

国外开放资源建设案例研究

陈雪飞, 黄金霞

(中国科学院文献情报中心, 北京 100190)

摘要: 开放资源建设是资源建设的重要内容之一,也是未来的发展方向。文章阐述了开放资源建设的内涵,以欧盟委员会和NIH作为样本机构,选取4个有影响力的开放资源建设案例进行研究,对样本案例从资源机制、服务机制、政策机制3个方面进行了考察。最后对样本案例的经验进行了总结,提出扩大资源类型、进行联盟建设、开发知识服务及产品的建议。

关键词: 开放获取资源; 资源建设; OPENAIRE; Zenedo; REPORTER; PMC

中图分类号: G201

DOI: 10.3772/j.issn.1673—2286.2014.12.004

在数字网络技术的推动下,用户对信息资源的需求越来越趋于数字化、开放化、知识化,科技学术论文开放获取迅速成为学术交流主流模式之一,开放数据正成为政府数据和科研数据的基本准则,开放信息开始成为新一代科技知识基础资源^[1]。欧盟研究与创新董事会发布的报告称,科研文献的开放获取要比预想的增长得更快,2011年发表的大约50%的论文现在都可以在线免费获取,全球科研成果正在向开放获取方式转变^[2]。布达佩斯宣言相关组织召开会议,明确提出10年后在世界上的任何国家或地区任何学科领域的同行评议学术论文均实现开放存储与开放获取^[3]。可以确定的是,当前的学术信息资源建设正面临开放信息环境的挑战和机遇。

为了迎合这一趋势,包括图书馆在内的组织和机构开始启动建设开放资源服务设施和系统,开放资源建设从创新期、巩固期、持续期发展到目前的战略期,OA相关问题已经能放到国家层面上讨论和解决,科研资助者、OA组织转为支持更深层次的OA知识服务^[4],开放资源建设在实践方面已经积累了众多优秀的经验。本文在阐述开放资源建设范畴的基础上,从建设机构角度选取国外若干开放资源建设案例进行考察,以期为

我国开展开放资源建设工作提供借鉴。

1 开放资源建设的内涵

开放学术信息资源体系涵盖的资源内容,从来源角度,包括开放状态的数字化学术信息资源以及用户创造的资源(UCC)等;从载体类型角度,包括开放期刊、机构知识库、开放会议、开放课件、开放报告、行业信息、产业信息、科学数据、科研工具、仪器设备以及广泛的经济社会和法律信息资源等^[5]。Peter Suber认为,原则上任何内容都可以OA,只是资源类型不同,OA的操作难度也不同。Peter Suber列举了已经有成功OA案例的一些学术信息资源,包括同行评议论文、预印本、学位论文、研究数据、政府数据、源码、会议资源、学术专著、教科书、新闻、存档记录和手稿、图像、开放教育资源以及数字化印本资源^[6]。

从Peter Suber的观点来看,开放资源的数量将不可估量。除了常规出版物外,非常规信息资源的开放获取与利用也渐受重视,例如实验日志的开放存储^[7]、开放获取基因组计划的诞生^[8]等,科研过程中产生的海量科学数据,互联网上以TB为单位的开放资源,在大数

据时代的推动下成为开放的可计算的知识。美国政府开发运营的开放政府数据网站data.gov体现了大数据时代所需要的基本特征：数以十万计的开放数据库，数以千计的数据分析工具，对任何第三方开放，欢迎来自其他国家、企业和个人的参与建设和创新^[9]。

开放资源的开放性、规模性、复杂性，导致开放资源的建设不能搞“数据独占”，必须支持全流程的开放性建设。因此，开放资源建设是对涵盖所有在线可开放获取资源的获取与组织、评价与遴选、管理与开放存储、开放许可与开放重用的过程，这一过程涉及资源、服务、政策、技术规范4方面的因素。其中“资源”是解决建设对象的问题，“服务”是解决实现功能的问题，政策属于开放建设中的顶层设计，是解决建设全流程管理的问题，“技术规范”是解决如何实现其他三个因素的问题，贯穿于其他三个因素的建设过程中，这四个因素构成了开放信息资源建设的内涵（如图1）。

