生命科学领域科研数据仓储特点及服务分析*

邹丽雪 欧阳峥峥 王辉 吴鸣
中国科学院文献情报中心 北京 100190

摘要: [目的/意义] 对生命科学领域的科研数据仓储进行调研与分析,探讨生命科学领域的科研数据管理服务。[方法/过程] 利用 re3data. org 开放数据仓储目录与注册系统,分析生命科学领域科研数据仓储的建设年代、国家、机构、学科领域、开放程度等分布情况,并选取 Genbank、Dryad、ArrayExpress、Purdue University Research Repository、Biosharing 和 dbGaP 6 个典型的数据仓储,从数据获取、重用、存储等方面深度分析其服务内容和模式。[结果/结论] 美英两国引领着生命科学领域科研数据仓储的建设与共享,在国家层面和资助机构层面均制定了科研数据相关政策; 国内可借鉴美英两国成熟的建设经验,加快制定战略规划和政策体系; 资助机构应发挥引导作用,在服务内容及模式上推动数据管理与共享,建设具有领域特色的高影响力的数据仓储并集成数据管理服务。

关键词: 科研数据仓储 生命科学 科研数据管理 科研数据服务
分类号: G239. 1
DOI: 10. 13266/j. issn. 0252 - 3116. 2016. 07. 009

随着科学研究向数据密集型科研快速发展,科研数据共享与重用的需求日趋强烈,科研数据仓储(data repository,DR) 作为科研数据存储、发布、开放共享的途径之一,得以快速发展,国际上一些基金资助机构、科研机构、期刊均纷纷制定数据相关政策,要求将科研数据提交至相应的数据仓储。科研数据仓储通过一定的数据提交机制,并进行数据质量审核,组织存储数据,同时对数据共享重用提出明确的规范要求。面向不同的学科领域和数据形态,科研数据仓储有着不同的特点[1],H. J. Nielsen 等的研究中提到应对科研数据进行学科领域的分析研究,以更好地提供数据服务[2]。

在生命科学领域,2005 - 2015 年这 10 年间,随着高通量测序技术的的快速发展与应用,在以基因组学和生物医学领域为代表的科研过程中产生了大量的数据,欧洲生物信息学研究所作为全球最大的生物数据库之一,目前已存储了 20PB(2×10^{16}b) 的数据,每年以 200% 的速率在迅速增长[3]。数据的膨胀驱动了数据仓储的快速发展,刘峰等[4]对科研数据仓储注册目录系统 databib 中注册的数据仓储按学科领域进行了分析,指出生物学领域数据仓储分布较为广泛。本文聚焦于生命科学领域的科研数据仓储,系统地分析其建设现状、特点及服务内容,并从政策制定与管理、建设内容、服务层面提出建议和对策。

1 生命科学领域科研数据的特点

生命科学领域的科研数据可分成通用数据和专门数据[5]。通用数据是指生物体或组织的核酸、基因、蛋白质序列等数据,具有量大、稳定、使用频率高等特点。专门数据是指特定主题的实验或临床所获取的数据,数据量少、变化较快、获取难度大并在一定程度上不可重复。该领域的科研数据特点与传统大数据特点一致,表现为数据量大、处理数据速度要求快、数据多源异构、数据整合分析复杂[6]。

本文以生命科学领域的科研数据仓储为研究对象,基于 Re3data. org 科研数据仓储注册与目录系统进行分析。截至 2015 年 8 月 20 日,该系统共注册了 1 314 个数

* 本文系中国科学院“科学数据管理服务试点建设”(项目编号: 院 1522) 和中国科学院文献情报中心青年人才项目“多源开放数据在科研领域的应用研究”(项目编号: 青 1305) 研究成果之一。
作者简介: 邹丽雪(ORCID: 0000 - 0002 - 2617 - 4151) ,馆员,硕士,Email: zoulx@ mail. las. ac. cn; 欧阳峥峥(ORCID: 0000 - 0001 - 5941 - 3561) ,馆员,硕士; 王辉(ORCID: 0000 - 0002 - 3775 - 3992) ,副研究馆员,博士; 吴鸣(ORCID: 0000 - 0002 - 5506 - 8657) ,学科咨询服务部主任,研究馆员。
收稿日期: 2016 - 02 - 03 修回日期: 2016 - 03 - 20 本文起止页码: 59 - 66 本文责任编辑: 杜杏叶

据仓储，其中生命科学领域（Life Science 分类）共 653 个，本文选定这 653 个数据仓储，对其建设时间、国家、机构、学科领域、开放程度进行统计，并对缺失的数据进行补充，系统地梳理其建设现状及特点，并就国家、科研机构及基金资助机构层面制定的政策进行解析。根据生命科学领域高影响力期刊 *Nature*、*Science* 指定用于存储不同类型数据的仓储，从中选取 6 个典型的数据仓储作为实例，深度分析其具体服务内容及模式。

