sherpa.ac.uk/id/repository/***”和“http://roar.eprints.org/***/”映射到唯一的开放仓储对象，GOAR_ID则作为新记录的唯一标识。如表2所示，序号1~19为仓储目录基础信息，20~21为仓储建设的机构信息，22~27为开放获取政策信息，28~29为仓储目录管理信息，30~32为溯源映射信息。

3.3　元数据采集与查重

元数据获取，是仓储目录建设的首要环节。本研究基于OAI元数据互操作协议对OpenDOAR、ROAR、Illinois大学图书馆OAI-PMH Data Provider Registry、OAI官方OAI-PMH Registered Data Providers和BASE五大开放仓储目录源进行元数据收割，利用开源ETL工具Kettle配置元数据采集工作流，定期对开放学术仓储元数据进行采集获取，并载入本地元数据仓储。截至2021年底，五大开放仓储目录共收割元数据29 551条，其中72.5%的数据带有OAI地址，后续整合将围绕有OAI地址的元数据构建GOAR开放仓储目录核心集，而OAI地址缺失的目录将通过人工筛选补充，并结合其他关键字段进行查重融合后构建形成GOAR开放仓储目录扩展集。

由于采集到的元数据来自不同的数据源，遵循的元数据标准不同，同一仓储在不同数据源中的元数据项表述也不同，同时不同来源的数据从字段上或记录上具有互补性，所以元数据融合主要是把不同来源的

表2 GOAR元数据描述规范

序 号	元 素	元素定义	序 号	元 素	元素定义
1	RepositoryTitle	仓储名称	17	City	城市
2	Repository_AlterTitle	仓储替代名称	18	Latitude	纬度
3	RepositoryType	仓储类型	19	Longitude	经度
4	RepositoryURL	仓储地址	20	OrganisationName	机构名称
5	OAI_PMH	OAI地址	21	OrganisationURL	机构地址
6	ContentTypes	内容类型	22	OA	是否开放获取
7	Subject	学科	23	ReuseConditions	重用条件
8	Software	开发软件	24	RobotHarvesting	是否允许机器收割
9	SoftwareVersion	软件版本	25	NotForProfitReuse	是否是非营利性重用
10	RecordCount	记录数	26	MetadataPolicyURL	元数据政策地址
11	Fulltext	是否支持全文	27	WithdrawnItemSearchable	撤销项目是否可搜索
12	FulltextCount	全文数量	28	CreatedDatetime	记录创建时间
13	Languages	语言	29	LastUpdatedDatetime	记录最后更新时间
14	Description	描述	30	SourceName	目录来源名称
15	Region	地区	31	SourceID	目录源ID
16	Country	国家	32	GOAR_ID	GOAR中仓储ID

元数据进行去重，保证同一仓储的多条重复元数据记录能够聚合归并为一条完整的记录[35]，同一仓储在不同数据源中元数据项能够相互补充，实现仓储元数据记录的唯一性和仓储元数据内容的丰富性。需要注意的是，在仓储元数据去重过程中应该保留原始记录，相互补充时应该标记相应的来源标签，这样有利于后期开放学术仓储元数据的维护和溯源。

开放学术仓储元数据的去重工作主要围绕“形式去重”和“内容去重”两个层次展开。“形式去重”可利用免费开源的数据清洗工具OpenRefine（又称GoogleRefine）对开放学术仓储元数据的OAI地址（OAI-PMH）和仓储名称（RepositoryTitle）两类元数据项进行相同空白填充、归类查重。“内容去重”一方面是按照一定周期利用OAI协议的Identify命令对开放学术仓储的基地址（baseURL）、仓储名称、状态（status）等基本信息进行跟踪采集，对已经发生更改的记录进行标记，保证开放仓储目录记录的动态性。HAL是法国最重要的国家开放获取仓储，它不仅收录了法国已发表的科研文献和未发表的科研预印本，也是欧盟多个开放科学项目的资源和元数据提供方[36]，在“内容去重”过程中发现有21个不同的基地址对应的开放仓储名称均为HAL，进一步查看通过OAI-Identify命令返回的结果，21个原始基地址已经整合为一，然后使用OAI-ListRecords命令检验这些开放仓储中的内容是否一致，完全相同即将这21个开放仓储进行标记归一。“内容去重”的另一方面是针对仓储元数据由于反复录入、输入错误或同一仓储在多来源中表示不一致造成的冗余（见表3、表4），最常见的就是不同仓储目录系统对于仓储名称中符号的处理，此类重复值可通过OpenRefine的指纹分类算法进行检测、合并。

