开放存取知识库及其数据采集规范的研究

万猛[1]，张永锋[2]，李振华[2]，霍东云[3]，赵弋洋[4]，王莲[4]
1. 教育部科技发展中心，北京 100080；2. 清华大学软件学院，北京 100084；
3. 北京赛时科技有限公司，北京 100084；4. 北京西普阳光教育科技股份有限公司，北京 100191

摘要
在建设学术大数据、促进学术共享的时代背景下，调查了目前开放存取知识库的发展现状，在数据规模、地区分布、系统软件等方面阐述了国内外的研究情况。以建立科研机构知识库为例，梳理了建立过程中的数据采集需求，并从数据属性、元数据标准、语义去重等方向分析了常用的数据采集规范。最后，综合考虑国内外开放存取知识库的发展现状，并结合我国发展开放存取知识库存在的问题和面临的挑战，提出了4点发展建议。

关键词
开放存取；知识库；数据；规范

中图分类号：TP3，G250.74　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2019041

Research on open-access repositories and data acquisition specifications

WAN Meng[1], ZHANG Yongfeng[2], LI Zhenhua[2], HUO Dongyun[3], ZHAO Yiyang[4], WANG Lian[4]
1. Center for Science and Technology Development, Ministry of Education, Beijing 100080, China
2. School of Software, Tsinghua University, Beijing 100084, China
3. ScientistIn Co., Ltd., Beijing 100084, China
4. Beijing Simpleware Education Technology Co., Ltd., Beijing 100191, China

Abstract
Under the background of building academic big data and promoting academic sharing, the current development status of open-access repositories was investigated and domestic and foreign research were summarized on data scale, regional distribution and system software. Taking the establishment of building academic institutional repositories as an example, the data collection requirements were analyzed and the commonly used data acquisition specifications were summarized from the aspects of data attributes, metadata standards and semantic deduplication. Finally, combined with the problems and challenges faced by China in developing open-access repositories, reasonable suggestions were put forwar.

Key words
open-access, repositories, data, specification

1 引言

随着大数据时代的到来，科学研究也进入了数据密集型阶段，科研数据的价值日益凸显。学术大数据是获取知识的捷径、科学研究的向导和终身教育的基础，更是国家大数据战略和建设“数字中国”的重要内容，对促进科学创新和加快社会发展有重大意义。在学术大数据的不断发展以及学术共享理念逐渐成为共识的背景下，开放存取（open-access，OA）这一全新的学术交流机制应运而生。

开放存取知识库是开放存取的一个主要实现途径。本文首先从国内外数据规模、地理分布、系统平台种类等方面概况和总结开放存取知识库的研究现状；接着，以面向一般科研数据的科研机构知识库为例，梳理数据采集的需求，总结数据采集过程中的相应规范；最后，从国内的现状出发，对我国开放存取知识库的发展提出建议。

2 开放存取知识库的研究现状

开放存取根据获取途径的不同，可以分为金色OA（gold open-access）和绿色OA（green open-access）。金色OA采用开放存取期刊（open-access journal）的方式，由作者支付版权费用，以实现面向读者的免费获取。绿色OA则采用开放存取知识库（open-access repositories）的形式，由作者将已出版或未出版的文献存储到知识库中，以实现免费获取。开放存取知识库主要分为机构知识库（institutional repositories，IR）和学科知识库（discipline repositories，DR）两种[1]。此外，近年来还出现了另一种开放存取途径，人们将出版的期刊文章通过盗版网站、学术社交网站等平台进行传播，这种通过非法途径免费传播的方式被称为黑色OA（black open-access）[2]，典型的黑色OA有Sci-hub、Research Gate等。

2.1 发展概况

开放存取知识库名录（the direct of open access repositories，OpenDOAR）是关于开放存取知识库的权威目录，其宗旨是通过对全球范围内的开放存取知识库资源进行系统的收集、描述、组织和传递，提高开放存取学术资源获取和使用的效益，推动开放存取运动的发展[3]。根据OpenDOAR统计，截至2015年9月，OpenDOAR收录的开放存取知识库已达到3 101个，到2017年3月增长到3 472个，到2019年5月则增长到了4 140个。在最新的4 140个开放存取知识库中，机构知识库有3 571个，占比为86.3%，学科知识库数量为338个，占比仅为8.2%。

由**图1**的数据可以看出，自2006年开放存取知识库数量迎来较大涨幅之后，每年的增加幅度为200~300个，且一直保持平稳增长。可见人们对开放存取知识库的重视程度在逐渐加强，并且越来越多的研究机构开始建立开放存取知识库。