2 生命科学领域数据仓储特点分析

2.1 建设年代分布

对 653 个生命科学领域数据仓储的建设时间进行分析（435 个数据仓储有明确的建设年代），发现最早的数据仓储建于 1903 年，是美国人口调查局建设的“United States Census Bureau”[7]。从建设年代分布看，1988 年之前只是零散地建立个别数据仓储，1988 – 1999 年间，数量波动性上升，2000 年大幅增多，在 2006 年达到峰值。数据仓储建设年代的转折点均与生命科学领域研究技术的突破以及研究数据的发展息息相关。如 1988 年美国联邦报告批准了人类基因组计划，建立了 NCBI[8]，为今后科学数据仓储的建设打下了基础；2000 年，人类基因组草图被绘制完成，随后便建立了 ArrayExpress[9]、dbSNP[10]、GBIF[11]、Worldwide Protein Data Bank[12] 等具有重要影响的数据仓储；2005 年之后，随着高通量测序技术的广泛应用，生命科学正式进入大数据时代，为满足新型大数据的需求，NIH 迅速建成了 dbGaP[13]、Sequence Read Archive[14] 等新型的数据仓储；2007 年之后，数据仓储的建设数量呈下降趋势，这可能与前期数据仓储趋于饱和与成熟且该领域暂未出现新型数据有关。图 1 以 1986 年之后建设的数据仓储为例，展示了其建设年代的分布情况：

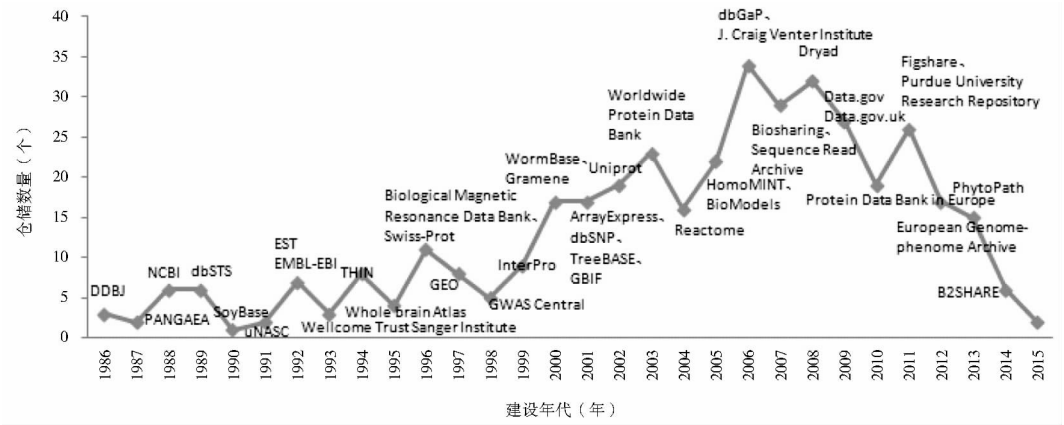

图 1 生命科学领域科研数据仓储建设时间分布

2.2 建设国家分布

对数据仓储建设国家进行分析（见图 2），美国、英国、德国在该领域有较大的优势，其中仅美国单独建设的就有 347 个，占所有数据仓储的一半以上，美国和英国合作建设的有 52 个，如 GenBank[15]、Dryad[16]、Protein Data Bank in Europe[17] 等典型数据仓储。在全球范围内，美国和中国是生命科学领域科研数据的产出大国[18]，但中国数据仓储的数量很少，仅有 19 个，且影响力远不如上述数据仓储。

美英两国引领着全球开放数据仓储的建设与共享，这与美英两国在国家层面制定了一系列与科学数据相关的的战略和政策息息相关[19-21]，具体见图 3。

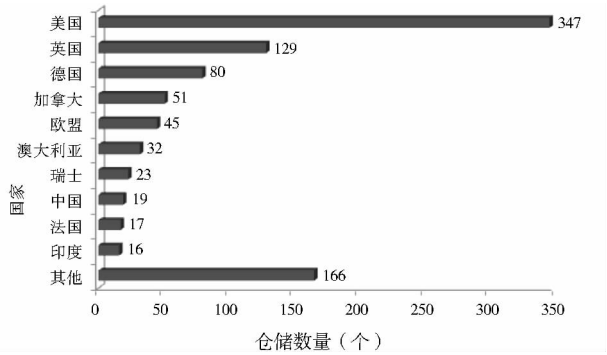

图 2 生命科学领域科研数据仓储建设的国家分布

2.3 建设机构分布

从数据仓储的建设机构看，美国国立卫生研究院在建设数量上具有明显的优势，且前 10 家机构中美国的基金资助机构和大学较多。英国建设数据仓储最多的机构是惠康基金会，德国则是马普学会。表 1 中的 10 家基金资助机构和研究机构大多都已制定了详细的数据政策作为其建设和服务的保障。