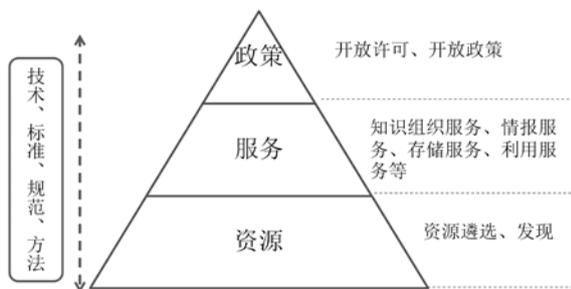


图1 开放信息资源建设要素图

2 案例研究

从技术时间线来看，开放资源建设已经进入一个全新的时期，从1.0时代的数字图书馆、数据库技术到2.0时代科研社交网络、开放共享创作平台的兴起，如今在大数据时代的推动下资源建设正走向科研成果开放集成、发现和计算。从地区来看，目前欧美地区的应用建设比较成熟，并且从国家和地区层面着力于建设和完善相关机制，以项目形式启动建设一系列开放资源服务系统。从资源角度，其建设模式分为元数据集成和全文集成两种模式^[10]，集成的学科类型则包括综合集成和学科集成。本文根据开放资源建设开展的影响力和规模性，以及目前开放资源发展到战略期这一宏观趋势^[4]，重点研究欧盟和美国NIH机构的开放资源建设进展，介绍欧盟

OPENAIRE机构知识库网络、知识仓储Zenodo、NIH在开放政府计划推动下启动的REPORTER系统、学科仓储PMC，其基本情况如下表1。

表1 欧盟和NIH开放资源建设项目类型

地区	建设机构	项目名称	集成类型	资源类型
欧盟地区	欧盟委员会	OPENAIRE ^[11]	元数据集成	综合知识资源整合系统
		Zenodo ^[12]	全文集成	综合知识仓储
北美地区	NIH	REPORTER ^[13]	元数据集成	学科知识资源整合系统
		PMC ^[14]	全文集成	学科知识仓储

2.1 欧盟及NIH开放资源建设体系比较

(1) 欧盟委员会

使公共资助的研究成果最大化开放获取是欧盟开放资源建设的战略目标和主要方向。2008年欧盟委员会提出FP7开放获取先导计划，要求由EC或FP7资助产生的科研成果在6个月或12个月内开放获取^[15]；到2014年，地平线2020计划取代FP7，除了延续FP7开放获取政策外，开始重视研究数据的开放和重用，形成包括出版物和研究数据在内的开放共享机制，如图2。

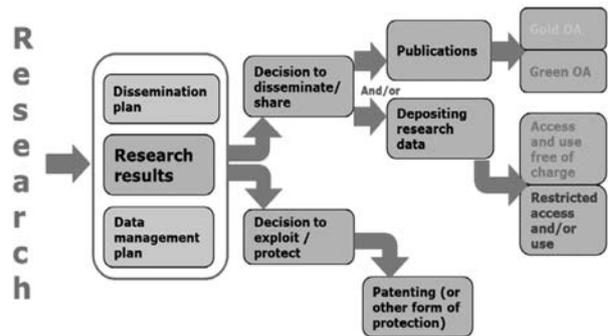


图2 地平线2020: 开放获取出版物和研究数据的利用和传播^[16]

2012年，欧盟委员会提交了《科学信息访问和保存建议》报告^[17]，认为要扩展开放获取的范围，将目前的政策应用到所有接受公共资助的研究。报告中提出对于科研成果开放建设的建议，包括建设开放数字信息基础设施，实现科研出版物以及科研数据的开放获取，科研资助机构、学术机构、出版商、研究人员通力支持参与等（图3）。

