

国外科学数据共享方案研究^{*}

彭秀媛^{1,2}, 王枫³, 周国民¹

(1. 中国农业科学院农业信息研究所, 北京 100081; 2. 辽宁省农业科学院信息中心, 沈阳 110161;

3. 中共辽宁省委党校信息中心, 沈阳 110004)

摘要: 本文从技术角度系统梳理国外现有主要科学数据共享方案, 将其归纳为四类: 基于数据知识库的共享方案、基于数据知识库和开放数据集的注册与目录系统的共享方案、基于机构知识库联盟的共享方案和基于科学数据基础设施的共享方案(以下简称“四类共享方案”), 通过对每类方案及其代表性案例的技术实现进行剖析, 对上述方案特点进行分析, 从服务角度对代表性案例进行比较分析, 总结方案带来的启示, 为我国科学数据共享工作进一步开展提供思路。

关键词: 科学数据; 共享; 数据知识库; 目录系统; 知识库联盟; 基础设施

中图分类号: S-3

DOI: 10.3772/j.issn.1673-2286.2017.09.008

科学数据的共享、开放和管理逐渐成为一种学术要求和学术态势, 成为国际组织、科研资助机构、科研机构、图书馆等关注的热点问题。多年来, 美国、澳大利亚、欧盟等发达国家和组织在数据共享研究方面开展了较好的理论与实践研究。本文从技术角度对国外科学数据共享建设方案进行全面梳理, 在各类方案中列举较有代表性的案例进行深入探析, 以帮助国内相关机构积极借鉴吸收国外先进经验, 为我国全面开展科学数据共享建设提供路径选择和方法支持。

1 基于数据知识库的科学数据共享方案

数据知识库(又称“数据仓储”“数据存储库”“数据资源库”等), 常用于存储和管理科学数据、支持科研活动及其知识创造。本文所指数据知识库主要包括机构知识库、学科知识库、项目知识库、综合知识库。机构知识库是从学术机构的角度组织和揭示数字化研究成果, 学科(主题)知识库是从学科(主题)的角度组织和揭示数字化研究成果。从发展历史看, 学科知识库的起步比机构知识库更早; 从发展规模看, 目前机构知识库的数量远超学科知识库。机构知识库与学

科知识库之间也存在交叉关系。

数据知识库从建库单位看, 分为机构独立建库和多机构联合建库; 从建库内容看, 分为单学科建设和多学科建设。随着“大数据”和“科学知识库2.0”的提出, 数据知识库将朝着综合库方向发展, 其建设内容也将向研究基础设施及ICT服务方面发展。

1.1 机构知识库

虽然学术界对机构知识库(Institutional Repository, IR)从不同维度进行阐释, 但学术出版和学术资源联盟对机构知识库的定义被很多学者接纳和认可, 认为“机构知识库搜集并保存学术机构智力成果, 促进学术交流, 同时也展现机构的学术价值”^[1]。2001年以来, IR在全球范围内迅速发展, 成为机构或领域数字知识创造与传播流程中的有机环节, 并朝着以科学研究为核心的整体数字信息资源与服务环境方向发展。目前, 以高校为代表的嵌入机构内部数据监管方案仍为主流的知识库建设方案, 随着需求增加, 数据暂存库方案、跨学科、可持续的机构数据监管方案不断涌现。

^{*} 本研究得到中国农业科学院科技创新工程(编号: CAAS-ASTIP-2016-AII)和大数据环境下农作物育种领域创新方法研究与示范(编号: 2015IM010400)资助。

1.1.1 嵌入机构内部的数据监管方案

2008—2012年牛津大学通过SDRSRD、EIDCSR、SUDAMIH、VIDaaS、ADMIRAL、DataFlow、DaMaRO等项目建设机构数据管理基础设施, 逐步形成一个模块化与联合化的数据管理基础设施方案, 并整合一系列服

