

国外科学数据监护平台调研分析*

汤子钰 欧石燕

(南京大学信息管理学院, 南京 210023)

摘要: 目前国内外均涌现出各种数据监护平台, 旨在为科学研究提供数据服务保障。但与国外发达国家相比, 我国数据监护平台的开发建设起步较晚且数量少, 缺乏成熟的建设经验。因此, 本文选择了20个国外代表性数据监护平台, 对其使用的数据生命周期模型、技术规范、组件、软件工具、功能等各方面进行了全面调研、分析和比较, 并在此基础上, 提出一个整体的数据监护平台构建框架, 为我国数据监护平台的开发与建设提供参考。

关键词: 数据监护; 数据监护平台; 调研; 数据监护平台框架

中图分类号: G250.76

DOI: 10.3772/j.issn.1673-2286.2019.04.006

在e-Science和大数据环境下, 数据密集型科学研究正在成为当代科学研究的主流范式, 科研活动逐步转向以科学数据为中心, 但同时面临着数据量大、数据类型繁多、数据价值密度低等诸多困难。数据监护将数据管理由静态、消极的保存上升到动态、积极的监护, 极大地提升了数据的应用价值, 受到学术界、企业界和政府等各方面的广泛关注, 国外相继涌现了大批数据监护平台。数据监护平台是为数据监护服务提供载体的基础设施平台, 是数据监护由理论走向实践的重要支撑。本文对国外代表性数据监护平台的技术规范、组件、软件工具、功能等各方面进行全面调研、分析和比较, 旨在为我国数据监护平台的开发与建设提供参考。

1 数据监护概述

数据监护萌生于20世纪90年代中期的数字资源长期保存研究与实践。随着研究的深入, 研究者们的关注重点从确保数据的完整、真实逐步转变为对数字资源的积极利用^[1]。2001年, 数字保存联盟(Digital Preservation Coalition)与英国国家空间中心(British National Space Centre)在伦敦联合举办了“Digital Curation: Digital Archives, Libraries and e-Science Seminar”国际研讨会, 首次提出数字监护(Digital

Curation)和数据监护(Data Curation)概念^[2]。这两个概念虽然前者多用于人文社科领域, 后者多用于自然科学领域^[3-4], 但并没有本质区别, 都是指主动管理处于生命周期中的科学数据, 从而促进数字资源共享的活动^[5], 在很大程度上可以互用。因此, 在本文中我们统一使用数据监护一词。

数据监护与原有的数据保存(Data Preservation)和数据存档(Data Archiving)这两个概念既有区别又有联系。英国数字数据监护任务组(Digital Data Curation Taskforce)认为, 数据保存是数据存档的一个方面, 而数据存档则是数据监护的基础和前提^[6]; 英国数字监护中心(Digital Curation Center, DCC)认为, 与数据保存相比, 数据监护在确保数据长期可用的同时, 能够为数据创建者和用户带来即时价值^[7]; 英国联合信息系统委员会(Joint Information Systems Committee, JISC)认为, 数据监护活动包含了数据存档和数据保存, 存档和保存只是数据监护多种活动中的一部分, 更为重要的是数据监护对数据进行增值活动以生成新的信息和知识^[8]; 国内学者杨鹤林^[9]认为, 数据存档从内容层面保证数据的再利用, 而数据保存从数据监护的技术层面确保数据的长期可用性; 张智雄等^[10]认为, 数据保存是消极被动的, 而数据监护则是一项积极主动的活动。综合上述观点可以看出, 数据监护更强调数据的生命周期

*本研究得到国家社会科学基金重点项目“基于关联数据的学术文献内容语义发布及其应用研究”(编号: 17ATQ001)资助。

管理,是贯穿数据整个生命周期过程的持续性数据管理活动,目标是数据的价值增值,因此在数据管理活动中具有更积极的意义^[11]。

数据监护的实施离不开数据监护平台的支撑。目前,有关数据监护平台的研究在欧美国家相对成熟和系统,理论与实践结合紧密。理论研究主要集中在数据管理、数据生命周期模型、平台基础设施和平台架构4个方面。如Bishoff等^[12]选取并分析了182个数据管理规划的具体内容,得出数据管理规划能够提高数据服务质量的结论;Pejša等^[13]以地震工程学数据监护平台NEES为例,构建了适合地震工程学领域的模型;Kerrien等^[14]从数据监护政策、数据类型、数据描述、项目合作、平台界面等方面,剖析了IntAct平台框架。在实践方面,国外数据监护平台实例众多、发展成熟,如美国的考古学平台tDAR^[15]、英国的海洋学平台MEDIN^[16]、澳大利亚的综合性数据监护平台ANDS^[17]等。

