基于re3data的中英科学数据仓储平台对比研究*

袁烨[1]　陈媛媛[2]
（1. 马来亚大学，吉隆坡 50088；2. 黑龙江大学信息管理学院，哈尔滨 150080）

摘要：以re3data为数据获取源，选取中英两国406个科学数据仓储为研究对象，从分布特征、责任类型、仓储许可、技术标准及质量标准等5个方面、11个指标对两国科学数据仓储的建设情况进行对比分析，试图为我国数据仓储的可持续发展提出建议：广泛联结国内外异质机构，推进多学科领域的交流与合作，有效扩充仓储许可权限与类型，优化技术标准的应用现况，提高元数据使用的灵活性。

关键词：科学数据；数据仓储平台；re3data；中国；英国
中图分类号：G250　DOI：10.3772/j.issn.1673-2286.2024.02.002
引文格式：袁烨，陈媛媛. 基于re3data的中英科学数据仓储平台对比研究[J]. 数字图书馆论坛，2024，20（2）：13-23.

在数据密集型研究范式下，科学数据已成为国家科技创新和发展的基础性战略资源，也是大数据时代最基本、最活跃的科技资源。随着FAIR数据原则的发布，对数据的开放性、可用性和可持续性的关注不断增强，对数据管理的紧迫需求突显。学术机构、资助单位等开始通过政策、倡议、网络和基础设施来应对数据管理需求的挑战。例如：美国国家科学基金会（National Science Foundation，NSF）在2011年发布了数据共享政策，要求资助获得者在数据管理计划中提供关于数据处理的信息[1]；我国国务院于2018年印发《科学数据管理办法》，明确了公共资金资助科学数据管理的职责、原则、方式和机制[2]。不过，由于科学数据来源广泛、数量庞大、种类繁多，科研人员正面临严峻的挑战，即如何有效地获取和处理这些数据。解决这一难题的关键在于建立一个集数据存储、描述、共享和获取功能于一体的科学数据仓储。因而涌现出不同类型、各具特色、质量参差不齐的科学数据仓储。然而，受到信息技术、政策支持、学术环境与研究氛围等因素的影响，科学数据仓储的发展在发达国家与发展中国家之间呈现出明显的差异。本文通过对中英两国科学数据仓储平台的对比研究，深入了解不同文化背景下的数据管理实践，为我国科学数据管理平台的建设与优化提供借鉴与启示。

1　研究背景

尽管我国已开始积极发展科学数据仓储，但相对于发达国家仍存在差距。在仓储建设方面，国内已经有学者展开相关研究。①国家或地区视角的调查研究。例如：张莎莎等[3]对英国数据仓储运行状况进行了调查，夏姚璜[4]则进行了中美科学数据仓储建设对比研究。②学科视角的数据管理平台分布研究。例如，王丹丹等[5]分析了全球社会科学数据管理平台建设概况，吴思竹等[6]结合可视化图表对医学科学数据仓储的分布、建设等情况进行分析。③仓储元数据项视角的综合分析。黄国彬等[7]从服务内容构成与服务实现模式两个方面对仓储元数据创建服务进行了调研。然而，尽管学者已经开展对中美科学数据仓储建设的对比研究，但研究成果

收稿日期：2023-11-15
*本研究得到国家社会科学基金项目“高校图书馆科研数据服务模式与服务系统研究”（编号：17CTQ041）资助。

无法全面体现中国的科学数据仓储建设与英美等发达国家的差距。因此，有必要考虑与更多国家进行对比，从更多维度更深入地理解和评估中国科学数据仓储建设的缺口，进而为科学数据管理和仓储建设提供更科学、系统的建议。

英国是早期推动开放获取的国家之一。南安普顿大学（University of Southampton）的ePrints仓储等标志着该领域的早期实践，也为全球研究成果的开放获取开创了先河[8]。英国研究基金会在2005年颁布了一项具有重大影响的政策，要求其所资助的研究项目成果在出版后的6个月内以开放获取的形式对公众开放[9]。该政策的实施进一步强化了英国在全球开放获取实践方面的引领地位。本文基于英国推动国际科研合作和开放科学实践、颁布具有示范性的科学数据管理政策与法规，以及建设全球知名的高水平科学数据仓储的经验，结合中英两国科学研究基础设施、法律许可等方面的差异，深入比较两国科学数据仓储的分布特征、责任属性、法律许可、技术标准和质量标准等，旨在为我国科学数据仓储的建设实践提供有益的参考。