务支撑数据管理全过程, 包括政策、培训、可持续发展、数据管理计划工具、数据管理与培训、数据创建与本地管理、数据存档与监护、数据发现与传播等服务(见图1)。底层技术架构建设基于科研人员的数据管理需求, 并支撑数据管理生命周期的数据服务。其中DataFinder为技术架构核心, 连接架构中的各独立模块^[2]。

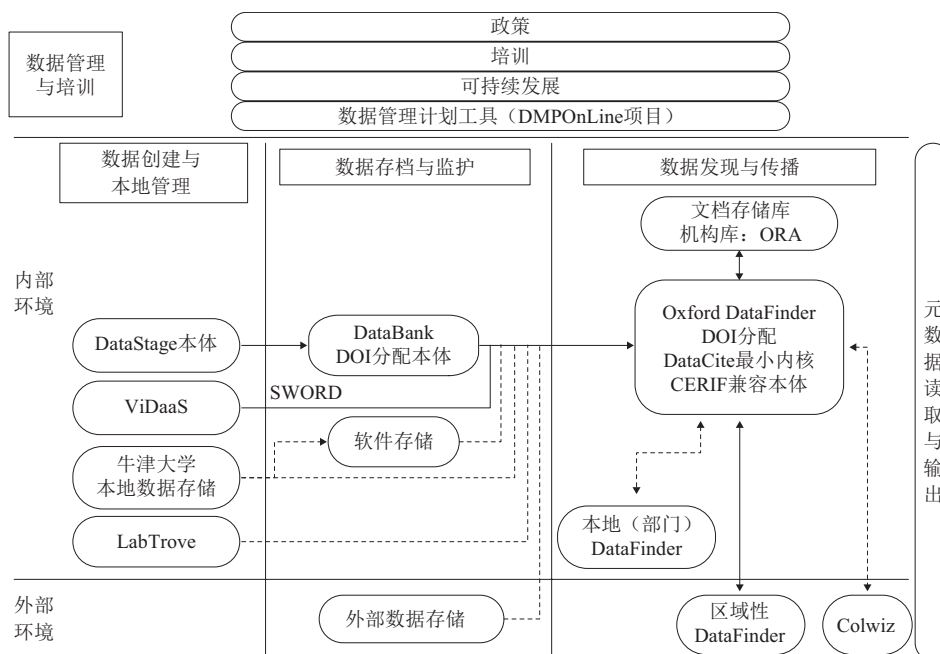


图1 牛津大学综合数据管理基础设施架构^[3]

1.1.2 数据暂存库方案

2007年, 康奈尔大学启动DataStaR (Data Staging Repository) 项目, 其定位于数据阶段型存储库, 以机构库为基础建设数据监护平台及完整服务方案, 通过集成不同领域特定的元数据模式, 提供跨学科支持; 通过多种格式的高质量元数据工具将数据由暂时、过渡性质的存储节点流动到各领域的长期存储库, 以供长期使用和保存(见图2)。DataStaR兼有一般机构库和学科库的特征。

次结构、模块化架构来规避底层系统的复杂性, 并通过多元协作机制实现大学图书馆、国家数据中心、国家实验室等不同机构与不同学科间资源共享(见图3)。

1.1.3 跨学科、可持续机构数据监管方案

2009年, 由约翰霍普金斯大学承接的Data Conservancy项目, 致力于构建一个能对复杂化和集成化科学数据进行持续性、跨学科、跨机构管理的数据监管平台, 以解决科研过程中面临的数据挑战, 是对基于数据监管的机构库建设的有益探索^[5]。Data Conservancy采用层