2.2 分布情况

开放存取知识库的地区分布不均，主要集中在欧洲（45.9%）、美洲（27.3%）和亚洲（19.4%）。大部分的开放存取知识库集中在发达国家，见**表1**。其中，美国的开放存取知识库数量为575个，是第二名英国的2倍多。而我国目前的数量仅为62个，距离美国、英国、德国、

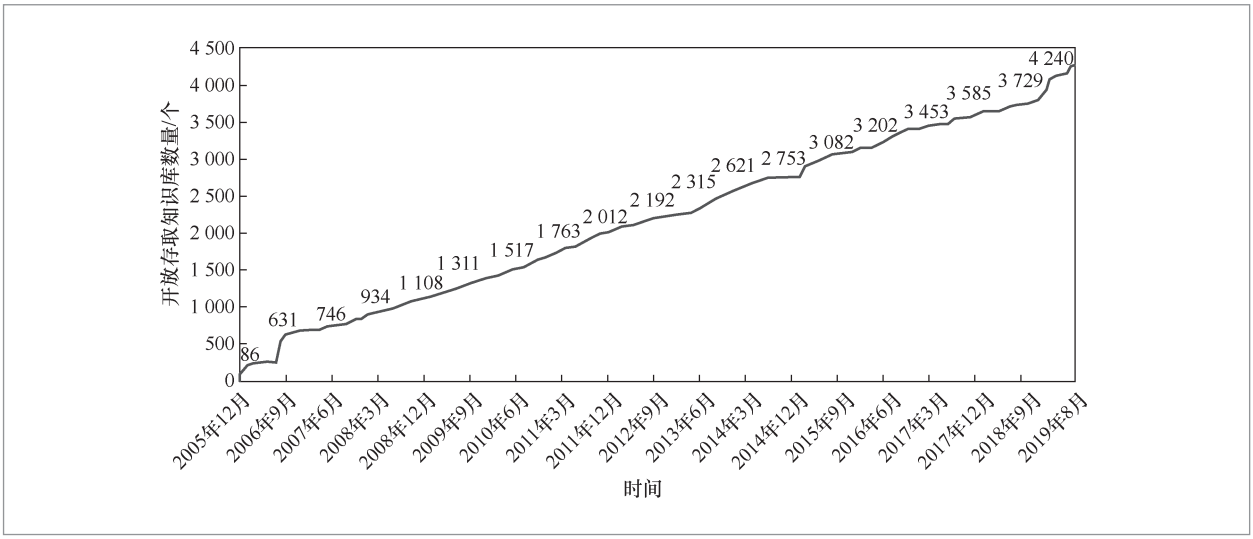

图1 OpenDOAR开放存取知识库数量增长情况

表1 OpenDOAR中开放存取知识库数量前10名的国家

序号	国家	数量/个	比例
1	美国	575	13.89%
2	英国	283	6.84%
3	德国	237	5.72%
4	日本	235	5.68%
5	西班牙	145	3.50%
6	法国	139	3.36%
7	意大利	139	3.36%
8	克罗地亚	116	2.80%
9	巴西	109	2.63%
10	波兰	106	2.56%

日本等国家还有一定差距。

一个国家开放存取知识库的数量可以在一定程度上反映国家对学术和科研的投入力度，也可以看出国家对开放存取这一形式的支持程度。我国目前的经济实力和科研实力都在逐步加强，加强开放存取知识库的建设可以更好地促进学术信息的交流和共享，提高国家的科技竞争力和影响力。

我国的开放存取知识库并未全部列入OpenDOAR中。事实上，我国正在逐渐创立自己的开放存取知识库群，如中国科学院机构知识库网格（Chinese Academy of Sciences Institutional Repositories Grid，CAS IR GRID）。CAS IR GRID以发展机构知识能力和知识管理能力为目标，快速实现对知识资产的收集、长期保存、合理传播利用，积极建设对知识内容进行捕获、转化、传播、利用和审计的能力，逐步建设包括知识内容分析、关系分析和能力审计在内的知识服务能力，开展综合知识管理。目前，CAS IR GRID共收录机构知识库114个，学者信息共计12 011位，学术成果共计985 767条，其中，包括期刊论文642 695条，会议论文112 257条，学位论文83 004条。

2.3 系统平台种类

开放存取知识库常用的系统平台可以分为开源软件和商业软件两类。比较著名的开源软件有DSpace、EPrints、Fedora等。DSpace系统由美国麻省理工学院图书