相较国外,我国针对数据监护平台的相关研究多数停留在理论层面,主要包括数据生命周期模型、数字资源仓储系统和国外代表性数据监护平台调研等方面。如殷沈琴等^[18]对DSpace、Fedora、Dataverse 3款数字资源仓储系统进行分析,评估各软件的基本功能、高级功能、使用的元数据规范和在线分析功能;杨鹤林^[19]分析了康奈尔大学DataStaR平台的运作模型,提出我国数据监护平台的构建应学习DataStaR的创新模式,以提高数据服务质量。实践方面,国内虽然也开发构建了一些数据监护平台,但数量较少,仅有北京大学开放研究数据平台、复旦大学社会科学数据平台、中山大学社会科学调查中心、中国科学院数据云等,且主要依赖于国外数字资源仓储系统(如哈佛大学的Dataverse^[20]、MIT和HP实验室的DSpace^[21])进行支撑。

总体来看,国内数据监护平台在理论研究和实践研究方面都与国外存在较大差距。因此,国内平台的开发与构建亟需学习国外较为成熟的理论与实践经验。

2 数据监护平台调研

为保证调研样本的代表性和调研结果的可靠性,我们根据领域覆盖面广、学科多样性强、平台详细技术信息可获得、数据存储量大的原则,本文选取了20个有代表性数据监护平台,分别来自英国、美国、澳大利亚、欧盟等国家或组织,涵盖人文社会科学、地球科学、物理科学、生物医学等学科领域。这些平台或是全

球性的服务平台,抑或是所在国家或领域比较权威的数据监护平台,具有知名度高、用户界面友好、服务功能完善、管理制度和构建技术成熟等特点。表1列举了所调研的数据监护平台的基本信息。

2.1 数据生命周期模型

数据监护平台与传统数据存储系统的最大区别在于,数据监护平台对数据整个生命周期的各个阶段都进行管理,从而保持数据的长期可用与价值增值;而传统数据存储系统主要是以数据检索与发现为目的,对数据进行简单描述和一次性静态存储,很少或根本不提供数据处理、数据分析等数据增值功能。

数据生命周期模型是规划监护活动并抽象描述数据监护活动各个阶段的概念框架^[22]。数据生命周期模型提供了在项目或机构中组织数据管理相关任务和活动的结构。不同组织机构根据需要提出不同的数据生命周期模型,从不同角度描述了数据从产生、收集、描述、存储、发现、分析到再利用的整个生命周期。本文调研的20个平台中,80%的平台都是基于数据生命周期模型进行数据管理,主要采用的模型有4个,具体介绍如表2所示。

(1) DCC生命周期模型。该模型由英国数据监护中心于2008年提出^[23],用于规划特定研究项目、组织或联盟内的数据管理活动,以确保按照正确的顺序管理科研资源^[24]。DCC模型将数据生命周期划分为6个主要阶段:概念化,创建和接收数据,评测和选择数据,长期保存和存储,访问、使用和重用,以及转换^[25]。

(2) OAIS (Open Archival Information System) 模型。该模型是由美国国家航空航天局(NASA)和美国空间数据系统咨询委员会(CCSDS)于1999年联合制定的标准^[26],其目的在于规范数字资源长期保存的概念和模型^[27],包括功能模型、信息模型和互操作模型^[28]。其中,功能模型涵盖数据收集、归档存储、数据管理、管理、保存规划、访问6个实体;信息模型用于解释数字保存系统中信息对象的结构和种类。三者共同作用,描述整个数据管理过程中不同信息对象的内容和结构^[29]。

(3) DDI (Data Documentation Initiative) 生命周期模型。该模型由数据文献计划联盟构建,是一个以社会科学数据为服务对象的线性模型。它将数据生命周期划分为概念研究、数据采集、数据处理、数据存档、数据发布、数据发现、数据分析和数据重用8个阶段^[30],旨