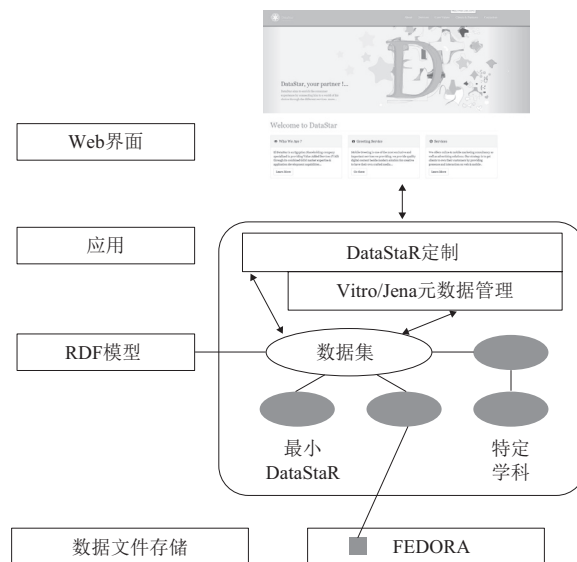
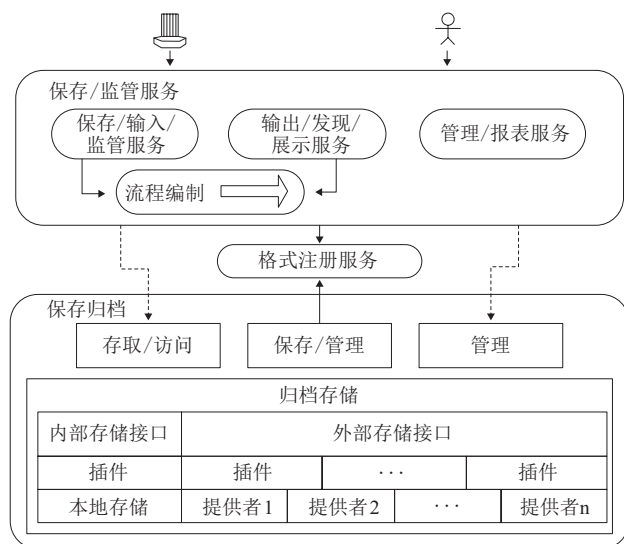


图2 DataStaR架构^[4]

图3 Data Conservancy架构^[6]

1.2 学科知识库

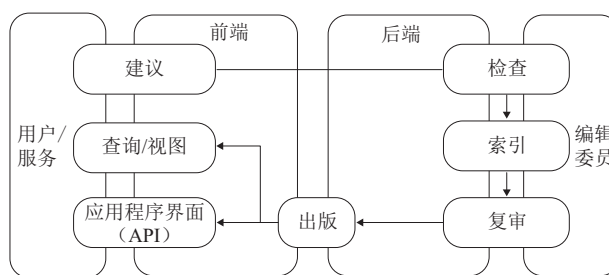
国外学科知识库大多基于需求由官方专业机构、协会、社区等多个单位联合组建，集中在单个或少数领域专业性较强。

2 基于数据知识库和开放数据集的注册与目录系统的共享方案

2006年，Oliver等对在线国际机构库目录进行汇总，包括现在仍广泛提供服务的Directory of Open Access Repositories (OpenDOAR) 和Registry of Open Access Repositories (ROAR)^[7]。随着数据知识库、开放科学数据集的日益增多，出现基于语义关联、本体等新技术数据知识库和开放科学数据集注册与目录系统。