馆（MIT libraries）和美国惠普公司实验室（Hewlett-packard labs）合作完成，经过两年多的努力，于2002年10月开始投入使用。DSpace是基于伯克利软件套件（Berkeley software distribution，BSD）开源协议的软件平台，免费提供给任意学术机构使用[4]，目前也是世界上开放存取知识库使用最广泛的系统平台。目前DSpace在开放存取知识库中的占有率高达43%。

EPrints是由英国南安普顿大学于2000年研发的通用免费开源软件。EPrints遵循通用公共许可（general public license，GPL）开源协议，在发行之初就得到了广泛传播，这也是第一个免费的开放存取机构库的系统软件。该软件的出现促进了其他类似软件的发展。目前，EPrints以13%的占有率成为第二大受欢迎的知识库系统软件。

为了进行更好的中文本地化适配，并且扩展现有的开源系统功能，中国科学院在2008年开发了新的系统软件CSpace，CSpace是基于DSpace1.4.2版本扩展的，并于2012年10月正式开源[5]。通过修改、添加新组件和模块，进行连续定制和扩展，它提供了更实用、更适合中文语言的功能和服务，并根据科研人员的需求不断改进。此外，面对研究环境中数字内容不断变化的背景，CSpace允许以不可编程的方式创建或定制内容类型感知模板和相关规则，以使其适应不同的内容管理和不断变化的需求。CSpace还提供了一系列其他有用的自定义选项，以方便在本地环境下简易部署[6]。

与DSpace相比，CSpace最初是DSpace的汉化版本，经过一次次的版本迭代，目前CSpace平台已经发布6.0版本，在中国科学院110多家研究所得到部署应用，并在中国农业科学院、中国铁道科学院、兰州大学等国内数十家科研机构、高校和科技型企业应用。CSpace6.0版本的新增特性包括知识整合、学习讨论厅、批量导入和批次管理等。以CSpace为支撑平台的CAS IR GRID现已累计采集和保存各类科研成果98万余份，含全文成果量80%以上，是国内较大规模的机构知识库群和较有影响力的机构知识管理平台，也是国际三大科技机构知识库之一。按照下载数排序，截至2019年5月31日，CAS IR GRID中前10名的机构库见**表2**。

3 数据采集规范的研究现状

开放存取知识库的种类很多，不同的知识库需要不同的数据采集规范。本节以一般科研数据的高校科研机构知识库为例，梳理和总结建立科研机构知识库的过程中采用的数据采集规范。从科研行为采集需求的梳理入手，分别阐述数据库采集字段的规范、元数据标准的确立以及语义去重的方法。

3.1 科研行为采集需求

科研机构知识库的数据来源主要有两种，一种是国内外高校、研究院的官网，另

表 2 CAS IR GRID 下载量前 10 名机构

序号	机构名称	下载量/次	数据量/条
1	半导体研究所	3 087 717	15 557
2	文献情报中心	1 544 048	8 535
3	力学研究所	1 475 556	17 577
4	生态环境研究中心	1 175 433	22 838
5	合肥物质科学研究院	1 006 240	16 016
6	水生生物研究所	873 957	12 494
7	沈阳自动化研究所	843 349	17 073
8	大连化学物理研究所	736 712	34 014
9	金属研究所	657 976	31 669
10	地理科学与资源研究所	472 921	27 518

一种是国内外的其他机构知识库或科研数据库。在整合不同数据源的学术数据时，为了统一不同学术数据库的数据采集规范，首先需要明确科研行为采集规范的内容和分类。科研行为需要采集的数据主要分为5类，分别为人员、科研机构、科研项目、学术活动、学术成果。这几类数据之间并不是孤立的关系，在整合到数据库中时，会形成一个相互关联的科研关系网络。因此应该按照一定的采集顺序进行采集，下面是采集顺序和各个类别的采集说明。

（1）采集科研机构信息

科研机构的组成包括学校、科研院、研究院等。人员、项目等内容都是依托科研机构进行的，因此科研机构的采集应该放在最前面，在其他类型数据采集后再建立逻辑关系。科研机构信息可通过科研机构的官网进行采集，也可通过开放式学术数据库的接口进行集成导入。

（2）采集人员信息

人员的组成包括高校、研究院的在职教师、研究员以及有科研成果的本科生、研究生等。人员信息可通过科研机构的官网和开放式学术数据库的接口进行简要采集，也可通过个人主页进行数据提取和整合。