2 数据获取及研究方法

re3data是一个旨在提供全球科学数据仓储的综合性注册平台。该平台由德国地球科学研究中心、卡尔斯鲁厄工业大学图书馆、普渡大学图书馆以及柏林洪堡大学图书与信息科学学院等合作伙伴共同创建[10]，涵盖来自不同学科领域的数据仓储，并为研究人员、资助机构、出版商和学术机构提供适当的存储库，用于永久性存储和访问数据集[11]。re3data提供了详细的数据仓储信息，包括数据仓储的名称、描述、所属学科、数据类型、访问方式、使用协议等，这些信息为对比研究提供了丰富的数据支持和参考依据，有助于深入分析和比较中英两国的科学数据仓储建设情况。re3data还制定了一系列的国际标准和指南，用于评估和认证科学数据仓储的质量和可信度，这些标准和指南可以成为评估中英两国科学数据仓储建设情况的参考依据，有助于提升对比研究的科学性和客观性。截至2023年10月30日，re3data索引3 160个科学数据仓储，共涉及88个国家或地区。美国以1 172个平台位居第一，其次是德国（504个）、加拿大（394个）、英国（321个），而中国以85个平台位居第9。在re3data项目启动时，只有少数科学数据仓储列表存在，仅列出基本信息，如仓储名称、责任机构和学科类型。2012年7月，1.0版本的科学数据仓储描述词汇文档发布。为确保词汇的开发透明度，该项目不仅支持用户在网站上对词汇进行评论，还发送电子邮件征询意见，并在2012年12月发布了2.0版本[12]。该词汇文档涵盖了以下方面：仓储分布，如仓储类型、内容类型、学科分布等一般信息；仓储责任，如机构责任类型、责任主体类型等；仓储许可，如数据许可、数据库许可等；仓储技术标准，如应用程序编程接口（Application Programming Interface，API）、软件使用等；仓储质量标准，涉及科学数据仓储认证、元数据标准等。根据描述科学数据仓储的上述5类词汇和11个分析指标，对406个科学数据仓储（英国321个，中国85个）的元数据项数据进行定量分析。其中：通过re3data的“country”浏览仓储类型、内容类型等9个分析指标，并根据分析指标进行二次检索；结合“country”与“subject”获取学科分布，并依据re3data控词进行统计；通过仓储URL熟悉责任机构并对责任主体类型进行归类，统计责任机构词频，得到主要责任主体。指标类别与数据来源映射关系如表1所示。

表1 指标类别与数据来源映射关系

指标类别	分析指标	re3data元数据	数据来源
仓储分布	仓储类型	Repository Type	re3data控词
	内容类型	Content Type	re3data控词
	学科分布	Subjects	re3data控词
仓储责任	机构责任类型	Institution Responsibility Type	re3data控词
	责任主体类型		URL
仓储许可	数据许可	Data Licenses	re3data控词
	数据库许可	Database Licenses	re3data控词
仓储技术标准	API	API	re3data控词
	软件使用	Software	re3data控词
仓储质量标准	仓储认证	Certificates	re3data控词
	元数据标准	Metadata Standards	re3data控词

3　中英科学数据仓储平台对比分析

3.1　仓储分布

（1）仓储类型。如表2所示，两国的主要数据仓储分为6种不同类型，普遍呈现出“学科性”和“机构性”两大主要特点，其中“学科性”仓储占主导地位。在英国的数据仓储中，学科类型的数据仓储占比高达62.3%，而机构类型仓储占比为22.7%，7.5%的数据仓储同时具备“学科性”和“机构性”。相较之下，中国的数据仓储中“学科性”特征更为显著，学科类型仓储占比高达83.5%，而机构类型仓储仅占5.9%。另外，4.7%的中国数据仓储既包含“学科性”又包含“机构性”。需指出的是，那些不具备明显学科和机构属性的仓储通常被用于存储一般性数据，以满足数据存储需求。此外，综合性数据仓储（学科与其他、机构与其他）还扮演着另一个重要角色，即存档相关的分析或实验控制数据，以补充特定类型仓储中的原始数据。

（2）内容类型。re3data把全球数据仓储的内容类型分为15种，中英两国科学数据仓储内容类型如图1所示。两国科学数据在所有主要内容类型上都有分布。英国数据仓储中，标准办公文件数量最多，共有200个，占比高达12.9%，其次是科学和统计数据格式、图像、原始数据以及纯文本等。中国数据仓储中，标准办公文件和结构化文本数量并列第一，各有31个，占比为8.9%，其次是结构化图形、源代码、科学和统计数据格式等。两国的数据内容类型数量都明显多于其仓储总量，这表明每个科学数据仓储通常包含多种不同类型的数据。在英国，仅有8个平台（2.5%）发布单一类型的内容，例如Diamond Data Catalog只发布存档数据、UK Reading Experience Database只发布纯文本数据，而其余的313个平台（97.5%）发布多种类型的数据。在中国的85个平台中只有National Wild Seed Resource Center发布单一类型的结构化文本数据，还有一个平台未提供明确说明，