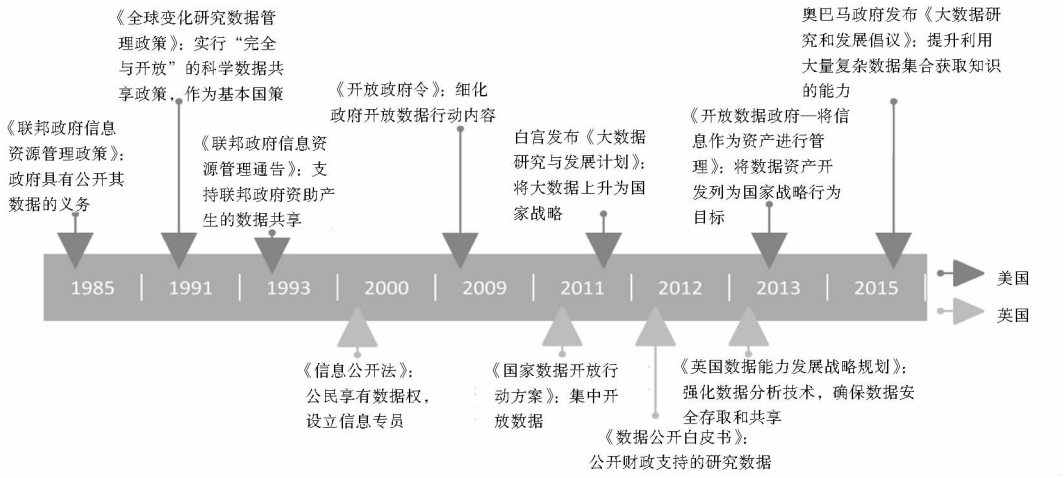

图3　美英两国开放数据政策一览

表1　数据仓储建设机构分布(部分)

建设机构	数量(个)	数据政策
National Institutes of Health(美国国立卫生研究院)	203	2003年发布“数据共享政策和实施指导”，要求：①从2003年10月1日开始所有向NIH申请经费在50万美元以上的科研人员都必须提交一个数据共享管理计划或者数据不共享的说明。②申请者可申请数据管理的经费支持。③NIH要求在项目完成后数据分种类汇交保存3年。④共享方式包括个人网站、机构网站或存储到数据库。⑤由项目工作人员对数据管理计划的实施进行评估[22]
European Molecular Biology Laboratory(欧洲分子生物学实验室)	63	欧洲生命科学生物信息基础设施发布“ELIXIR2014－2018计划”，旨在为生物信息建立一套欧洲的有效的数据基础设施，计划内容包含：①建立分布式基础设施，确保核心数据源能安全提供，提供有效工具控制数据访问。②发展数据管理和重用标准。③提供技术平台如云服务。④与用户社区合作确保高效可持续服务，提供综合系统的培训，支持生物大数据的创新[23]
National Science Foundation(美国国家科学基金会)	51	①NSF 2010年发布“项目管理指南”，要求：从2011年1月18日起，所有项目申请书中必须包含一份不超过两页的数据管理计划，并制定了通用的数据管理计划模板。②其下属的生物科学研究理事会NSF-BIO根据学部特殊性细化数据汇交政策，形成生命科学部的数据管理计划模板，管理计划中应详细描述数据收集、格式及采用标准、数据存储和长期保存的策略及设施、数据传播方式及载体、数据共享与公开获取政策、数据管理中的角色与责任[24]
European Commission(欧盟委员会)	48	GRDI2020项目发布了“全球科学数据基础设施：重大数据挑战”报告，提出了构建全球科学数据基础设施的政策挑战和建议，包括：①须开发科学数据基础设施支持开放链接的数据空间、支持数据与文献间的互操作、支持数据密集型研究、支持多学科与跨学科的研究及科学生态系统。②须开发新型数据工具、数据模型及查询语言。③培养专业的数据专家和研究团队[25－26]
U. S. Department of Health & Human Services(美国卫生和公共服务部)	40	2014年发布“科学数据开放获取政策”，提供指南和框架，用以帮助制定设施计划，共享联邦资助的数据，制定数据行动计划，研究数据元素的标准化及定义[27]
United States Department of Agriculture(美国农业部)	37	2014年制定“OneUSDA数字战略”，内容：①确保数据开放、准确、清晰地描述、结构化，机器可读、支持移动端使用数字化服务。②使用专业数据分类标准对常用数据进行标准化。③提供政策方面的指导，制定开放数据策略描述共享的信息框架。④建立虚拟动态开放数据库[28]。另外，制定了“开放数据政策传播计划”，推动开放数据及Web API共享[29]
University of California(美国，加州大学)	26	①提供数据管理指南，详细地描述数据类型、文件格式、元数据、数据安全存储、共享、引用规范。②开发了数据管理计划在线撰写工具DMP TOOL，通过向导的方式协助创建基金资助机构要求的数据管理计划，提供协同式工作环境，进行DMP的交流与共享
University of Georgia(美国，佐治亚大学)	18	图书馆开展社会科学数据相关服务，面向大学提供数据管理、工具、仓储、访问及指导撰写数据管理计划
Wellcome Trust(英国，惠康基金会)	17	2010年发布“数据管理和共享政策”，要求：①申请资助项目时需提交数据管理与共享计划。②数据共享的成本由项目提供经费。③共享计划的实施由资助机构进行验收、评估和监督。④用户使用数据时应注明数据来源，遵从原始数据要求的条款[30]
Max Planck Digital Library(德国，马普学会)	14	①2003年马普学会在柏林会议上通过《柏林宣言》指出开放获取的内容不仅包括原始的科学研究成果，还应包括科学数据[31]。②规定各研究所应对科研数据的安全和保存制订详细的指南，原始数据应由学会或研究所至少保留10年[32]