经过对开放仓储元数据OAI地址和仓储名称的形式去重和内容去重处理后，开放仓储元数据由原始的29 551条整合为11 660条，构成GOAR开放仓储目录核心集。同时，在元数据“内容去重”的实践探索中笔者发现了一些存在的问题。例如，伊利诺伊大学的开放仓储目录系统Illinois大学图书馆OAI-PMH Data Provider Registry中，一部分OAI地址的请求已经无法响应，进一步分析发现，这些开放学术仓储本身是从开放学术期刊集成平台独立出来形成的网址门户，现已终止运行或已全部整合交由政府平台进行维护，这也从侧面证实了各国政府机构对开放仓储的建设越来越重视，如Illinois大学图书馆OAI-PMH Data Provider Registry中OAI地址前缀为IMLSDCC的114个开放仓储均是独立的期刊平台，现全部失效，已整合到https://www.imls.gov/。对于类似这样的失效数据，会在状态（Status）中专门标记，以便后续人工再进行跟踪关注，确保提交目录的有效性和时效性。

表3 RepositoryTitle重复情况的样例

RepositoryTitle重复	形 式	样 例	数据源
由于数据反复录入，导致同源重复	-	-	-
同一实体在不同数据源中表示的不一致	竖杠	Scholarship，Research，and Creative Work at Bryn Mawr College | Bryn Mawr College Research Scholarship，Research，and Creative Work at Bryn Mawr College	ROAR OpenDOAR
	空格	Scholarly Commons @Baptist Health South Florida Scholarly Commons @ Baptist Health South Florida	ROAR OpenDOAR
	横杠	Scholarly Commons University of the Pacific Scholarly Commons - University of the Pacific	OpenDOAR ROAR
	@	Virginia Tech Computer Science Technical Reports Computer Science Technical Reports @Virginia Tech	ROAR OpenDOAR
	冒号	Simon Fraser University：Institutional Repository Simon Fraser University Institutional Repository	ROAR
	大小写	Pandemos PANDEMOS	ROAR OpenDOAR
	语种符号	Tesis Electronicas de la Universidad de Chile Tesis Electrónicas de la Universidad de Chile	OpenDOAR ROAR
	命名方式	Digital Repository of The University of Toledo The University of Toledo（UT）：Digital Repository	ROAR BASE

注：“-”表示无此项

表4 OAI-PMH重复情况的样例

OAI-PMH重复	形 式	样 例	数据源
表示不同但实际指向同一实体	末尾多斜杠	http://scholar.uwindsor.ca/do/oai http://scholar.uwindsor.ca/do/oai/	OpenDOAR ROAR
	下划线替代冒号	http://cemi.socionet.ru/oai/ecoorg_org1/oai.cgi http://cemi.socionet.ru/oai/ecoorg:org1/oai.cgi	ROAR Illinois Registry
	横杠替代斜杠	http://www.hirsla.lsh.is/lsh/oai/request http://www.hirsla.lsh.is/lsh-oai/request	Illinois Registry ROAR
输入错误	中间多斜杠	http://archive.nyu.edu//request http://archive.nyu.edu/request	OpenDOAR ROAR
安全链接	http与https	http://scholarworks.smith.edu/do/oai/ https://scholarworks.smith.edu/do/oai/	ROAR OpenDOAR

3.4 元数据整合与映射

由于各开放仓储目录系统在元数据信息描述详尽程度、重点描述维度等都存在差异[37]，因此，在元数据融合过程中还需要对多来源元数据的元数据项，即描述字段进行互相补充，目的是形成包含字段信息较为丰富的厚元数据，细化资源揭示粒度[38]，保障开放仓储目录的建设质量，为后续数据分析、数据挖掘等数据增值服务奠定基础。经过查重处理后的部分仓储元数据仍然存在关键字段值缺失问题，如OAI地址缺失会导致后期无法正常通过OAI命令获取信息，针对该问题笔者团队利用搜索引擎等渠道进行了人工追踪、筛选、补充，但仍有部分OAI地址无法获取，对此类仓储元数据予以保留并将其与核心集元数据进行整合、查重，构成GOAR开放仓储目录扩展集，最终得到包括核心集在内的开放仓储共15 903个。