表2　中英两国科学数据仓储类型

仓储类型	英 国		中 国	
	数量/个	占比/%	数量/个	占比/%
学科类型	200	62.3	71	83.5
机构类型	73	22.7	5	5.9
学科与机构	24	7.5	4	4.7
其他	16	5.0	4	4.7
学科与其他	7	2.2	1	1.2
机构与其他	1	0.3	0	0.0
总计	321	100.0	85	100.0

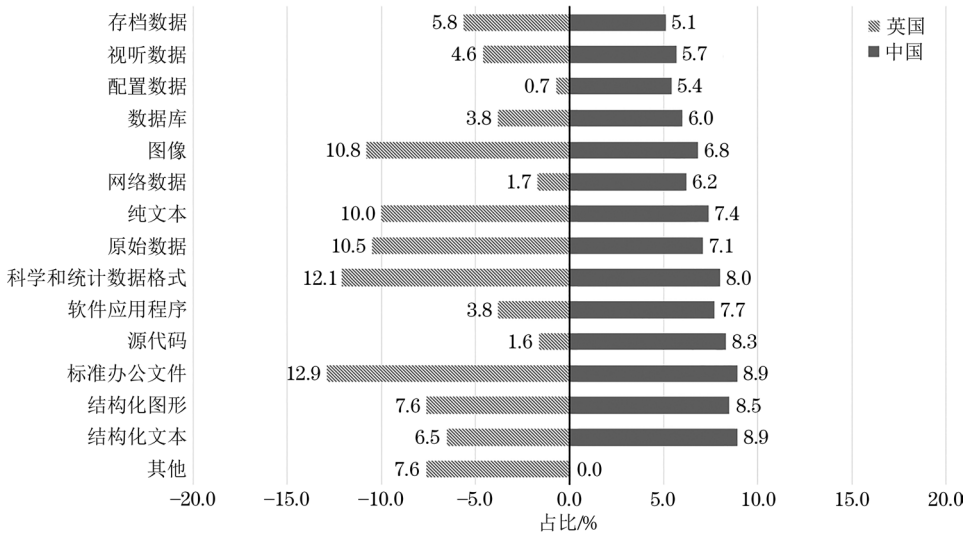

图1　中英两国科学数据仓储内容类型

其余平台都发布多种不同类型的数据。总体而言，中国仓储的内容类型相对于英国更加均衡。

（3）学科分布。按照德国研究基金会制定的学科分类框架，re3data收录的科学数据仓储可划分至4个主要学科大类[13]，涵盖了14个一级学科和45个二级学科。中英两国科学数据仓储的学科分布如表3所示。就学科

表3　中英两国科学数据仓储的学科分布

学科大类	一级学科	二级学科	英国		中国	
			数量/个	占比/%	数量/个	占比/%
人文和社会科学	人文科学	古代文化	10	12.5	0	0.6
		历史	20		0	
		美术、音乐、戏剧和媒体研究	13		1	
		语言学	12		0	
		文学研究	8		0	
		非欧洲语言和文化、社会和文化人类学、犹太研究和宗教研究	4		0	
		神学	5		0	
		哲学	4		0	
	社会与行为科学	教育科学	8	10.3	0	2.5
		心理学	8		0	
		社会科学	26		2	
		经济学	16		2	
		法理学	5		0	
生命科学	生物学	基础生物和医学研究	101	25.3	28	41.2
		植物科学	22		16	
		动物学	32		21	
	医学	微生物学、病毒学和免疫学	37	18.9	14	16.5
		医学	65		9	
		神经科学	13		3	
	农业、林业、园艺和兽医学	农业、林业、园艺和兽医学	19	3.1	7	4.4
自然科学	化学	分子化学	7	3.7	0	0.6
		化学固体和表面研究	2		0	
		物理和理论化学	3		0	
		分析化学	5		0	
		生物化学和食品化学	5		1	
		聚合物研究	1		0	
	物理学	凝聚态物理学	3	4.4	0	3.8
		离子体物理学	7		1	
		粒子、核与场	6		2	
		天体物理学和天文学	11		3	
	数学	数学	9	1.5	0	0.0
	地球科学	大气科学和海洋学	30	14.8	11	27.9
		地质学和古生物学	5		7	
		地球物理学和大地测量学	15		13	
		地球化学、矿物学和晶体学	9		5	
		地理学	18		4	
		水研究	13		4	
工程科学	机械与工业工程	机械与工业工程	2	0.3	0	0.0
	热能工程/过程工程	工艺工程和技术化学	2	0.5	0	0.6
		热能技术、热机和流体力学	1		1	
	材料科学与工程	材料工程学	1	0.7	0	0.6
		材料科学	3		1	
	计算机科学、电子和系统工程	系统工程	3	3.3	0	1.3
		电气工程	2		0	
		计算机科学	15		2	
	建筑工程与建筑学		4	0.7	0	0.0