2.4 学科领域分布

目前，数据仓储主要分为机构仓储、学科仓储、多学科仓储以及特定项目仓储这 4 种类别，653 个数据仓储中有 529 个为多学科类别，占 81.01%，这表明该领域的数据仓储在开放性和学科领域的广度上具有优势。具体学科领域分布见图 4。其中，医学、基础生物和医学研究领域数据仓储数量最多，而农林园艺、神经科学方面的仓储数量相对较少。

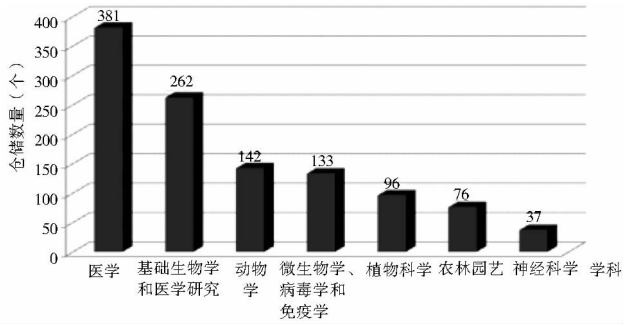

图 4 生命科学领域数据仓储的学科领域分布

2.5 数据仓储开放程度

数据仓储的开放程度可从数据获取、重用、存储 3 方面进行评价，开放程度分为完全开放、受限制、不开放 3 个层次。653 个数据仓储中，88.82% 支持完全开放获取，66.16% 支持完全开放重用，7.20% 支持完全开放存储（见表 2）。可见支持开放存储的比例还很小，分析其原因，一则可能与该领域数据仓储多是由基金项目资助建设的数据库，存储数据基本上都是项目产生的数据有关；另外，除非基金资助要求或期刊要求，科研人员主动存储和共享数据的意愿还不是太强烈；再者，开放存取对数据类型、格式和质量等要求较高，许多数据仓储尚未制定相应的政策保障机制。

表 2 生命科学领域数据仓储的开放程度

开放程度	获取	重用	存储	仓储案例
完全开放（不受任何限制）	88.82%	66.16%	7.20%	PANGAEA [33]、Figshare [34]、Protein Data Bank in Europe [17]
受限制（需要注册或进行身份认证）	10.87%	33.84%	58.96%	European Genome-phenome Archive [35]、ComBase [36]
不开放（限机构或项目内部人员）	0.31%	–	33.84%	United States Transuranium & Uranium Registries [37]

3 生命科学领域数据仓储实例

参考 *Nature*、*Science* 等期刊对存储不同类型的科研数据指定的数据仓储，如将 DNA 和 RNA 序列存储至 Genbank、DNA DataBank of Japan（DDBJ）等，分子结构数据存储至 Worldwide Protein Data Bank（wwPDB）等、微阵列数据存至 Gene Expression Omnibus（GEO）、ArrayExpress，生态数据存至 Dryad，基因型和表型数据存储至 dbGap 等，这些数据仓储都是存储生命科学领域各种类型数据的典型代表。此外，结合一些特色数据仓储如嵌入了数据管理计划服务的 Purdue University Research Repository（PURR），集成了生命科学领域注册标准的 BioSharing，本文选取了 6 个数据仓储为典型代表，深度分析其具体服务内容与模式。

3.1 GenBank

GenBank 是美国国立卫生研究院建立的基因序列数据库，用于收集所有公开可获取的 DNA 序列，数据主要来源于作者直接提交或由生物医学相关的文献中检索已发表的序列数据，并与日本的 DDBJ（DNA DataBank of Japan）和欧洲的 EMBL（European Molecular Biology Laboratory）每天进行数据交换。

（1）GenBank 的数据可用 Entrez 检索核苷酸序列标识符和注释，可由序列数据检索至相关的蛋白质序列、三维结构及文献数据，提供 Blast 进行序列局部比对检索，数据可通过 NCBI 提供的 e-utilities 以及 FTP 服务器下载。