在开放仓储基本元数据描述信息完整的基础上，对各来源开放仓储元数据进行整合。具体而言，采取质量优先原则，在五大源开放仓储元数据整合时首先以质量最为可靠的OpenDOAR作为重点优选来源和首选元数据作为主数据入库，再辅之ROAR、BASE与OpenDOAR来源的目录元数据项进行字段级补充融合，最终形成一条丰富完整的厚元数据记录[39]。在SourceName中保存来源目录名称，SourceID中保存源目录ID。多个源目录名称和源目录ID之间用英文分号隔开且一一对应，以便建立与其他仓储目录的关联和溯源。进一步分析经过整合后的GOAR开放仓储核心集11 660个仓储元数据发现，其中5 864个仓储独立来自上述五大来源的单一来源，而同时被2个、3个、4个和5个目录收录的仓储数量分别是3 447个、1 690个、520个和139个。由此也体现出本研究开展多来源仓储目录整合和映射关系的必要性和实际价值。此外，47%的元数据提供了描述信息Description，包括对开放学术仓储内收录的资源类型、学科覆盖范围、浏览界面支持语言、是否支持RssFeeds内容更新提醒订阅服务等内容，这些信息对于利用开放学术资源开展知识服务很有价值。如表5所示，经过OpenDOAR、ROAR与Illinois大学图书馆OAI-PMH Data Provider Registry 3个来源的开放仓储“Academica-e”元数据整合，对OAI地址进行验证保留质量较高的OpenDOAR基地址并最终形成一条新的厚元数据记录，不仅给出开放仓储名称、网站地址、仓储类型，而且揭示了开放仓储的学科覆盖、资源类型、界面支持语言、经纬度等细粒度描述信息，最后还继承了在OpenDOAR和ROAR中对应的溯源ID信息。

表5 多来源元数据整合样例

元数据项 仓储目录	ROAR	OpenDOAR	Illinois Registry	融合新增（GOAR）
RepositoryTitle	Academica-e	Academica-e	Academica-e	Academica-e
OAI_PMH	http://academica-e.unavarra.es/oai	http://academica-e.unavarra.es/oai/driver	http://academica-e.unavarra.es/oai/request	http://academica-e.unavarra.es/oai/driver
RepositoryType	institutional	institutional	-	institutional
OrganisationURL	http://www.unavarra.es/	http://www.unavarra.es/	-	http://www.unavarra.es/
RepositoryURL	http://academica-e.unavarra.es/	http://academica-e.unavarra.es/	-	http://academica-e.unavarra.es/
RegistryID	4 466	2 347	-	goar_2246
Description	This site provides access to the research output of the institution. The interface is available in Spanish	This site provides access to the research output of the institution. The interface is available in Spanish	-	This site provides access to the research output of the institution. The interface is available in Spanish
Region	-	Europe	-	Europe
Country	ES	ES	-	ES
Subject	-	Science General > Agriculture, Food and Veterinary Technology General Technology General > Computers and IT	-	Science General > Agriculture, Food and Veterinary Technology General Technology General > Computers and IT
SourceName	-		-	ROAR; OpenDOAR
SourceID	-		-	4 466; 2 347
Languages	-	Spanish English	-	Spanish English
Location	-	42.805 9, -1.629 9	-	42.805 9, -1.629 9

注：“-”表示无此项

4 总结与展望

我国已提出“十四五”时期要加快构建国家科研论文和科技信息高端交流平台。开放科学创新生态将对高端交流平台建设[40]产生重要影响。开放科学的目标是构建开放创新生态，而开放科研论文和科技信息等数据资源内容是该生态系统的基础。开放学术资源的深度整合利用对于提高科研创新和知识发现能力尤为重要，整合开放仓储目录将是高端交流平台建设的重要组成部分。因此，本文从资源收录情况、目录系统功能和技术选型等维度，综合调研分析了五大具有代表性的国际开放仓储目录，分析了当前开放仓储目录的建设现状和不足。在此基础上，论述了国际开放仓储目录整合的目标与思路，以及元数据描述规范设计、采集处理和元数据融合方法，初步形成了收录范围更全面、元数据项更丰富和时效性更强的开放仓储目录。

在后续研究中，笔者将探索建立一套开放学术仓储元数据动态更新融合机制和常态化监控机制，实现对开放学术仓储可访问性、资源更新情况、活跃度的动态及时跟踪，在此基础上发布全球开放仓储目录发布系统，方便用户从国别、仓储类型、学科类型等不同维度检索浏览开放学术仓储，目录系统需要针对不用仓储类型给出最优的学术资源获取方式，帮助用户获得所需资源，并且可以参考BASE系统发布年度报告，对全球范围的开放学术仓储发展现状进行扫描，为开放科学运动的发展提供支持。与此同时，也将基于整合后的仓储目录开展多来源仓储中异构开放数字资源元数据的采集收割、内容融合，并多层次实现开放资源检索、发现和挖掘利用研究与实践。

参考文献

[1] 赵艳枝，龚晓林. 从开放获取到开放科学：概念、关系、壁垒及对策 [J]. 图书馆学研究，2016（5）：2-6.