大类而言，生命科学和自然科学仓储在数量方面居于领先地位，其次是人文和社会科学、工程科学。在生命科学领域，生物学是最主要的学科，侧重于基础生物和医学研究。值得强调的是，中国的生物学仓储占比高达41.2%，远超英国。自然科学领域中，地球科学仓储最多，英国占比为14.8%，而中国占比为27.9%。中国仅有1个化学仓储，且没有数学领域仓储。在人文科学领域，英国仓储覆盖了13个学科，共有139个，占比为22.8%；相比之下，中国仓储的学科分布较为零散，涵盖的学科相对较少，相关仓储占比仅为3.1%。至于工程科学领域，两国仓储相对较少，英国占比为5.5%，计算机科学、电子和系统工程仓储数量较多。在中国，工程科学领域仓储占比为2.5%，机械与工业工程及建筑工程与建筑学领域尚未有相关仓储。综上，中英两国的数据仓储主要集中在生物学、医学和地球科学领域。英国的仓储涵盖了各个学科，而中国仓储还未覆盖23个二级学科，尤其在人文科学领域差距显著。

3.2　仓储责任

（1）机构责任类型。re3data注册表的元数据项提供了机构类型（非营利性、商业性等）以及科学数据仓储责任信息[14]。中英两国均以非营利性仓储为主，占比分别为中国95.4%和英国90.8%。机构责任类型与此一致，具体来说：英国的科学数据仓储中，受公共资金资助的占比28.18%，一般性的占比41.15%，受商业赞助的占比1.44%，技术型的占比29.23%；中国仓储的责任类型分布有相似性，其中公共资金资助仓储占比（23.76%）略低于英国，一般性仓储占比（45.86%）稍高，商业赞助仓储占比为0.55%，技术型仓储占比为29.83%。这些数据表明，科学数据仓储在不同国家可能会承担不同类型的责任，这对跨国合作和数据共享的相关政策和实践具有重要的参考价值。

（2）责任主体类型。根据仓储URL中的信息将责任主体分为政府及政府性质委员会、研究所、研究理事会、私人公司、数据中心、基金会、高校、国际性机构8类（见图2）。英国的科学数据仓储中，占比最高的责任主体类型是国际性机构，占比高达54.4%，其次是高校，占比为18.2%。在中国的科学数据仓储中，占比最高的责任主体类型是研究所，占比为27.5%，其次是国际性机构，占比为22.7%。

通过对具体机构名称进行词频统计，可以确定在科学数据仓储建设中作出突出贡献的具体机构。中英两国科学数据仓储主要责任主体如表4所示。在英国，牛津大学、伦敦大学学院、剑桥大学、爱丁堡大学及曼彻斯特大学等高校是科学数据仓储建设的主要责任主体。惠康信托基金会在资金方面提供支持，并协助建设。各重要学科如自然环境、医学、生物技术和生物科学研究理事会等积极参与建设。联合信息系统委员会（Joint Information Systems Committee，JISC）通过提供数字资源和网络技术设施等，协助高校应对数字化挑战。在中国，中国科学院及其所属研究所作为责任主体占比为23.9%，是科学数据仓储建设的核心机构。中华人民共和国科学技术部和国家科技基础条件平台中心是提供关键资金和技术支持的单位，其他机构的贡献较为有限。