（2）在数据重用上没有任何限制。

（3）GenBank 接受科研人员直接提交序列数据，数据内容必须包括源生物信息和注释信息，提交方式可通过 BankIt（以 www 表格在线提交）、Sequin（输入数据处理后发邮件提交）、Tbl2asn（命令行程序自动创建序列记录提交完整基因组和大批量序列）等工具进行提交。

3.2 Dryad

Dryad 由美国国家进化分析中心等机构建立，其最初由进化生物学和生态学的主要期刊和科学团体提出，鼓励与数据一同提交手稿，进行存储。目前已有 1 1047个数据包，451 种期刊与之合作进行存储数据，包括生命科学领域目前主要的数据期刊如 *Genomic Data*、*BMC Research Notes*、*Open Health Data*、*GigaScience*、*F*1000 *Research*。

（1）所有数据都与出版期刊及其他数据仓储的数据进行关联，并与 TreeBASE 和 Knowledge Network for Biocomplexity 合作进行元数据互收割及数据检索，向

DataCite、Google scholar、Mendeley 等外部系统提供元数据检索,所有数据可免费下载。

(2) 与 DataONE 合作,为数据分配数字对象标识符(digital object identifier, DOI),便于引用与共享。除了处于保护期暂不公开的数据外,所有数据和元数据均可通过 CC0 协议进行复用。

(3) 不限制所提交的数据文件格式,鼓励提交 ASCII 和 HTML 数据格式,鼓励采用现有标准或进一步发展标准,提交过程简便。对论文发表前提交的数据允许短时间内禁止共享,在论文发表后数据同时发表,并与 CLOCKSS 合作对数据进行长期保存。

3.3 ArrayExpress

ArrayExpress 是由 EMBL-EBI、EC、NIH、NSF 于 2001 年联合建立的,用于存储功能基因组学数据,数据来源为科研人员直接提交或从 GEO 数据库中导入。高通量测序实验的实验描述和处理数据存储在 ArrayExpress 中,原始数据由 European Nucleotide Archive (ENA) 管理。目前,该数据库中共有 62 491 条实验数据。

(1) 每一条数据均匹配了收藏号,可按关键词及收藏号检索,按物种、技术、实验类型分类,并提供 ENA 原始数据链接,数据可直接下载,并可链接至分析软件如 Genomespace、Bioconductor 等进行可视化分析。

(2) 可接受所有芯片和测序技术产生的功能基因组学数据,微阵列数据提交需遵循基因芯片实验最小信息标准(minimum information about a microarray experiment,MIAME),测序数据提交需遵循高通量测序实验最小信息指南(Minimum Information about ahigh-throughput SeQuencing Experiment,MINSEQE),提交内容需包括元数据、原始数据文件、加工的数据文件,提供 Annotare 工具协助完成数据标准的要求来提交数据,所有数据在文章正式发表后才公开。

3.4 NCBI dbGaP

dbGaP 由 NIH 2006 年建立,用于存储和共享全基因组关联研究的编码基因型、表型相互作用的数据。2014 年 8 月,NIH 发布了“基因组数据共享政策”,dbGaP 是该政策指定的存储库[39]。

(1) dbGaP 中可获取的数据包括 3 种——授权的个人基因组数据、NIH 孤独症组学研究数据、精神分裂症遗传学研究数据,采用 Entrez 检索系统,可进行“结果关联检索”,提供“PheGenI”根据基因型及表型检索,可通过“Genome Browser”查看染色体测试位点分析结果。所有上传文件(包括研究方法、调查问卷及分析图表)均可开放获取[39],访问个体基因型和表型数据需要授权。

(2) 每一项数据被分配一个唯一的入藏号,面向公共领域可开放重用,引用时需至少包含入藏号,并向数据获取委员会提交获得数据使用认证,数据分析得到的结果在数据集禁止释放日期之前不能出版。

(3) 存储 NIH 资助的研究数据需要与项目办公室或基因组项目管理人员联系申请注册“研究”,邀请 PI 进入系统和提交模块,按照数据模板上传数据,经预览及批准后发布。数据模板中提供各类数据文件格式要求及指南。dbGaP 提供了 GRAF 工具用于查找 SNP 基因型数据相关的主题,TransEAV 工具用于将 EAV 数据转换成可提交的矩阵表格形式。

3.5 PURR

PURR 由普渡大学图书馆组织建立,面向普渡大学研究人员提供数据管理服务,支持数据集和软件工具的发布,并嵌入了数据管理计划服务,目前该仓储中在线发布了 329 个数据集,建立了 63 个项目空间、1 个在线工具[40]。

(1) 所有数据集在发布之后可在线下载。图书馆通过 PURR 完成数据对象的描述,管理元数据的采集,并在图书馆在线检索中建立索引,支持基本的数据集检索与浏览[41]。

(2) PURR 通过 CC 协议支持数据重用和引用,科研人员也可推荐其他许可协议。每一个数据集都分配唯一的 DOI,并提供具体引用格式,另外,与 DataCite 合作,为每一个数据集注册唯一的 DOI,并提供具体引用格式,便于数据发现及引用。