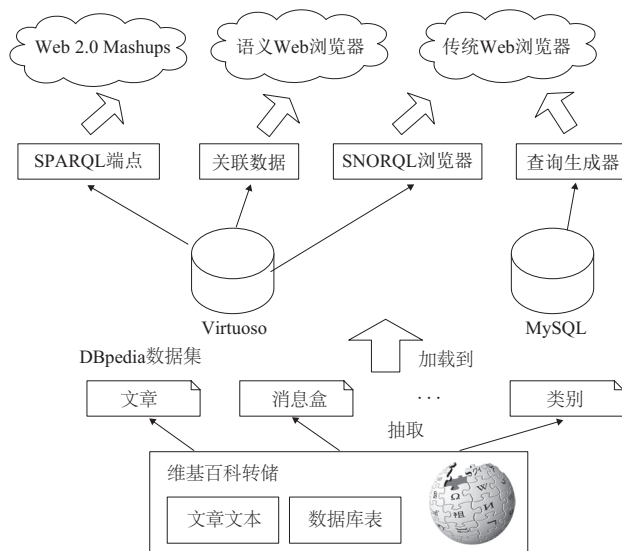
2.1 全球科学数据知识库注册表

全球科学数据存储库注册表re3data.org作为可靠的科学数据服务新兴基础设施，涵盖所有学科的科学数据知识库，并提供科学数据可见性服务，促进文化共享。其方案如图4所示，re3data.org提供了有关科学数据知识库的详细信息，并通过独特的图标帮助科研人员识别相关知识库，以进一步访问和选择数据集。2013年3月，re3data.org和Databib宣布合并，继续提供服务。截至2017年3月3日，re3data.org共收录1 821个数据仓储^[8]。

图4 re3data.org方案^[9]

2.2 基于关联数据的语义知识图

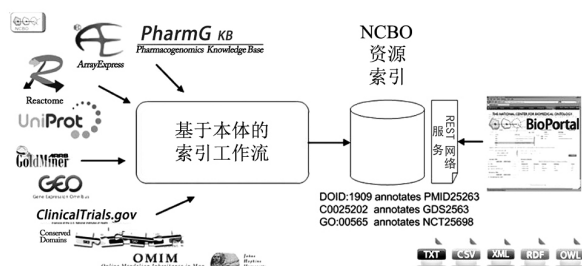
关联数据指在语义网使用统一资源标识符和资源描述框架发布、分享、连接各类数据、信息和知识，通过HTTP协议揭示并获取数据^[10]。DBpedia是基于维基百科、语义Web和关联数据技术的创新型知识库，是数据网的代表之一^[11]。DBpedia从Wikipedia抽取结构化数据并将抽取内容转变为结构化知识，在关联开放数据原则指导下与网络其他数据互联，使数据以关联数据的形式网络共享（见图5）。

图5 DBpedia组建方案^[12]

2.3 基于本体的资源索引

国家生物医学中心的本体资源索引(National Center for Biomedical Ontology Resource Index, NCBO Resource Index)目的是共享和应用生物医学领域不断增加的公开数据，其基于本体语义编码提供探索性、多术语、语义增强的资源搜索，其方案如图6所示。NCBO Resource Index应用基于现有541个本体，使用基于本体的索引工

作流集成生物医学资源, 其将资源索引存储在数据库中, 通过REST Web对外提供多种服务, 并基于BioPortal门户提供搜索和导航资源索引^[13]。

图 6 NCBO Resource Index 方案^[13]

3 基于机构知识库联盟的共享方案

机构知识库联盟是一定区域范围内不同成员机构为更好地促进相互间的学术交流与合作,通过知识跨库检索与内容交换,实现数据备份、保存、故障恢复等功能,从而构建的一种共享联盟机制。从资源获取角度,IR联盟模式包括集中式、分布式、综合式三种。

3.1 集中式IR联盟模式

此模式下联盟机构成员将元数据和数字内容提交到集中的服务器,进行统一保存、管理、发布和利用,集中管理联盟机构数字资产。采用这种模式的机构知识库联盟案例有英国白玫瑰研究在线^[14]、美国华盛顿研究图书馆联盟数字知识库^[15]、芬兰应用科学院的Theseus^[16]。

3.2 分布式IR联盟模式

此模式下联盟机构成员按照共同的数据交互标准和协议构建自身机构库,通过收割元数据建立统一资源检索平台。采用这种模式的机构知识库联盟案例有美国校际社会科学数据共享联盟、美国HELIN数字知识库^[17]、俄亥俄州数字知识库联盟^[18]、英国JISC资助的ePrints UK项目^[19]。