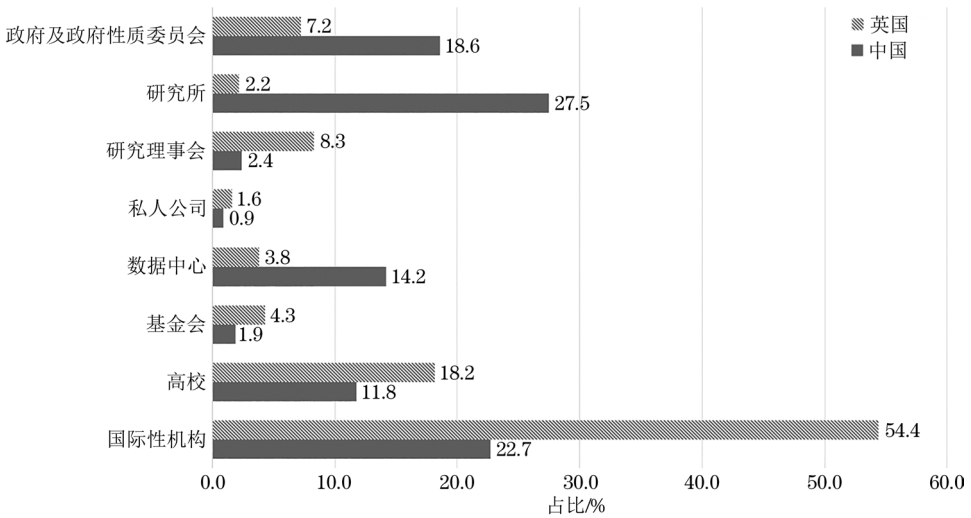

图2　中英两国科学数据仓储的责任主体类型

表4　中英两国科学数据仓储的主要责任主体

英国仓储			中国仓储		
责任主体	数量/个	占比/%	责任主体	数量/个	占比/%
惠康信托基金会	31	4.4	中国科学院	22	23.9
牛津大学	19	4.2	中华人民共和国科学技术部	6	6.5
伦敦大学学院	14	3.1	国家科技基础条件平台中心	3	3.3
剑桥大学	13	2.8	中国科学技术大学	3	3.3
自然环境研究理事会	12	2.6	北京大学	2	2.2
医学研究理事会	12	2.6	国家基因组科学数据中心	2	2.2
生物技术与生物科学研究理事会	12	2.6	国家自然科学基金委员会	2	2.2
爱丁堡大学	12	2.6	华中科技大学	2	2.2
曼彻斯特大学	10	2.4	南京大学	2	2.2
联合信息系统委员会	10	2.4	天津大学	2	2.2

3.3　仓储许可

作为数据管理的关键工具，数据许可和数据库许可在焦点和范围上存在略微差异[15]。数据许可涉及数据的本质，包括数据的内容、结构以及使用权限；而数据库许可更关注整个系统，包括数据库管理系统（Database Management System，DBMS）许可、数据库服务器使用许可、数据库集群许可等。中英两国科学数据仓储的许可类型与数量如表5所示。英国的数据许可共500个，中国的数据许可共99个，表明每个仓储同时拥有多种类型的数据许可。两国的主要数据许可类型都为CC和Copyrights。CC基于开放内容许可协议，为版权保护和知识共享提供了一种灵活的方式。有数据库许可的英国仓储总计78个，占比为24.3%；中国有14个，占比为16.5%。同样的，CC和Copyrights也是两国数据库许可的主要类别。综合比较来看，两国的数据许可分布比数据库许可更广泛。在许可类型中，CC和Copyrights是两国仓储许可中最常见的类型。

3.4　仓储技术标准

中英两国科学数据仓储的技术标准包括API和软件使用两个方面，具体数据如表6所示。

（1）API。科学数据仓储中的API基于标准的通信协议和数据格式，允许不同软件或系统有效地共享和交换数据[16]。英国仓储的API覆盖率达到70.7%，而中国仅为27.1%。两国的科学数据仓储提供了7种类型的API：文件传输协议（FTP）、网络通用数据格式（NetCDF）、元数据采集协议（OAI-PMH）、表征状态转移（REST）、简单对象访问协议（SOAP）、SPARQL

表5　中英两国科学数据仓储的许可类型与数量

数据许可			数据库许可		
类 型	英国数量/个	中国数量/个	类 型	英国数量/个	中国数量/个
Apache License 2.0	12	0	Apache License 2.0	4	2
BSD	4	0	BSD	2	0
CC	148	23	CC	24	3
CC0	21	4	CC0	5	2
Copyrights	110	48	Copyrights	13	3
ODC	19	2	ODC	2	0
OGL	30	0	Public Domain	1	0
Public Domain	34	4	其他	27	4
其他	122	18			

表6　中英两国科学数据仓储的技术标准

API			软件使用		
类 型	英国仓储数量/个	中国仓储数量/个	类 型	英国仓储数量/个	中国仓储数量/个
FTP	59	11	CKAN	13	0
NetCDF	8	0	DSpace	13	0
OAI-PMH	36	2	Dataverse	2	4
REST	74	8	EPrints	23	0
SOAP	8	0	Fedora	3	0
SPARQL	7	0	MySQL	17	6
其他	35	2	dLibra	1	0
			其他	67	5
			unknown	125	23

（SPARQL Protocol and RDF Query Language）、其他。FTP和REST是两国仓储的主要API类型：FTP适用于大规模数据传输，但不提供加密服务[17]；而REST由于传输效率高、简单实用，在互联网和分布式系统中得到广泛应用[18]。此外，OAI-PMH是一种开放和协作式的标准，适用于数字档案馆和图书馆等领域，在英国仓储中也有一定比例的应用。相比之下，我国仓储的API覆盖率明显低于英国，同时API类型的多样性也有待提高。

（2）软件使用。re3data的元数据项包括13种软件，中英两国的科学数据仓储共涉及8种不同的软件。这些软件的覆盖范围都相对有限。英国仓储的软件使用覆盖率为43.3%，近40%的软件状态为“unknown”，表示相关信息未知或未提供。中国仓储的软件使用覆盖率为17.6%，有27.1%的软件状态为“unknown”。英国仓储广泛使用的软件包括EPrints、MySQL、CKAN和DSpace。EPrints广泛应用于高校、研究机构和图书馆，用于创建和管理数字存储库；MySQL是一种开源的关系型数据库管理系统；CKAN大多被政府机构和组织采用，用于创建开放数据门户[19]；DSpace通常被高校图书馆、研究机构和文化遗产机构采用，用于建立数字存储库，以促进学术研究成果的传播、文化遗产的保存和数字内容的管理[20]。相比之下，中国仓储使用的软件种类较少，只有少数仓储使用MySQL和Dataverse。Dataverse是开源的研究数据管理平台[21]，可以帮助研究机构提高研究数据管理的效率和可信度。相较而言，英国仓储在软件使用方面提供了更广泛的选择，而中国仓储使用的软件相对有限。未来，我国仓储应考虑采用不同类型软件，以提高科学数据管理的效率和质量。