[2] Budapest Open Access Initiative [EB/OL]. [2021-12-10]. https://www.budapestopenaccessinitiative.org/read/.

[3] OA2020 Progress Report [EB/OL]. [2021-12-10]. https://oa2020.org/progress-report/.

[4] Part I: The Plan S Principles [EB/OL]. [2021-12-10]. https://www.coalition-s.org/plan_s_principles/.

[5] 黄金霞，赵展一，王昉. 从开放科学路线图分析到开放科学道路决策方法设计 [J]. 农业图书情报学报，2020，32（12）：5-19.

[6] Open Access Directory [EB/OL]. [2021-12-10]. http://oad.simmons.edu/oadwiki/Main_Page.

[7] Implementation Roadmap for the European Open Science Cloud [EB/OL]. [2021-12-10]. https://ec.europa.eu/info/news/implementation-roadmap-european-open-science-cloud-2018-mar-14_en.

[8] UNESCO Recommendation on Open Science [EB/OL]. [2021-12-10]. https://en.unesco.org/science-sustainable-future/open-science/recommendation.

[9] 王颖，张智雄，钱力，等. ChinaXiv预印本服务平台构建 [J]. 数字图书馆论坛，2017（10）：20-25.

[10] 张学文，陈凯华. 数字时代的开放科学：理论探索与未来展望 [J/OL]. 科学学研究，2021：1-10 [2022-01-14]. DOI:10.16192/j.cnki.1003-2053.20210409.001.

[11] 王翠萍，王佳佳. 科研数据知识库注册目录系统调查与分析 [J]. 情报资料工作，2017（5）：56-62.

[12] 张莎莎，黄国彬，耿骞. 基于re3data的英国科学数据发布平台研究 [J]. 数字图书馆论坛，2017（6）：16-24.

[13] 刘峰，张晓林，孔丽华. 科研数据知识库研究述评 [J]. 现代图书情报技术，2014（2）：25-31.

[14] 夏姚璜. 基于re3data的中美科学数据仓储对比研究 [J]. 图书馆学研究，2018（6）：17-26.

[15] 管凤贞，范丽婷，林菲菲，等. 基于OpenDOAR的中国开放存取知识库建设现状及主要内容研究 [J]. 晋图学刊，2019（4）：22-33.

[16] 杨丽娜，马建玲，李慧佳. 资源环境领域开放获取仓储目录的分析研究 [J]. 数字图书馆论坛，2017（9）：23-27.

[17] 郑一波，陈瑞，曾建勋. 数字环境下联合目录体系创新研究 [J]. 数字图书馆论坛，2021（10）：8-15.

[18] 肖曼，黄金霞，王昉，等. 领域特色资源的开放共享建设机制探析——以OAinONE项目为例 [J]. 数字图书馆论坛，2019（9）：2-8.

[19] 王蕾，方安，杨雨生，等. 多源期刊元数据汇聚研究——以世界卫生组织西太平洋地区医学索引为例 [J]. 数字图书馆论坛，2021（1）：47-53.

[20] PINFIELD S，SALTER J，BATH P A，et al. Open-access repositories worldwide，2005–2012：past growth，current characteristics，and future possibilities [J]. Journal of the Association for Information Science and Technology，2014，65（12）：2404-2421.

[21] SUMMANN F，CZERNIAK A，SCHIRRWAGEN J，et al. Data science tools for monitoring the global repository eco-

system and its lines of evolution [J]. Publications, 2020: 8 (2): 35.

[22] HITCHCOCK S, BRODY T, HEY J, et al. Digital preservation service provider models for institutional repositories: towards distributed services [J/OL]. D Lib Magazine, 2007, 13 [2022-01-14]. DOI:10.1045/may2007-hitchcock.

[23] ABDULLAH N. Global visibility of Asian universities' open access institutional repositories [J]. Malaysian Journal of Library & Information Science, 2010, 15 (3): 53-73.

[24] BHARDWAJ R. Open research data repositories: a content analysis to comprehend data equitable access [J]. Journal of Scientometric Research, 2019, 8 (3): 135-142.

[25] 武学超, 罗志敏. 开放科学时代大学科研范式转型 [J]. 高教探索, 2019 (4): 5-11.

[26] 肖冬梅. 开放存取资源整合及集成服务平台分析 [J]. 高校图书馆工作, 2008 (2): 27-29.