3.3 综合式IR联盟模式

此模式介于集中式和分布式模式间或是两种模式的综合。采用这种模式的机构知识库联盟案例有日本山

口县大学共享知识库、美国德州数字图书馆知识库^[20]、芬兰Doria^[21]。

4 基于科学数据基础设施的共享方案

科学数据基础设施是一个包含软件与硬件,支持科学数据分享与重用,促进全球科研合作,加速科学创新的数据生态系统。美国联邦研究的资助者使用信息基础架构来描述一种研究环境,可以支持先进的数据采集、数据存储、数据管理、数据集成、数据挖掘、数据可视化,以及其他分布在互联网上超出单个机构范围的信息处理服务^[22]。当前各国科学数据基础设施建设大多基于学科领域,且朝着支持世界范围内跨学科合作与共享的全球化科学数据基础设施方向发展。

4.1 生物多样性虚拟电子实验室

由“欧盟第七研发框架计划”资助,卡迪夫大学以及欧洲15个伙伴机构共同建立的生物多样性虚拟电子实验室(Biodiversity Virtual e-Laboratory, BioVeL),通过强大的科学基础设施实现海量、跨学科数据的收集、组织、共享处理、分析和建模,为探索科学家提供开放获取和研究的工具和软件环境,支持生物多样性数据密集型科学,促进国际社会对生物多样性的合作研究^[23]。由图7可知,BioVeL基础设施包括:生物多样性目录,提供数据注释分析服务;R studio构建R程序环境;Taverna workbench,创建数据访问和分析流程的工作台;工作流库,存储现有的工作流;BioVeL门户,用于科研人员和合作者执行、管理和共享工作流;文本组件。以上每个基础设施均可独立使用或者作为整个基础设施的一部分,已经建立Biome-BGC生态系统指标监管工作流等多个生物学不同领域的工作流供科学家使用。

4.2 欧洲协作数据基础设施

欧洲协作数据基础设施 (European Collaborative Data Infrastructure, EUDAT) 作为面向欧洲的科学社区, 汇集了来自15个国家的35个联盟伙伴 (包括研究社区、国家数据和高性能计算中心、技术提供商、资助机构等)。作为一个以用户为驱动, 可信、安全、跨学科和可持续的通用数据基础设施, 旨在为来自任何学

科的欧洲科研人员提供一个可信的环境,提供查询、存储、复制、暂存、共享和计算等一系列研究开放数据服务,使科研人员在社区内共享数据,有效地进行研究,其建设方案如图8所示。

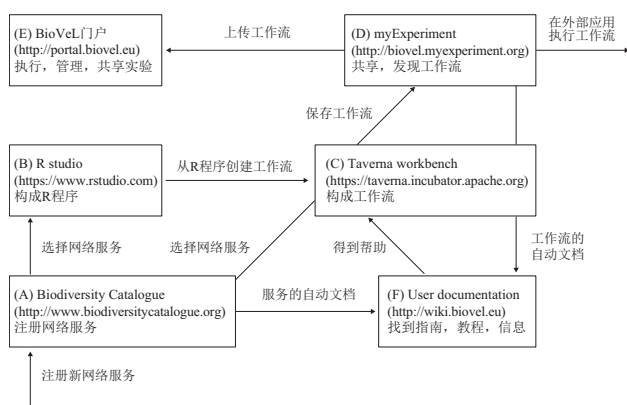


图7 BioVeL方案及流程^[24]

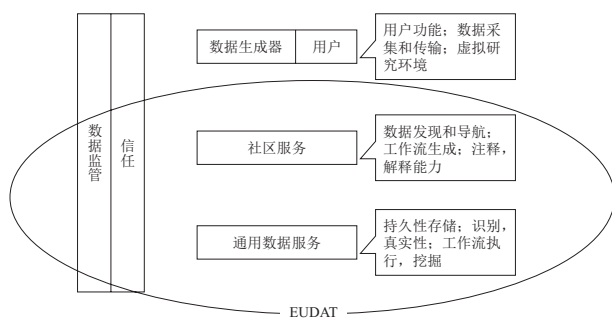


图8 EUDAT方案^[25]