3.5　仓储质量标准

（1）仓储认证。评估认证是确保科学数据仓储质量的有效手段。虽然已经出现多种认证标准和机制，如可信数据知识库审查与认证（TRAC）、数据认可印章（DSA）、世界数据系统（WDS）和CoreTrustSeal，但就re3data元数据项中的“Certificates”而言，数据仓储的认证覆盖情况并不理想。全球范围内总共有3 160个数据仓储，仅有311个获得了认证，认证覆盖率仅为9.8%。英国仓储的认证覆盖率为8.1%，共有17个仓储通过CoreTrustSeal认证。中国仓储的认证覆盖率为15.3%，共有6个仓储通过CoreTrustSeal认证。综上所述，CoreTrustSeal作为国际公认的认证标准，在中英两国都得到了广泛的认可。目前，中国科学数据仓储的认证率高于全球平均水平，这在一定程度上反映了数据平台建设的质量和可信度。

（2）元数据标准。中英两国科学数据仓储的元数据标准如表7所示。英国共有143个数据仓储（44.5%）明确说明了采用的元数据标准（16种），中国共有40个数据仓储（47.1%）明确说明了所使用的元数据标准（11种）。英国仓储使用较多的元数据标准包括Dublin Core、DataCite Metadata Schema、ISO 19115和DDI。ISO 19115是用于描述地理信息和地理数据的元数据标准[22]，而DDI则是专注于社会科学领域的元数据标准[23]。中国仓储使用较多的元数据标准包括Repository-Developed Metadata Schemas、DataCite Metadata Schema和Dublin Core。Repository-

表7　中英两国科学数据仓储的元数据标准

元数据标准名称	元数据标准定义	英国仓储数量/个	中国仓储数量/个
CF Metadata Conventions	Climate and Forecast Metadata Conventions	3	0
CIF	Crystallographic Information Framework	3	0
DCAT	Data Catalog Vocabulary	2	0
DDI	Data Documentation Initiative	11	3
DIF	Directory Interchange Format	1	1
Darwin Core		2	1
DataCite Metadata Schema		36	8
Dublin Core		54	7
FGDC/CSDGM	Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata	3	2
Genome Metadata		0	1
ISA-Tab	Investigation/Study/Assay Tab-Delimited Format	1	3
ISO 19115		16	4
MIDAS-Heritage	Monument Inventory Data Standard-Heritage	1	0
OAI-OR	Open Archives Initiative Object Reuse and Exchange	3	0
QuDEx	Qualitative Data Exchange Format	1	0
RDF Data Cube Vocabulary		3	0
Repository-Developed Metadata Schemas		0	9
SPASE Data Model	Space Physics Archive Search and Extract Data Model	0	1
其他		3	0

Developed Metadata Schemas指的是平台自行开发的元数据模式或元数据框架，用于描述信息的规则和结构[24]。总体而言，中国仓储的元数据标准覆盖率虽然高于英国，但仍未达到50%。中国仓储使用的元数据标准类型相对较为集中，而英国仓储使用的元数据标准更为丰富。Dublin Core因其简单性、互操作性和扩展性等特点成为最广泛使用的标准[25]，其次是DataCite Metadata Schema。

4　启示与建议

4.1　广泛联结国内外异质机构

我国的科学数据仓储以自主建设为主，国际合作开发率相对较低。同时，这些仓储的开发主要依赖中国科学院及其附属研究所，其他机构的参与度有限。然而，这些科学数据仓储通常采用非营利的公益性服务模式，为了确保长期运营与发展，必须稳定、持续地进行资金、技术和人力资源投入与保障[26]。因此，有必要积极开展与国内外数据组织的合作与交流，以维持其可持续发展。首先，鼓励科研机构与企业建立战略合作关系，促进科研成果在产业应用中的转化。可通过数据管理工具的研发、数据共享平台的开发、数据驱动的业务模型的建立等充分发挥双方优势，推动科研成果更快速、有效地转化为创新产品和服务。其次，关注并遵循国际数据管理领域的标准和最佳实践，参与并推动全球性数据仓储与共享网络的建设，如国际地球观测项目（International Geospatial Initiative）或国际脑计划（International Brain Initiative）等，促进我国数据管理技术与国际标准接轨，提高我国数据仓储的国际影响力。