（3）采集项目信息

项目信息包括国家各级别的资助项目，如重大项目、面上项目、青年基金支持项目等。国家项目可以从国家自然科学基金委员会（NSFC）数据库、海研网站、知网项目信息库等处进行定向采集和导入。

（4）采集学术活动信息

会议和期刊是主要的学术活动形式，可通过会议或期刊的官网查看其基本信息，也可直接通过开放学术数据的集成接口进行查看。

（5）采集学术成果信息

论文是学术活动直接的科研成果，此外还有著作、获奖、专利等。学术成果与人员、科研项目、科研机构和学术活动等都有直接的关系，可以在会议和期刊的官网上采集，也可从科研项目数据库中获取。

3.2 数据属性规范格式

确定了需要采集的数据类别之后，还需要对应每个类别，确定属性名称字段和格式。不同的学术数据库往往有不同的属性字段，本节以学术成果中的论文为例，制定数据字段，并且规范字段的格式。

通过调研Web of Science、NSFC、中国知网（CNKI）等学术数据库对论文的字段描述，对3个数据库的字段取并集，并根据实际需求对字段进行缩减后，总结出一套较普遍的论文字段和属性，见**表3**。

对于其他类别，如人员、科研机构等，其属性字段也按照同样的思路进行调研。在确定了所有字段后，需要选择数据源。从数据量、开放存取、数据获取难易度等几个角度综合考虑，尽可能选取数据量大、能直接获取原文、反爬虫措施较弱、网站稳健性高的站点，NSFC、CAS IR GRID等开放存取学术数据库是较好的数据源选择。

3.3 元数据标准

元数据也被称为数据的数据（data about data），一般是提供关于信息资源或数据的一种结构化的数据，用于组织、描述、检索、保存、管理信息和知识资源。元数据可以使信息描述和分类实现格式化，从而为机器处理创造了可能。在不同的应用场景下，元数据有不同的标准。在开放存取知识库中，元数据主要分为3种：描述型元数据、管理型元数据和结构元数据[7]。

总之，科研机构知识库的元数据标准类型是多种多样的，其中经常采用的是

都柏林核心（Dublin core，DC）元数据和数据引用元数据框架（datacite metadata schema）。DC元数据标准由于其高度的普适性和扩展性，是描述科研数据最常用的元数据标准。目前流行的机构库系统（如DSpace、EPrints、Fedora等）都对DC元数据提供支持。DC元数据的简化形式（simple Dublin core）共包含15个元素，见**表4**。

制定了元数据标准之后，还要规范描述元数据的方法。资源描述框架（resource description framework，RDF）是由WWW提出的对万维网（World Wide Web）上资源进行描述的一个框架，为互联网上的信息描述提供了一种规范。RDF由主语、谓词、宾语的三元组形式组成，其中，主语一般由统一资源标识（uniform resource identifiers，URI）表示，谓词描述实体具有的相关属性，宾语为属性对应的属性值[8]。

RDF采用XML文件的形式。RDF的强大之处在于，在确定了主语之后，谓词和宾语可以根据需要自由使用。而最常见的谓词和宾语是DC元数据标准。DC元数据标准的简单形式有15个属性，对应着15个谓词和宾语。采用DC元数据标准后，RDF基本可以表示所有网络资源。目前主流的机构库系统软件都支持基于DC元数据标准的RDF格式的XML文件，通过这一形式的

表3 论文类型采集字段和属性

字段	类型	长度/byte	字段	类型	长度/byte
题名	Varchar	255	页码	Varchar	255
作者	Varchar	255	摘要	LongText	–
作者单位	Varchar	255	doi	Varchar	255
期刊名称	Varchar	255	ISSN	Varchar	255
会议名称	Varchar	255	卷	Int	10
类别	Varchar	255	期	Int	10
分类代码	Varchar	255	会议地点	Varchar	255
来源数据库	Varchar	255	会议代码	Varchar	255
通信作者	Varchar	255	唯一获取号	Varchar	255
会议日期	DateTime	–	出版日期	DateTime	–

表4 DC元数据简化形式的15个元素

序号	英文名称	中文名称	定义
1	title	题名	赋予资源的名称
2	creator	创建者	创建资源内容的主要责任者
3	subject	主题	资源内容的主题描述
4	description	描述	资源内容的说明
5	publisher	出版者	可以获得资源并使其可用的责任者
6	contributor	其他责任者	对资源的内容做出贡献的其他实体
7	date	日期	与资源生命周期中的一个事件相关的时间
8	type	类型	资源内容的特征或类型
9	format	格式	资源的物理或数字表现形式
10	identifier	标识符	在特定的范围内给予资源的一个明确的标识
11	source	来源	对当前资源来源的参照
12	language	语种	描述资源知识内容的语种
13	relation	关联	对相关资源的参照
14	coverage	覆盖范围	资源内容涉及的外延与覆盖范围
15	rights	权限	有关资源本身所有的或被赋予的权限信息