5 国外科学数据共享方案分析及启示

5.1 国外科学数据共享方案特点分析

国外科学数据共享案例种类繁多,大多基于某个学科领域、单个或多个机构(以高校居多),应用最终以数据知识库和门户形式呈现,体现各自的需求和特点,不具有普适性。共享案例积极探讨了科学数据管理与共享新技术;重视跨学科领域的科学数据共享及利益相关者的多元合作;注重国际交流与标准化;持续改进共享平台功能,不断提高共享数据服务能力。

在基于“数据知识库—数据目录系统—机构知识库联盟—科学数据基础设施”的科学数据共享方案演变过程中,方案总体呈现如下特点:建设对象由大科学数据延伸到长尾数据;建设内容由通用的基础设施向专业的基础设施转变;建设方式由单个项目建设向联

合建设转变;发现机制由检索页面向集成目录变化;应用形式由数据检索向数据重用拓展,共享工作由科学研究流程下游向上游转移。

5.2 从服务角度进行国外科学数据共享方案比较分析

面向数据生命周期,总结四类共享方案提供的科学数据服务,主要包括管理规划服务、数据存储服务、数据发布服务、数据发现服务、数据展示服务、数据引用服务、数据访问服务、信息处理服务、培训服务。将四类共享方案从服务角度进行比较分析,有助于利用其构建各具特色的数据服务。由表1可知,数据收集、数据描述、数据发布、数据发现、数据访问为数据共享方案通用服务;基于数据知识库的共享方案更侧重于管理规划、培训服务;基于数据知识库和开放数据集的注册与目录系统共享方案侧重于数据发现、数据访问服务;基于科学数据基础设施的共享方案更侧重于数据分析、可视化服务。

5.3 国外科学数据共享方案启示

(1) 准确把握共享需求,制定共享方案。上述四类共享方案的实质是满足不同的数据共享需求。本文所指需求包括功能性需求、非功能性需求(性能需求、遵循标准等)以及设计约束等。因此,我国在制定共享方案时,需要在数据共享范围和对象既定的环境下,准确把握共享需求,选择共享模式和技术制定共享方案。还可借鉴上述方案提出的理论方法,如科学数据共享概念框架、科学数据管理生命周期理论、需求调查方法(访谈、问卷等)、数据资产框架^[26]、数据可重用性评估框架^[27]、数据重用研究框架、理想数据共享处理框架等。

(2) 创新数据服务力,促进共享发展。数据相关者对科学数据服务需求的不断增加,促使数据服务向数据全生命周期拓展,逐渐成为数据共享方案支撑的重要部分。上述方案在创新服务模式和内容的基础上,开发了灵活的数据服务工具来支持科学数据的处理、分发与出版、检索、分析及集成应用。这些在实践中产生的工具,绝大部分为开源软件,可供国内共享方案整合利用,能在一定程度上减轻科研人员工作量,提升其服务能力,拓宽与国外科学数据共享研究的交流渠道。

表 1 从服务角度进行国外科学数据共享方案比较分析

	管理 规划	数据 收集	数据 描述	数据 存储	数据 标识	数据 发布	数据 引用	数据 发现	数据 访问	数据 分析	可视化	培 训
DataStaR方案	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Data Conservancy方案		✓	✓	✓		✓	✓	✓	✓			✓
牛津大学综合基础设施方案	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓
re3data.org方案		✓	✓		✓	✓	✓	✓	✓			
DBpedia方案		✓	✓	✓	✓	✓	✓	✓	✓			
NCBO Resource Index方案		✓	✓		✓	✓	✓	✓	✓			
EUDAT方案		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓
BioVeL方案		✓	✓			✓		✓	✓	✓	✓	