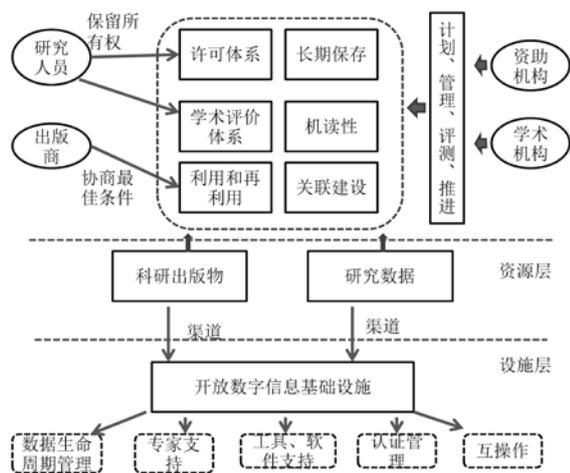


图3 欧盟委员会提出的开放资源建设架构

(2) NIH

NIH作为国家级的科研资助机构和研究机构，一直致力于开放获取在政策和实施层面的推进。与欧盟类似，NIH也同时推动出版物和研究数据的开放获取，其中，公众获取政策要求在符合版权法的范围内，所有由NIH资助产生的同行评议文章必须在出版后12个月提交到PMC以供开放获取^[18]。NIH还开发了适应于单年度申请金额在50万美元以上项目的数据共享政策^[19]，要求申请者提供数据共享计划，使NIH资助产生的最终研究数据，尤其是那些unique data要能为其他研究者所用。这些数据来自于基础研究、临床研究、调查研究、其他类型研究等。一旦数据可用，研究人员既不可以设置任何限制问题或方法，也不能以合作者作为获取数据的条件，需要出台适当的文件以确保他人使用，防止误用、误解和混淆。

开放资源建设方面，2004年NLM面向开放获取后，其网站的访问人数增加了一百倍^[20]。PUBMED由收费索引转向在线免费获取，PUBMED CENTRAL的诞生及在欧美地区的分布式建设，以及研究项目资源集成建设——REPORTER，成为开放获取进程中具有里程碑式的事件。另外，NIH设立了许多开放获取的数据库，包括基因型与表型数据库dbGaP，扩大基因和临床数据开放获取的SHARc^[21]，千人基因组计划全部数据免费对外开放，数据总量达到200TB，是世界上最大的人类基因变异数据集，由亚马逊旗下的云计算公司——“亚马逊网络服务”存储^[22]。2014年，NIH推动生物医药大数据开放教育资源项目^[23]，致力于推动生物医药开放教育资源(OER)的利用，提

高从业人员应用和分析生物医药大数据能力，并支持MOOCS发布和传播。

2.2 资源机制

资源机制指资源的发现、获取、遴选，以向用户提供需求内容的机制。这4个开放资源建设项目都面向科研项目资源，但其资源机制不尽相同。

OPENAIRE是由欧洲38个合作国共同支持开发的开放知识库网络，对欧盟委员会通过FP7或欧洲研究委员会(ERC)资助的成果提供在线免费获取。但OPENAIRE本身并不提供资源内容本身，而是通过建设泛欧网络架构和技术架构，实现资源在元数据层的集成和存储(图4)。目前OPENAIRE提供beta版的集成搜索引擎，资源类型包括出版物、数据、项目、研究人员、机构、数据来源，学科领域涉及健康、能源、环境、部分信息通讯技术和科研基础设施、社会科学、人文等。Zenodo由OpenAIRE和CERN创建，European Commission资助，为研究成果提供一站式存储服务。因此，尽管基于和OpenAIRE一样的理念，但不同于OpenAIRE收集知识库的资源元数据记录，Zenodo作为OpenAIRE知识库网络中的一员，由研究人员上载数据。而且，Zenodo支持“长尾”研究成果的存储，接收任何类型、任何阶段、任何学科、任何格式、任何许可的数据。

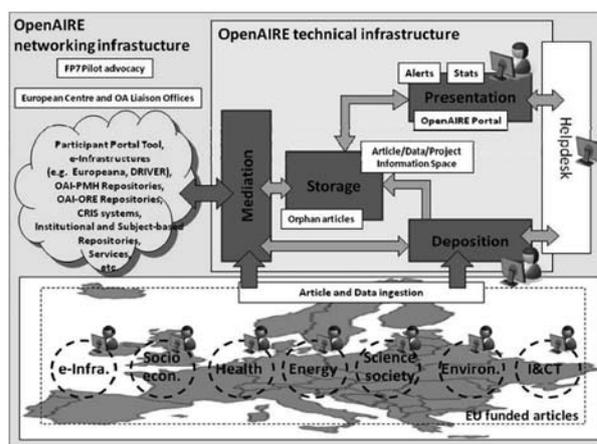


图4 OpenAIRE架构^[24]