4.2　积极推进学科领域的交流与合作

我国的数据仓储主要集中在生物学、地球科学和医学领域，尚未覆盖其他23个细分学科。为实现科学数据仓储的生态化发展，确保数据来源的全面性与覆盖学科的均衡性，可采取以下举措：①设立专项基金，

支持高校成立人文社科平台或数据中心，负责协调和推动人文社科数据的采集、整合和共享；②与相关学科专业协会共同制定人文社科领域的数据标准，以确保数据的互操作性和可持续性；③发挥生命科学与自然科学领域数据优势，推动跨学科研究项目的合作，或者实施数据集成项目等，以消除学科壁垒，促进数据共享和跨学科研究。这些措施将有助于科学数据仓储的生态化发展，保障我国仓储数据的全面性与数据来源的可持续性。

4.3　有效扩充仓储许可权限与类型

数据库许可管理数据库系统的结构和性能，是确保数据合法、安全和有效利用的关键工具。然而，当前我国的数据库许可覆盖范围较窄，类型也较单一。为此提出以下建议：①制定多元化数据库许可策略，以满足不同用户群体的需求，这些策略可以包括开放数据库、学术研究数据库、教育数据库、商业合作数据库等；②通过不同类型的许可，适应不同用户的需求，提高科学数据仓储的可用性；③建立数据库许可门户网站，提供详细的数据库许可政策和可用数据库的信息，以此增强用户对许可政策的了解，提高数据仓储的可访问性；④倡导并推动开放数据库共享倡议实施，通过与学术期刊、科研机构等合作，鼓励用户将其数据分享至开放数据库，促进知识开放共享和数据库多样性提升。

4.4　持续优化技术标准的应用现况

目前，我国的科学数据仓储的技术标准应用情况并不乐观，API类型有限，仅有17.6%的仓储支持软件使用。然而，API和软件使用对数据仓储开放起到了推动和支持的作用，涉及平台的访问与共享、数据整合以及用户体验等多个关键方面。因此，亟需改善技术标准应用现状以提升数据利用率、释放数据价值。首要措施是积极拓展多样化的API，以满足各类用户及应用程序的不同需要。这包括支持多种API类型，如NetCDF、SOAP、SPARQL等，从而确保数据平台的互操作性，推动数据的广泛应用和共享。此外，为提高软件的使用率，需要开发软件集成工具，提供示例代码和教程，并开展API及软件培训。定期的软件更新和维护也至关重要，应确保支持的软件与数据仓储兼容，从而提高软件的性能和稳定性。

4.5　提高元数据使用的灵活性

当前，我国仓储使用的元数据标准类型相对单一，缺乏丰富性。在保持标准集中性的同时，确保应用的灵活性尤为重要。针对该问题，提出以下措施。①建立创新元数据标准管理体系。采用现代技术和方法，如智能合同或分布式账本技术，实现元数据标准的实时更新和扩展，以确保其与不断演进的数据环境保持同步。②构建开放式的元数据标准平台。鼓励各部门和利益相关者积极参与标准的制定和更新，提供开发工具和资源，支持自定义标准的创建和管理。③引入智能辅助标准协商工具。协助各个部门协商和更新元数据标准，提供分析和建议，以促进快速的标准制定和调整。通过这些措施，有效提高元数据使用的灵活性，提升数据管理效率，鼓励更广泛的数据共享和重复使用。

5　结语

基于我国与英美数据仓储发展的差距，本文从分布特征、责任类型、仓储许可、技术标准及质量标准等方面对中英两国406个仓储进行调查与对比分析，并提出我国仓储建设的可行举措，期望给予我国数据开放获取有价值的借鉴与参考。然而，本文仍存在一些不足：仅对re3data现有指标进行调查，缺乏对引申指标的研究。下一步应结合分析结果与经验，以提升科学数据仓储的影响力为切入点，引入知识转移与技术转化、文化与教育价值、社会参与度与媒体关注度等指标，全方位、多维度地开发科学数据仓储影响力的评估模型，促使各界更准确地评估和理解仓储在社会、文化、教育等方面的实际影响，从而以更高的标准和更广阔的视野推动我国科学数据仓储的进步与发展。

参考文献

[1] ARCHANA S N, PADMAKUMAR P K. The status quo of Indian data repositories indexed in re3data registry[J]. Digital Library Perspectives, 2023, 39 (4): 496-516.

[2] 中国政府网. 国务院办公厅印发《科学数据管理办法》[J]. 中国科技财富, 2018 (4): 5.

[3] 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.

[4] 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.

[5] 王丹丹，任婧媛. 国外主要社会科学数据管理平台建设研究及启示[J]. 图书情报工作，2023，67（3）：131-139.

[6] 吴思竹，李赞梅，崔佳伟，等. 基于全球研究数据注册仓储Re3data.org的医学科学数据仓储建设[J]. 中华医学图书情报杂志，2018，27（9）：20-31.