注：“✓”表示科学数据共享方案包含此项服务。

(3) 研究可持续性策略, 保持方案生命力。数据共享可持续策略是当前共享方案面临的问题和挑战。上述方案从技术可持续策略和经济可持续策略两方面探讨可持续的长期发展策略。其中, 技术可持续策略体现在技术可靠性、适应性、可扩展性, 以及经济适用的成本模式研究。如长期数字保存成本模型^[28]、英国大学的成本模型^[29]等。经济可持续策略方面, 由于目前科学数据共享建设主要为项目投资, 可持续性差, 因此需要不断创新经费来源模式, 提出并实践会员制、增值性服务等新模式。这些研究成果为我国数据共享可持续策略研究奠定基础。

6 结语

本文在文献调研基础上, 从技术角度系统梳理国外科学数据共享实践, 除上述归纳的四类主要共享方案外, 还包括数据出版架构、个人或团队网站共享方案等非主流实践。与国外相比, 国内有代表性的共享方案主要是综合性科学数据共享平台(如科学数据共享工程系列平台、中国科学院数据云、复旦大学社会科学数据平台等高校建设的共享平台), 在共享方案制定所需理论方法、共享数据服务工具、共享可持续发展策略等方面探讨较少。因此, 本研究为我国数据共享工作的进一步开展提供建设思路和技术支撑, 促进我国科学数据共享建设。

参考文献

[1] SPAR C.The case for institutional repositories:a SPARC position paper[EB/OL].[2016-12-10].http://sparc.arl.org/resources/papers-guides/the-case-for-institutional-repositories.

[2] WILSON J A J,MARTINEZ-URIBE L,FRASER M A,et al.An institutional approach to developing research data management infrastructure[J].The International Journal of Digital Curation,2011,6(2):274-287.

[3] RUMSEY S.Building a research data catalogue:knowing what you've got and where it is data finder at oxford[EB/OL].[2016-08-20].http://damaro.oucs.ox.ac.uk/outputs.xml.

[4] KHAN H,CARUSO B,CORSON-RIKERT J,et al.DataStaR:using the semantic web approach for data curation[J].The International Journal of Digital Curation,2010,6(2):209-221.

[5] History of DC[EB/OL].[2016-01-09].http://dataconservancy.org/about/historyofdc/.

[6] CHOUDHURY S,HANISCH R.The data conservancy:building a sustainable system for interdisciplinary scientific data curation and preservation[EB/OL].[2017-03-20].https://jscholarship.library.jhu.edu/bitstream/handle/1774.2/34018/47_Choudhury_DataConservancy.pdf.

[7] OLIVER K B,SWAIN R.Directories of institutional repositories: research results & recommendations[EB/OL].[2017-04-02].http://www.ifla.org/IV/ifla72/index.htm.

[8] re3data.org[EB/OL].[2017-01-22].http://www.re3data.org/search.

[9] ULRICH R.Re3data.org:making research data repositories visible and discoverable[EB/OL].[2017-03-03].https://www.fosteropenscience.eu/sites/default/files/pdf/1374.pdf.

[10] BERNERS-LEE T.Linked data[EB/OL].[2017-02-22].http://www.w3.org/DesignIssues/LinkedData.html.

[11] The DBpedia data set[EB/OL].[2016-12-05].http://wiki.dbpedia.org/Datasets.

[12] AUER S,BIZER C,KOBILAROV G,et al.DBpedia: a nucleus for a web of open data[C]//The Semantic Web.Lecture Notes in Computer Science.Berlin:[s.n.],2007,4825:722-735.