REPORTER由NIH开发，对NIH以及美国其他健康和公共服务机构资助的项目的信息进行集成，是一个完善的、功能全面的、资源多样的项目资源系统。

此外,除项目的具体信息外,这些项目产生的出版成果和专利信息也提供使用。REPORTER是NIH“政府公开计划”的一个重要组成部分,以展现NIH活动的透明度,促进数据收集的质量,提高资源利用率。除了REPORTER项目、出版成果、专利数据集外,REPORTER中的信息还来源于一些已有的数据库——

eRA数据库、Medline、PubMed Central、NIH院内数据库以及iEdison。PMC是NIH期刊论文的仓储,由于其开放存缴政策的要求,资源主要来源于作者提交以及出版商按照PMC的格式进行自动发布。资源的内容包括生物医药领域的开放获取论文、NIH资助的研究成果论文等。表2总结了4个项目的资源属性。

表2 4个项目的资源属性

项目名称	资源来源	资源学科范围	资源内容类型	资源特色
OpenAIRE	作为第三方收割知识库元数据	FP和ERC资助的7大学科	出版物、项目、数据、人员、机构、数据源的元数据信息	泛欧科研信息开放集成
Zenodo	资源所有者提交	任何学科	数据集、图片、海报、演示报告、出版物、软件、视频等不限	支持长尾科学成果存储
REPORTER	NIH资助项目数据库及相关数据库互链	生物医药领域	项目、出版物、专利、试验等元数据信息	美国生物医药领域跨库、跨资源类型项目科研产出集成和关联
PMC	作者或出版商提交	生物医药领域	期刊论文	最权威的生物医药领域出版物开放仓储

2.3 服务机制

开放资源建设的目的在于基于资源向用户提供增值服务,方便用户利用知识,促进开放知识资源的传播和利用。服务类型包括知识组织服务、情报服务(包括资源评价、计量、分析、可视化等服务)、长期保存服务、再利用服务、Web 2.0服务等。

知识组织方面,OpenAIRE作为知识库网络门户需要向知识库成员执行一套统一的元数据体系,以便进行内容收割、聚合和展示。因此,OpenAIRE面向主要的科研产出——出版物和研究数据分别开发了相应的指南(面向CRIS系统的指南正在制定中),定义了出版物和研究数据的描述规范。其中,出版物在DC元数据schema下,采用info:eu-repo通用词汇表,以促进开放获取领域内标准词表的推广使用。对于数据集,OpenAIRE采用DataCite元数据schema,通过实施这一指南,旨在促进增强出版物的创建以及为关联数据基础设施的建设奠定基石^[25]。Zenodo作为OpenAIRE知识库成员之一,也遵循这一标准的描述规范。另外,OpenAIRE及Zenodo也积极推动知识关联,OpenAIRE后续启动OpenAIRE plus项目,开展国家层面上出版物与相关信息(如数据集、多媒体、作者

信息、资助信息等)的关联。Zenodo则基于OpenAIRE泛欧倡议,扩展研究成果与数据集和资助信息在欧洲和国家背景下的链接,并为科学团体创建特定集合,例如,一个EC资助的项目可能会为他们所有的研究成果创建一个协作空间,并能分配一系列许可,包括Creative Commons,每个数据集和出版物都被分配一个DOI^[26]。REPORTER的元数据非常详尽,每个元数据是否有值取决于来源的数据库。NIH应用“研究、条件、疾病分类系统”(RCDC)^[27]对项目进行自动描述和分类,RCDC词汇表汇总了MESH等7个词汇表,并采用专家选词和对词汇设置权重的方法,一定程度实现了语义的丰富化和关联化。PMC采用了NIH开发的JATS标签集,互操作和通用性较强。作为项目资源集成工具,REPORTER跨库知识链接服务非常强大,实现了项目产出知识及相关知识的关联整合以及平台与外部资源的关联。PMC则实现从参考文献直接链接到PubMed、CrossRef的跨库链接。

在利用服务方面,OpenAIRE使用OAI-PMH2.0协议开放出版物元数据。Zenodo提供丰富的接口来链接对象和资金信息,可通过OAI-PMH对任何上传的数据或创建的集合进行全文收割,以方便把数据集展示给PubMed Central或本地机构^[26]。REPORTER并不