[27] 徐恩元, 李娜. 基于OpenDOAR的OA知识库政策初探 [J]. 图书馆论坛, 2010, 30 (6): 268-272.

[28] 董文鸳, 袁顺波. 全球学科知识库发展现状扫描 [J]. 图书馆, 2015 (4): 40-43.

[29] 朱莲花. 学术资源共享系统WEKO介绍 [J]. 情报探索, 2011 (3): 99-101.

[30] 周志峰. 基于资源目录网站的机构库分布研究 [J]. 图书与情报, 2009 (6): 97-103.

[31] MOULAISON SANDY H, DYKAS F. High-quality metadata and repository staffing: Perceptions of United States-based OpenDOAR participants [J]. Cataloging & Classification Quarterly, 2016, 54 (2): 101-116.

[32] 肖珑, 陈凌, 冯项云, 等. 中文元数据标准框架及其应用 [J]. 大学图书馆学报, 2001 (5): 29-35.

[33] 刘美杏, 徐芳. 古道线性文化遗产信息资源关联数据模型构建及其实证研究 [J]. 图书馆学研究, 2019 (14): 40-50.

[34] 曹雨晴, 冯东昕, 张学福, 等. 开放环境下农业科技信息资源开放共享——以FAO的实践为例 [J]. 农业图书情报学报, 2020, 32 (12): 50-58.

[35] 赵捷, 董微. 面向发现服务的图书馆元数据集成管理系统构建研究 [J]. 数字图书馆论坛, 2018 (7): 11-21.

[36] 高芳, 王艺颖. 法国开放科学顶层设计与实践进展分析及启示 [J]. 全球科技经济瞭望, 2021, 36 (5): 1-11.

[37] 丁遒劲, 苏静, 曾建勋. 国家元数据库及其协同构建框架研究 [J]. 情报理论与实践, 2020, 43 (10): 88-92.

[38] 丁遒劲. 面向资源发现的联合目录体系构建研究 [J]. 农业图书情报学报, 2021, 33 (8): 71-78.

[39] 马袁燕. 面向发现服务的文献元数据集成整合研究 [J]. 图书馆, 2019 (1): 76-81.

[40] 黄金霞, 王昉, 姜恩波, 等. 融入开放科学生态的高端交流平台建设 [J]. 数字图书馆论坛, 2021 (12): 9-14.

作者简介

张云玲, 女, 1996年生, 硕士, 研究方向: 开放学术资源整合。
罗婷婷, 女, 1985年生, 硕士, 助理研究员, 研究方向: 知识组织、大数据融汇治理、信息管理与信息系统。
赵瑞雪, 女, 1968年生, 博士, 研究员, 研究方向: 信息管理与信息系统、数字图书馆、知识组织与知识服务。
鲜国建, 男, 1982年生, 博士, 研究员, 通信作者, 研究方向: 大数据融汇治理、知识组织、知识图谱, E-mail: xianguojian@caas.cn。

Research on Integration of the International Open Access Repositories

ZHANG YunLing[1] LUO TingTing[1,2] ZHAO RuiXue[1,2] XIAN GuoJian[1,3]
(1. Agricultural Information Institute of CAAS, Beijing 100081, P. R. China; 2. Key Laboratory of Knowledge Mining and Knowledge Services in Agricultural Converging Publishing, National Press and Publication Administration, Beijing 100081, P. R. China; 3. Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081, P. R. China)

Abstract: The directory of open access repository is an instruction and index of open access repository, which is the basis for the utilization, discovery, and sharing of open academic resources. By discussing current situation and development of five international dominant open access repository registry construction, such as OpenDOAR, ROAR, BASE. We find that there are still deficiencies in the international open access repository registry revealing, such as the repository registry coverage is not complete, the registry metadata items are not abundant enough,the registry update timelines needs to be improved, and the revealing system function is relatively simple. Therefore, this paper proposes the integration of open access repository metadata based on OAI-PMH, including the design of metadata description pattern, the use of ETL tools to harvest metadata, the data cleaning tool OpenRefine to "form de-duplication" and OAI-Identify to obtain the results for "content de-duplication". Finally, this paper established a path to match and integrate metadata items of multi-source heterogeneous repository directory, formed a more richer and comprehensive global open access repository directory, and suggested the future research direction in three aspects: establishing a dynamic update and integration mechanism, a regular monitoring mechanism and a directory issuing system.

Keywords: Open Access Repository; Directory Integration; OAI-PMH; Metadata Integration; Open Access

（收稿日期：2022-01-04）