(3) PURR 支持普渡大学的研究人员发布项目研究数据,并可创建项目管理空间,提供“数据采集表”来帮助科研人员确定数据是否适合存储。PURR 可接受所有的文件类型,提交的数据集在两个工作日内,由存储专员和学科馆员审核、添加标签后进行发布。

(4) PURR 中嵌入了数据管理计划服务,帮助用户制定资助机构要求的数据管理计划,如帮助评估数据需求、组织、管理、共享数据,提供自我评估工具(DMP Self-Assessment Tool)以问卷的形式帮助用户了解数据管理计划中应包含的内容,并提供现成的样本文件,样本文件可直接放在申请书中。自 2011 年以来,普渡大学超过 1 000 个项目申请书中的数据管理计划使用了 PURR。另外,学科馆员提供在线参考咨询服务,服务内容涵盖制定数据管理计划、组织和管理数据、发现和使用研究数据。

3.6 BioSharing

BioSharing[42] 由英国牛津大学于 2007 年建立，主要集成了生命科学领域注册的数据和试验元数据的标准、数据仓储以及数据政策，并将三者进行关联。此外，BioSharing 还监测标准的开发及在数据仓储中的实现和应用，促进标准和数据库的协调一致，减少重复。目前该平台上所收集的内容包括标准 622 个、数据库 702 个、政策 23 个。

BioSharing 数据标准类型包括术语文件标准(Terminology artifact，345 篇)、模型格式(Model/format，192 篇)和报告指南(Reporting guideline，85 篇)3 种。在每一条记录下可查看与该记录相关的标准、数据仓储、政策、出版物，数据访问和重用条件、涉及的工具以及提供的支持帮助，并与机构、出版商和数据期刊合作，如 *Biomed central*、*F*1000 *Research*、*Scientific data*，帮助定义数据政策，集成相关标准和数据仓储。数据重用使用 CC 开放许可协议，数据存储需注册，可存储标准、数据仓储、数据相关政策，在存储数据后，BioSharing 团队进行完备性检查后即可发布。数据标准和数据仓储开发与维护人员可进行注册认证，使标准可被发现和重用。

4 生命科学领域数据管理服务启示

4.1 制定国家层面的战略规划和政策体系

数据政策是数据共享的基础，美英两国政府近年来密集发布开放数据相关的政策及战略规划来支持和推动数据共享，提升数据利用能力。生命科学领域一系列国际通用与高影响力数据仓储大部分由美英两国建立。我国虽然也发布了《科学数据共享工程建设规划》《科学数据共享条例》《国家科技计划项目科学数据汇交办法》和《科学数据分类分级共享及其发布策略》等政策[43]，但与美英两国的规划和政策体系相比，仍不够完善。未来还需加快出台国家层面的战略规划，在借鉴美英经验的同时，立足于我国生命科学数据的现状与需求，梳理与提炼出政策要素内容，针对专业领域或不同的数据类型制定相应的开放数据政策与数据安全政策，从国家层面构建完善的政策体系。

4.2 资助机构在政策制定上应发挥带头作用

国外基金资助机构如美国 NSF 和 NIH、英国生物技术和生物研究理事会 BBSRC、惠康基金会[44] 等纷纷制定了领域或机构内的数据政策。我国国家自然科学基金委员会 2014 年发布了受资助项目科研论文的开放获取政策[45]，但针对科学数据目前还没有具体的政策。国内的基金资助机构应发挥带头作用，引导国内科研机构对科学数据进行管理与共享，在政策的具体内容上可借鉴国外资助机构的模式如对数据提交内容、数据存储时间、数据公开及传播方式、数据使用规范、数据管理成效评估及经费支持等方面综合考虑。研究机构相关部门如图书馆可协助研究机构制定与实施相关政策，增强科研人员的数据管理意识与相关技能，并为其提供相关数据管理服务。

4.3 数据仓储在服务内容上应多元化

在生命科学领域，中国是数据产出大国，但建设的数据仓储无论是在数量上还是影响力上，都与数据产出不成正比。未来我国在数据仓储建设的服务内容上，可借鉴美英两国成熟的数据仓储建设经验，在数据获取方面，除建立快捷准确的检索系统和下载方式外，可通过将数据与出版期刊的文献进行链接、向外部系统如 DataCite 提供元数据检索、在图书馆检索系统中建立索引等方式提高数据被检索获取的可能性。在数据重用方面，数据仓储应为数据分配标识符或入藏号，对数据使用规范和引用格式作出说明，还可考虑将数据链接至外部分析软件，方便数据复用。在数据存储方面，应对数据提交内容、格式、涉密数据保护等作出说明，提供不同数据的标准，鼓励采用标准提交数据，可提供工具软件辅助完成数据提交，在数据质量层面可邀请领域专家参与，建立审核与发布及长期保存机制。