提供接口服务,但可通过ExPORTER站点进行整库下载(EXCEL格式)重用。PMC支持全部元数据的自动获取,但对于全文内容仅支持OA子集全文的FTP批量获取。情报服务方面,OpenAIRE开通资源监测功能,对资源进行实时分类统计、分析,并引入可视化图表呈现;REPORTER也提供项目及资助信息、出版物的可视化分析,并可以选择不同可视化类型,以及通过谷歌地图接口呈现机构的地理位置。存储服务方面,PMC和Zenodo作为仓储都提供存储服务,Zenodo为每个用户

提供2GB的存储空间。Web 2.0交互方面,基本都提供RSS服务,但其他交互方式不多。

2.4 政策机制

政策机制是指为了保障资源系统有效运行,促进资源开放利用和传播而制定或采取的采集、保存、利用政策或许可协议等。下面分别对四个资源项目的数据权利政策、保存政策、利用政策进行分析(表3)。

表3 资源政策分析

项目名称	数据权利政策	保存政策	利用政策
OpenAIRE	鼓励所有者保留所有权	要求所有资助成果必须存储到指定知识库中。	—
Zenodo	要求提交之前不存储在所有权变更	1.永久保存(仓储生命期内)。2.仅作功能性保存。3.每天备份,提供多重副本。4.定期校验稳定性和可靠性。	1.数据必须指定一个许可类型。 2.数据可以指定一个时滞期,到期后自动转为公众获取。 3.所有元数据许可类型为CC0。
REPORTER	—	未知	符合总统“开放政府计划”和信息自由法案(FOIA)
PMC	—	所有由NIH全部或部分资助产生的、或由NIH成员撰写的同行评议在接受发表时向PMC存缴其同行评议的最终稿。	—

3 结语

开放获取资源的迅速增长带来了开放资源建设的迫切需求,同时也对传统资源建设的模式提出了挑战,开放获取成为大规模科学文献计算的可能条件^[28]。根据本文介绍的开放资源建设案例的经验可知:首先,面向科研活动的信息服务体系不再局限于传统出版物,研究数据以及其他任何形式的知识在开放信息环境下变得更易获取和利用,面向科研活动全过程和知识生命周期全过程的资源建设成为现实。其次,开放资源打破了资源专有壁垒,开放资源开始上升到国家层面甚至国际层面进行联盟建设和战略布局,资源得到了最大化整合和共享。再次,目前基于描述性元数据的文献集成与系统互操作仍是主流,信息资源主要呈现为文档和自由文本形式的文献对象,缺乏结构化、数据化、语义化描述与知识组织^[29],而开放资源建设的优势和核心在于应用开放知识组织体系,从而将开放信息资源体系转

换为可计算的开放知识资源^[5]。因此需要支持资源层的开放集成与内容层的知识发现,如开放性综合集成平台、知识库、知识搜索引擎、资源嵌入式开放服务产品等的开发。另外,强制性政策的推动是开放获取的重要助推剂,要求研究成果的开放获取和存储,推动商业资源、非公开科研信息资源的开放,也是重要使命之一。

参考文献

- [1] 张晓林.开放获取、开放知识、开放创新推动开放知识服务:30会聚与研究图书馆范式再转变[J].现代图书情报技术,2013,230(2):2-10.
- [2] Open access to research publications reaching “tipping point” [EB/OL]. [2013-08-22]. <http://www.openaccess.nl/news/382-open-access-to-research-publications-reaching-tipping-point>.
- [3] Budapest open access initiative. Ten years on from the Budapest open access initiative: Setting the default to open [R/OL]. [2014-03-30]. <http://www.opensocietyfoundations.org/openaccess/boai-10->