规范设计，极大地方便了数据库或机构库之间的数据交流。

3.4 语义去重标准

由于数据源的多样性，在采集数据的过程中会采集到来自不同数据源的重复数据，因此需要对重复数据进行去重。针对5种不同的数据类别，需要制定不同的去重标准。由于数据的质量不统一，对统一类别也应采取多种去重方式并行的方法，防止某字段的缺失导致的去重失败。

针对论文数据的去重，可以有以下两种方式。

- 区分标题和会议/期刊名称。论文的标题相同的概率不高，即使相同也很难出现在同一个会议或期刊上，因此可以从逻辑上进行区分。然而由于论文标题较长，对比的效率和正确率会受一定的影响。
- 区分唯一标识码和DOI。由于论文的唯一标识码不会重复，故可直接进行比较，但是由于数据源的质量不同，有可能收集不到此条数据。

人员信息的去重是去重工作的核心。对于姓名相同、工作机构也相同的人员，可以认为是同一个人。由于英文的特殊性，在判定姓名相同时，需对姓、名、中间名进行切分，有以下情况可以讨论：

- 姓、名、中间名全部相同，则认为人名相同；
- 姓、名相同，中间名存在缩写情况，若缩写与全称第一字母对应，则认为人名相同；
- 姓、名相同，但有一方中间名缺失，则认为人名相同[9]。

由于数据量较大，因此很难总结出完美的去重标准。若出现极端特殊情况，如同一机构内有相同姓名的两人，应该人工处理解决。数据清洗的效果在一定程度上也会影响去重的效果，因此做好清洗工作，采用官方标准名称和规范十分重要。

4 我国相关研究的启示

近年来，随着大数据研究的热潮兴起，学术大数据展现着越来越重要的作用。通过建立开放存取知识库来提交、保存、管理和组织科学研究过程中的原始数据，实现科研数据的共享，促进学术交流，增进学术繁荣发展，逐渐成为科研工作者以及高校图书管理人员的共识。通过对开放存取知识库现状的调研及与国外知识库的储量对比，可以看出，我国相关领域的研究还处在探索与发展阶段。近年来，中国机构知识库推进工作组主办了“中国机构知识库学术研讨会”，主要研讨我国机构知识库的建设问题。综合目前国内研究存在的一些问题，对我国的研究启示可以分为以下几个方面。

4.1 增强高校对建设机构库的重视

从OpenDOAR的整体数据来看，我国的机构库数量处于较低水平。近年来，我国的科技实力和创新实力都明显增强，应有更多的研究成果和机构库出现。截至目前，有很多高校的机构知识库还处在建设阶段，也有很多高校的机构知识库建成后因疏于维护，无法访问[10]。我国可以灵活出台鼓励政策，从根本上解决高校建设机构知识库动力不足、重视程度不够的问题。

4.2 建立多个机构知识库群

我国目前较为成熟的机构知识库群只有CAS IR GRID，该网站仅收录基于CSpace系统的机构库，事实上还存在很多

DSpace系统的机构库。很多高校不重视建立机构知识库，很大的原因是建成之后的机构知识库知名度较低，点击率也较低，难以形成正向的反馈机制。建立机构知识库群时，“组团取暖”的形式可以提高机构库成员的点击率和知名度，有效带动新兴机构知识库的发展。

4.3 增加版权意识，防止黑色OA的负面影响

目前，比较流行的黑色OA途径源于国外，国内的开放存取知识库发展较慢，这也导致黑色OA还没有起到较大的影响。国外的黑色OA的流行度也给我国的OA发展敲响了警钟，在建立科研机构知识库的过程中，用户必须遵守国际版权法，明确数据集的所有权以及数据上传、下载、传播的许可权限。机构库建设的从业人员也应及时参加相关版权培训，以保障机构知识库建设的安全和可持续发展。

4.4 增加软件系统的种类和水平

我国目前除了CAS IR GRID使用自己扩展和研发的CSpace之外，其他机构库主要使用的是DSpace，而且基本采用默认设置。这使得我国的机构知识库软件平台的功能十分单一，缺少定制化和个性化的功能。建议在增加对系统平台的多样性尝试之外，也可尝试修改个性化的配置，以增强多样性。