4.4 数据仓储中应嵌入数据管理相关服务

美英两国建设的数据仓储在服务模式上的创新，对于国内数据仓储开展服务具有很好的借鉴意义。我国一方面可向 PURR 学习，将数据仓储嵌入科研数据生命周期，提供数据管理计划服务，为科研人员提供在线的数据协作环境，助力其动态存储科研数据。另一方面，在数据仓储中提供数据管理相关服务如在线参考咨询服务、专业培训、工具软件使用指导等，辅助科研人员更好地利用数据仓储进行数据管理与共享。在此模式下，图书馆可发挥重要的作用，国外很多图书馆已开展数据管理服务，国内如中国科学院文献情报中心也展开了数据管理服务的相关研究[46]，并面向中国科学院大学生命学院、地球科学学院等科学数据高产出学院开展了科学数据管理课程教学实践。而国内图书馆则可进一步努力尝试科研数据管理与服务，其数据仓储平台可集成图书馆提供的数据管理服务，从而丰富服务模式层面的建设。

参考文献：

[1] 刘晶晶，马建华. 论科研数据开放共享的三种途径[J]. 情报杂

志,2015,34(10):146-150.

[2] NIELSEN H J,HJORLAND B. Curating research data: the potential roles of libraries and information professionals[J]. Journal of documentation,2014,70(2):221-240.

[3] MARX V. Biology: the big challenges of big data[J]. Nature,2013,498(7453):255-260.

[4] 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.

[5] 丁建华,彭政,王飞. 生物数据仓库研究及应用[J]. 计算机工程与应用,2005,41(12):192-194.

[6] MAY M. Life science technologies: big biological impacts from big data[J]. Science,2014,344(6189):1298-1300.

[7] United States Census Bureau[EB/OL]. [2015-12-20]. http://www.census.gov/.

[8] NCBI[EB/OL]. [2015-12-20]. http://www.ncbi.nlm.nih.gov/.

[9] ArrayExpress-functional genomics data[EB/OL]. [2015-12-20]. http://www.ebi.ac.uk/arrayexpress/.

[10] dbSNP[EB/OL]. [2015-12-20]. http://www.ncbi.nlm.nih.gov/projects/SNP/index.html.

[11] Global Biodiversity Information Facility[EB/OL]. [2015-12-20]. http://www.gbif.org/.

[12] Worldwide Protein Data Bank[EB/OL]. [2015-12-20]. http://www.wwpdb.org/.

[13] dbGaP[EB/OL]. [2015-12-20]. http://www.ncbi.nlm.nih.gov/gap.

[14] Sequence Read Archive[EB/OL]. [2015-12-20]. http://www.ncbi.nlm.nih.gov/sra.

[15] GenBank overview[EB/OL]. [2015-12-20]. http://www.ncbi.nlm.nih.gov/genbank/.

[16] Dryad[EB/OL]. [2015-12-20]. http://datadryad.org.

[17] Protein Data Bank in Europe[EB/OL]. [2015-12-20]. http://www.ebi.ac.uk/pdbe/.

[18] 朱伟民,朱云平,杨啸林. 生命科学信息工程设施以及在中国的实现[J]. 中国科学:生命科学,2013,43(1):80-88.

[19] 张勇进,王璟璇. 主要发达国家大数据政策比较研究[J]. 中国行政管理,2014(12):113-117.

[20] 唐源,吴丹. 国外医学科学数据共享政策调查及对我国的启示[J]. 图书情报工作,2015,59(18):6-13.

[21] 张涵,王忠. 国外政府开放数据的比较研究[J]. 情报杂志,2015,34(8):142-146.

[22] 张瑶,顾立平,杨云秀,等. 国外科研资助机构数据政策的调研与分析——以英美研究理事会为例[J]. 图书情报工作,2015,59(6):53-60.

[23] ELIXIR Programme 2014-2018[EB/OL]. [2015-12-25]. http://www.elixir-europe.org/about/elixir-programme-2014-2018.

[24] 司莉,邢文明. 国外科学数据管理与共享政策调查及对我国的启示[J]. 情报资料工作,2013(1):61-66.

[25] 欧盟 GRDI2020 发布《全球科学数据基础设施:重大数据挑战》报告[EB/OL]. [2015-12-25]. http://www.gisti-think-bank.ac.cn/admin/upload/20111101-20110311.pdf.

[26] 司莉,封洁. 科学数据的保存与维护:国际组织的动向[J]. 图书馆,2015(4):6-10.

[27] HHS public access policy for research data[EB/OL]. [2015-12-25]. http://www.hhs.gov/open/plan/hhs-public-access-policy-for-research-data.html.

[28] Digital strategy at USDA[EB/OL]. [2015-12-25]. http://www.usda.gov/wps/portal/usda/usdahome?navid=DIGITAL-STRATEGY.

[29] United States Department of Agriculture open data policy communications plan[EB/OL]. [2015-12-25]. http://www.usda.gov/documents/odp-communications-plan.pdf.

[30] Policy on data management and sharing[EB/OL]. [2015-12-25]. http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm.