[13] JONQUET C,LEPENDU P,FALCONER S,et al.NCBO resource

- index:ontology-based search and mining of biomedical resources[J].
Web Semantics:Science Services & A Gents on the World Wide
Web,2011,9(3):316-324.
- [14] White rose research online[EB/OL].[2016-08-15].<http://eprints.whiterose.ac.uk/>.
- [15] WRLC digital repository[EB/OL].[2016-09-15].<http://dspace.wrlc.org/>.
- [16] Theseus.fi[EB/OL].[2016-06-15].<http://publications.Theseus.fi/>.
- [17] HELIN digital commons[EB/OL].[2016-09-15].<http://helindigital-commons.org/>.
- [18] OhioLINK digital resource commons[EB/OL].[2016-08-10].<http://drc.ohiolink.edu/>.
- [19] Jisc[EB/OL].[2017-01-02].http://www.jisc.ac.uk/index.cfm?name=fairsynthesis_eprintsuk.
- [20] Texas digital library[EB/OL].[2016-07-18].<http://www.tdl.org/repositories/>.
- [21] Doria home[EB/OL].[2016-08-13].<http://www.doria.fi>.
- [22] Wikipedia.Cyberinfrastructure[EB/OL].[2016-12-15].<http://en.wikipedia.org/wiki/Cyberinfrastructure>.
- [23] BioVeL.Biodiversity virtual e-Laboratory[EB/OL].[2017-01-15].
<http://www.biovel.eu/>.
- [24] HARDISTY A,BACALL F,BEARD N,et al.BioVeL:a virtual laboratory
for data analysis and modelling in biodiversity science and ecology[J].
Bmc Ecology,2016,16(1):49.
- [25] REETZ J.EUDAT-open data services for research[2017-02-22].
https://zenodo.org/record/34650/files/Reetz_SciOps2015.pdf.
- [26] Research data management project:a DAF investigation of research
data management practices at the University of Northampton[EB/OL].
[2017-02-05].<http://nectar.northampton.ac.uk/2736/1/Alexogiannopoulos20102736.pdf>.
- [27] Reusing scientific data:a research framework[EB/OL].[2016-11-15].
http://www.sci.utah.edu/images/Workshops/cscw2010/faniel_2.pdf.
- [28] SHEHAB E,LEFORT A,BADAWY M,et al.Modelling long term
digital preservation costs:a scientific data case study[C]//the 11th
International Conference on Manufacturing Research.[S.l.]:[s.n.],
2013:133-138.
- [29] BEAGRIE N,CHRUSZCZ J,LAVOIE B. Keeping research data safe:
a cost model and guidance for UK universities[EB/OL].[2016-11-02].
<http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf>.

作者简介

彭秀媛,女,1986年生,博士,助理研究员,研究方向:农业科学数据共享、农业物联网技术,E-mail: peng_xiuyuan@163.com。

王枫,男,1982年生,硕士,讲师,研究方向:数据管理、数据分析,E-mail: delwww@126.com。

周国民,男,1969年生,研究员,博士生导师,研究方向:农业科学数据共享系统与技术、农业网络信息智能搜索、农业远程诊断方法与技术、果树信息化,E-mail: zhouguomin@caas.cn。

Research on Foreign Scientific Data Sharing Scheme

PENG XiuYuan^{1,2}, WANG Feng³, ZHOU GuoMin¹

(1.Agricultural Information Institute, CAAS, Beijing 100081, China; 2.Information Center, LAAS, Shenyang 110161, China;

3.Information Center, Party School of the Liaoning Committee of CPC, Shenyang 110004, China)

Abstract: The demand and form of scientific data sharing are expanding, and the data sharing scheme presents a trend of diversification. The scientific data sharing scheme has an important guiding role in the practice of scientific data sharing. From the angle of technical, article systematic study the existing scientific data sharing scheme, which is divided into four categories: sharing scheme based on the data repositories, sharing scheme based on the registration and directory system of the data repository and open data sets, sharing scheme based on institutional repository alliance, and sharing scheme based on scientific data infrastructure. Based on each type of solution and technology analysis of typical cases, analysis of the characteristics of the above scheme. From the aspect of service, comparative analysis the representative case. Summarized the revelation to provide the construction ideas for the further development of data sharing.

Keywords: Scientific Data; Sharing; Data Repositories; Directory System; Repository Alliance; Infrastructure

(收稿日期: 2017-07-12)