5 结束语

随着大数据时代的到来，数据成为重要的生产要素和战略资源，数据的价值和复用率不断提升，并逐步形成数据开放共享的社会氛围。学术大数据的开放存取已经成为一种全球的潮流，变成一种影响全球学术交流、信息共享的运动。近年来，我国在开放存取知识库方面的研究正快速发展，但是与世界先进水平相比还有一定的差距。加强高校对机构知识库的建设可以有效提高我国的开放存取建设的效果，增加学术交流和促进学术共同繁荣。

参考文献:

[1] 李月明. 基于OpenDOAR的开放存取知识库分析与研究[J]. 图书馆, 2017(7): 46-48, 98.
LI Y M. Analysis and research on open access repository based open DOAR[J]. Library, 2017(7): 46-48, 98.

[2] BJORK B C. Gold, green, and black open access[J]. Learned Publishing, 2017, 30(2).

[3] 何琳. OpenDOAR和机构知识库发展现状[J]. 图书馆工作与研究, 2009(2): 30-33.
HE L. OpenDOAR and development of institutional repository[J]. Library Work and Study, 2009(2): 30-33.

[4] 杨武健, 王学勤. DSpace机构知识库系统的分析与研究[J]. 现代情报, 2006(11): 220-222, 225.
YANG W J, WANG X Q. Analysis and research of dspace institutional repository system[J]. Journal of Modern Information, 2006(11): 220-222, 225.

[5] 祝忠明, 马建霞, 卢利农, 等. 机构知识库开源软件DSpace的扩展开发与应用[J]. 现代图书情报技术, 2009(Z1): 11-17.
ZHU Z M, MA J X, LU L N, et al. Developing an institutional repository platform via extending dspace[J]. New Technology of Library and Information Service, 2009(Z1): 11-17.

[6] ZHU Z M, ZHANG W Q, LIU W, et al. CSpace-a more practical and customizable repository platform serving local needs[J]. Polymer-Plastics Technology and

Engineering, 2010, 49(7): 662-671.

[7] 吴玲芳. 用于机构知识库的元数据研究[J]. 现代情报, 2009, 29(8): 128-130, 134.
WU L F. Study on metadata used in institutional repository[J]. Journal of Modern Information, 2009, 29(8): 128-130, 134.

[8] 杜小勇, 陈峻, 陈跃国. 大数据探索式搜索研究[J]. 通信学报, 2015, 36(12): 77-88.
DU X Y, CHEN J, CHEN Y G. Exploratory search on big data[J]. Journal on Communications, 2015, 36(12): 77-88.

[9] 李建伟, 宋文, 汤怡洁, 等. 科研本体知识库数据建设研究[J]. 现代图书情报技术, 2013(11): 15-21.
LI J W, SONG W, TANG Y J, et al. Research on data building for knowledge base based on scientific research ontology[J]. New Technology of Library and Information Service, 2013(11): 15-21.

[10] 朱立禄, 宋世俊, 王琳. 国内外机构知识库建设现状及建议[J]. 现代情报, 2017, 37(3): 109-115.
ZHU L L, SONG S J, WANG L. The development status of worldwide institutional repositories and some corresponding measures[J]. Journal of Modern Information, 2017, 37(3): 109-115.

作者简介

万猛（1975- ），男，博士，教育部科技发展中心研究员，主要研究方向为信息管理与信息系统、科技评价与管理、教育大数据等。

张永锋（1994- ），男，清华大学软件学院硕士生，主要研究方向为云存储、网络信息爬取等。

李振华（1983- ），男，博士，清华大学软件学院副教授、博士生导师，主要研究方向为云计算、云存储、移动互联网等。

霍东云（1981- ），男，北京赛时科技有限公司联合创始人兼首席技术官，主要研究方向为大数据、云计算、移动互联网等。

赵弋洋（1975- ），男，博士，北京西普阳光教育科技股份有限公司首席科学家，主要研究方向为物联网、定位、移动互联网等。

王莲（1984- ），女，北京西普阳光教育科技股份有限公司高级经济师，主要研究方向为物联网、大数据、产业经济学等。

收稿日期：2019-05-31
通信作者：李振华，lizhenhua1983@gmail.com
基金项目：国家重点研发计划基金资助项目（No.2018YFB1004701）
Foundation Item: The National Key Research and Development Program of China (No.2018YFB1004701)