[31] 刘细文,熊瑞. 国外科学数据开放获取政策特点分析[J]. 情报理论与实践,2009(9):5-9.

[32] 司莉,李月婷. 国外科学数据保存政策的分析与启示[J]. 信息资源管理学报,2014(2):45-50.

[33] PANGAEA[EB/OL]. [2016-01-06]. http://www.pangaea.de.

[34] Figshare[EB/OL]. [2016-01-06]. http://figshare.com/.

[35] European Genome-phenome Archive[EB/OL]. [2016-01-06]. https://www.ebi.ac.uk/ega/.

[36] Combase[EB/OL]. [2016-01-06]. http://www.combase.cc/index.php/en/.

[37] United States Transuranium & Uranium Registries[EB/OL]. [2016-01-08]. http://www.ustur.wsu.edu/.

[38] NIH genomic data sharing policy[EB/OL]. [2016-01-08]. https://gds.nih.gov/03policy2.html.

[39] TRYKA K A, HAO L, STURCKE A, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP[J]. Nucleic acids research, 2014,42:975-979.

[40] Purdue University Research Repository[EB/OL]. [2016-01-10]. https://purr.purdue.edu/.

[41] 王辉, WITT M, 窦天芳. 普渡大学研究仓储及其支持的科学数据管理服务[J]. 现代图书情报技术,2015,31(01):9-16.

[42] BioSharing[EB/OL]. [2016-01-10]. https://biosharing.org/.

[43] 周文能. 科学数据共享工程"十一五"建设规划[J]. 太原科技,2006(6):1-3.

[44] Funders' data policies[EB/OL]. [2016-01-10]. http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies.

[45] 国家自然科学基金委员会关于受资助项目科研论文实行开放获取的政策声明[EB/OL]. [2016-01-10]. http://www.nsfc.gov.cn/publish/portal0/tab38/info44471.htm.

[46] 盖晓良,张闪闪,顾立平. 国际信息服务机构的数据管理服务政策分析[EB/OL]. [2015-12-02]. http://ir.las.ac.cn/handle/12502/7817.

作者贡献说明:
邹丽雪: 撰写论文主体内容,收集、整理数据并进行统计分析,修订论文;
欧阳峥峥: 设计研究方案,提供信息源,提出论文修改意见,修订论文终稿;
王辉: 设计研究方案,修订论文;
吴鸣: 提出论文修改意见,参与论文修订。

Research on the Characteristics and Services of Research Data Repositories in Life Science

Zou Lixue Ouyang Zhengzheng Wang Hui Wu Ming
National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract: **[Purpose/significance]** This paper investigates and analyzes the research data repositories in life science and discusses the research data service in this field. **[Method/process]** Using re3data. org, we analyzed the status of research data repositories in life science in terms of set - up year, country, institution, academic field, openness. Besides, we took Genbank, Dryad, ArrayExpress, Purdue University Research Repository, Biosharing and dbGaP as six typical cases to demonstrate the services. **[Result/conclusion]** America and United Kingdom lead the construction of data repositories in life science, owing to the good policies of governments and funders. China should learn the abundant experience and establish the strategic plan and policy system. Meanwhile, the funders should play the role of guidance to promote the data management and sharing, build the specific and high - impact data repositories in life science and provide services.

Keywords: research data repository life science research data management research data service

(上接第 58 页)

作者贡献说明:
高冉: 论文框架设计,论文第一、二、四部分主笔,课程前 6 课时教学,学生作业参考文献评价,统稿;
张波涛: 论文第二、三部分主笔,课程后 4 课时教学,学生作业内容质量评价;
茹海涛: 课程框架及内容指导,论文最终审校。

Practice and Evaluation of Embedding Information Literacy Education in Environment Science Special English Class by the Theory of Threshold Concept

Gao Ran[1] Zhang Bo-tao[2] Ru Haitao[1]
[1] Library, Beijing Normal University, Beijing 100875
[2] College of Water Sciences, Beijing Normal University, Beijing 100875

Abstract: **[Purpose/significance]** This paper explores the practical teaching methodthat information literacy education is embedded in the environment science special English class by the theory of threshold concept. The course effect is evaluated by investigating the class assignment and the following master dissertation proposal. **[Method/process]** This class includes the research information resources introduction, literatures searching method and results analysis, references managing skills, research paper writing skills, and the processes introduction from submission to publication. The master dissertation proposals of attending class graduate students are compared with the control group to evaluate the course. The correlation between the class assignment and the following master dissertation proposals is also analyzed. **[Result/conclusion]** The results show that: the embedding class has positive significance for improving students' research capabilities in different aspects, such as research degree, writing regulation, reference selective periodical and citation criteria; postgraduate coursework and specification degree of thesis proposal document reference has a good linear relationship. The embedded course can play a positive role in improving students' scientific research ability.

Keywords: embedded education curriculum information literacy environment science threshold concepts