[7] 黄国彬，王涛. 综合型科学数据仓储元数据创建服务研究[J]. 图书情报工作，2021，65（21）：131-140.

[8] 谷秀洁. 开放型机构知识库著作权管理研究[M]. 上海：上海交通大学出版社，2013.

[9] 龙艺璇，赵昆华，王胜兰，等. 从开放获取到开放科学：科研资助机构的选择与挑战[J]. 信息资源管理学报，2021，11（4）：70-79.

[10] ELGER K K. New features of the re3data registry of research data repositories[C]//Agu Fall Meeting. AGU Fall Meeting Abstracts，2016.

[11] GONZÁLEZ-BARAHONA J M，ROBLES G. Revisiting the reproducibility of empirical software engineering studies based on data retrieved from development repositories[J]. Information and Software Technology，2023，164：107318.

[12] PAMPEL H，VIERKANT P，SCHOLZE F，et al. Making research data repositories visible：the re3data.org registry[J]. PLoS One，2013，8（11）：e78080.

[13] ALTENHÖNER R，BLÜMEL I，BOEHM F，et al. NFDI4Culture-consortium for research data on material and immaterial cultural heritage[J]. Research Ideas and Outcomes，2020，6：57036.

[14] RÜCKNAGEL J，VIERKANT P，ULRICH R，et al. Metadata schema for the description of research data repositories：version 3.0[EB/OL]. [2023-11-06]. https://gfzpublic.gfz-potsdam.de/rest/items/item_1397899_6/component/file_1398549/content.

[15] 李辉，曾文，付宏，等. 科学数据出版研究现状及启示[C]//2019年北京科学技术情报学会学术年会：“科技情报创新缔造发展新动能”论坛，2019.

[16] 阿儒涵，吴丛，李晓轩. 科研数据开放的国际实践及对我国的启示[J]. 中国科学院院刊，2020，35（1）：11-18.

[17] WITT M，PAMPEL H，ZHANG X L，et al. Re3data.org-a global registry of research data repositories[C]//EGU General Assembly Conference Abstracts，2014.

[18] 崔佳伟，吴思竹，邬金鸣，等. 科学数据仓储元数据标准研究与启示[J]. 数字图书馆论坛，2019（6）：19-28.

[19] 赵华，赵瑞雪，金慧敏，等. 农业科技大数据仓储建设与服务[J]. 数字图书馆论坛，2020（8）：48-55.

[20] 王辉，WITT M. 基于re3data的科研数据仓储全景分析[J]. 图书情报工作，2017，61（22）：69-76.

[21] 陈昕，郑晓欢，潘博雅，等. 中国科学院科学数据中心体系建设实践及展望[J]. 中国科学数据：中英文网络版，2023，8（1）：139-157.

[22] 完颜邓邓. 国外科学数据仓储元数据实践调查及启示[J]. 新世纪图书馆，2016（5）：81-84.

[23] 贾欢，李泽锋，刘越男. 多学科科学数据仓储元数据方案比较研究[J]. 档案管理，2022（4）：61-64.

[24] 赵怡萌，邱春艳. 科学数据的方法元数据建设与应用现状研究：以生物学领域为例[J]. 图书馆学研究，2021（15）：54-63.

[25] 张勇，苏学，谢振峰. 面向科技大数据的元数据仓储建设实践探索[J]. 情报工程，2020，6（6）：84-96.

[26] 姜璐璐，张泽钰，李宗闻，等. 全球科学数据仓储平台的建设实践现状与展望[J]. 中国科学数据：中英文网络版，2023，8（1）：175-195.

作者简介

袁烨，男，博士研究生，研究方向：科研管理，E-mail：s2179270@siswa.um.edu.my。
陈媛媛，女，博士，教授，研究方向：数据治理、数据服务。

Comparative Study of Research Data Repositories in China and the United Kingdom Based on re3data

YUAN Ye[1]　CHEN YuanYuan[2]
（1. University of Malaya, Kuala Lumpur 50088, Malaysia; 2. School of Information Management, Heilongjiang University, Harbin 150080, P. R. China）

Abstract: This paper utilizes re3data as the data source, selecting 406 research data repositories from China and the United Kingdom as the research objects. Comparative analyses are conducted across five aspects, including distribution characteristics, responsibility types, repository licenses, technical standards, and quality criteria, encompassing 11 indexes. The aim is to provide suggestions for the sustainable development of data repositories in China. Recommendations include fostering extensive connections with heterogeneous institutions both domestically and internationally, promoting interdisciplinary exchanges and collaboration, effectively expanding repository license permissions and types, optimizing the current application status of technical standards, and enhancing the flexibility of metadata utilization.

Keywords: Research Data; Research Data Repository; re3data; China; The United Kingdom

（责任编辑：王玮）