- recommendations.
- [4] 张建勇,黄金霞,李麟.论开放资源建设的发展(2001-2013)[J].图书情报工作,2013,57(21):12-17.
- [5] 孙坦.开放信息环境:学术图书馆信息资源建设的重定义与再造[J].中国图书馆学报,2013,39(205):9-17.
- [6] SUBER P. OPEN ACCESS [M]. London: The MIT Press, 2012: 97-99.
- [7] BOHLE S. Open access: Online repository for lab notebooks [J/OL]. Nature, 2014, 506(159) [2014-04-10]. <http://www.nature.com/nature/journal/v506/n7487/full/506159e.html>.
- [8] CALLAWAY E. Open-access genome project lands in UK [EB/OL]. Nature [2014-04-10]. <http://blogs.nature.com/news/2013/11/openaccess-genome-project-lands-in-uk.html>.
- [9] 谢文.山重水复寻新路:七谈大数据时代[EB/OL]. [2014-04-10]. http://blog.sina.com.cn/s/blog_513a2b8001017ddp.html.
- [10] 陈雪飞,黄金霞,张建勇,等.国外开放科技资源项目实证研究及启示[J].图书馆建设, 2013(6):18-21,25.
- [11] OpenAIRE [EB/OL]. [2014-04-10]. <https://beta.openaire.eu/search>.
- [12] Zenodo [EB/OL]. [2014-04-10]. <http://zenodo.org/>.
- [13] REPORTER [EB/OL]. [2014-04-10]. <http://projectreporter.nih.gov>.
- [14] PMC [EB/OL]. [2014-04-10]. <http://www.ncbi.nlm.nih.gov/pmc/>.
- [15] FP7 [EB/OL]. [2014-04-10]. <https://beta.openaire.eu/open-access-inf7-seventh-research-framework-programme>.
- [16] European Commission. Guidelines on Open Access to Scientific Publications and Research Data [R/OL]. [2014-04-10]. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.in Horizon 2020.
- [17] COMMISSION RECOMMENDATION [R/OL]. [2014-04-10]. http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf.
- [18] NIH Public Access Policy Details [R/OL]. [2014-04-10]. <http://publicaccess.nih.gov/policy.htm>.
- [19] NIH Data Sharing Policy [R/OL]. [2014-04-10]. http://grants.nih.gov/grants/policy/data_sharing/.
- [20] Johnson R K. Will Research Sharing Keep Pace with the Internet? [J]. The Journal of Neuroscience, 2006, 26(37): 9349-9351.
- [21] NIH Launches Extensive Open-Access Dataset of Genetic and Clinical Data [EB/OL]. [2014-04-10]. http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v22.p8.
- [22] 1000 Genomes Project data available on Amazon Cloud [EB/OL]. [2014-04-10]. <http://www.nih.gov/news/health/mar2012/nhgri-29.htm>.
- [23] Department of Health and Human Services [EB/OL]. [2014-04-10]. <http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-14-009.html>.
- [24] CASTELLI D, MANGHI P. The OpenAIRE Project: Open Access Infrastructure for Research in Europe [R/OL]. [2014-04-10]. <http://ercim-news.ercim.eu/en80/es/openaire>.
- [25] OpenAIRE. OpenAIRE Guidelines: For Data Archives [R/OL]. [2014-04-10]. https://guidelines.openaire.eu/wiki/OpenAIRE_Guidelines_For_Data_Archives.
- [26] Zenodo Launches! [EB/OL]. [2014-04-10]. <http://www.openaire.eu/en/component/content/article/9-newsevents/457-zenodo-is-launched>.
- [27] NIH. Categorization Process [R/OL]. [2014-04-10]. <http://report.nih.gov/rcdc/process.aspx>.
- [28] LYNCH C. Open Computation: Beyond Human-Reader-Centric Views of Scholarly Literatures [M]// JACOBS N. Open Access: Key Strategic, Technical and Economic Aspects. Chandos Publishing, 2006: 185-193.
- [29] 曾蕾. DCMI 2013听会点滴[J].数字图书馆论坛,2013(12):19-23.

作者简介

陈雪飞,女,1987年生,研究方向:开放资源,E-mail: chenxuefei@mail.las.ac.cn。

黄金霞,女,1972年生,研究方向:开放资源建设、知识组织,E-mail: huangjx@mail.las.ac.cn。

Best Practice Study on Open Access Resources Development

CHEN XueFei, HUANG JinXia

(Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Open access resources development is one of the most important components in the collection development of libraries, which is also the future trends. 4 Open access resources development cases from European Commission and NIH were analyzed as samples according to their resources mechanisms, services designs and policy systems. Especially, at the basis of the demonstration experience of these outstanding practices, we should continue to expand resources scopes, pay attention to alliance and cooperation, and develop knowledge tools and products.

Keywords: Open access resources; Resources development; OpenAIRE; Zenedo; REPORTER; PMC

(收稿日期: 2014-12-11)