表1 调研的数据监护平台基本信息

平台名称	全 称	访问网址	成立时间	负责机构	资金来源	学科领域	数据量
tDAR	Digital Archaeological Record	https://www.tdar.org/	2008年	数字遗产非营利组织	安德鲁·梅隆基金会、JISC、美国国家科学基金会（NSF）等	人文社会科学（考古学）	约3万份文献
ADS	Archaeology Data Service	http://archaeologydataservice.ac.uk/	1996年	英国约克大学考古数据服务（ADS）管理协会	欧盟、英国艺术与人文研究委员会（AHRC）、JISC、英国自然环境研究委员会（NERC）	人文社会科学（考古学）	超过2万份文献
BODC	British Oceanographic Data Centre	https://www.bodc.ac.uk/	1989年	英国国家海洋学中心	安德鲁·梅隆基金会、NERC、JISC、NSF	地球科学（海洋科学）	超过9 000个数据集
SeaDataNet	SeaDataNet	http://www.seadatanet.org/	2004年	欧盟	欧盟	地球科学（海洋科学）	超过4 084个数据集
CCDC	The Cambridge Crystallographic Data Centre	https://www.ccdc.cam.ac.uk/	1987年	英国剑桥晶体数据中心	英国科学技术信息办公室、英国科学与工程研究协会（IET）	物理科学（晶体科学）	超过87万条记录
NEEShub	The Network for Earthquake Engineering Simulation	https://nees.org/	2004年	地震工程仿真网络（NEES）联盟	NSF	物理科学（地震工程学）	-
NCEI	National Centers for Environmental Information	https://www.ncei.noaa.gov/	2015年	美国国家海洋大气部门	美国政府	地球科学（环境科学）	超过20PB的数据资源
UKCCSRC	The UK Carbon Capture and Research Centre	https://ukccsrc.ac.uk/	2012年	英国国家地球科学数据中心	英国工程与物理科学研究协会（EPSRC）、英国商业能源和工业战略部（DECC）	地球科学（能源科学）	-
GBIF	Global Biodiversity Information Facility	http://www.gbif.org/	2001年	全球生物多样性信息机构（GBIF）理事会	全球60个国家、50多个全球性组织	生物科学	超过35 220个数据集
FishNet	FishNet	http://www.fishnet2.net/	1999年	美国自然历史博物馆	美国国家生物信息基础设施（NBII）计划、NSF	生物科学	超过2 000万条鱼标本
BioModels	BioModels	http://www.ebi.ac.uk/biomodels-main/	2006年	欧洲分子生物学实验室	欧洲分子生物学实验室（EMBL）、美国国立卫生研究院（NIH）	生物科学	超过14万个生物模型
Protein Data Bank	Protein Data Bank	http://www.wwpdb.org/	1971年	全球蛋白质数据银行（PDB）协会	美国国家卫生研究院和能源部、NSF	生物科学	超过13万条记录
eCrystals	eCrystals	http://ecrystals.chem.soton.ac.uk/	2003年	英国南安普敦大学	JISC	物理科学（晶体科学）	超过1 026条记录
Dryad	Dryad	http://www.datadryad.org/	-	-	欧盟、NSF	医学等其他科学领域	约14 000份文献
PURR	The Purdue University Research Repository	https://purrr.purdue.edu/	-	美国普渡大学	-	综合领域	约628个数据集

续表

平台名称	全 称	访问网址	成立时间	负责机构	资金来源	学科领域	数据量
DataShare	Edinburgh DataShare	http://datashare.is.ed.ac.uk/	2007年	英国爱丁堡大学 信息服务中心	JISC	综合领域	约1 933条记录
ANDS	Australian National Data Service	http://www.ands.org.au/	2009年	澳大利亚 莫纳什大学	澳大利亚政府	综合领域	超过13万个 数据集
Dataverse Network	Dataverse Network	http://dataverse.org/	2006年	美国哈佛大学	美国国会图书馆、NIH、NSF	人文社会科学	约5万个数据集
ICPSR	Inter-University Consortium for Political and Social Research	http://www.icpsr.umich.edu/icpsrweb/	1962年	美国密歇根大学	政治与社会研究校际 联盟 (ICPSR)	人文社会科学	超过25万份文献
UKDA	UK Data Archive	http://www.data-archive.ac.uk/	1967年	英国艾塞克斯 大学	英国经济与社会研究委 员会 (ESRC)、JISC	人文社会科学	超过7 211个 数据集

表2 数据生命周期模型基本信息

数据生命 周期模型	数据生命周期划分	应用的数据监护平台	
		数量/个	名 称
DCC	6个主要阶段: 概念化, 创建和接收数据, 评测和选择数据, 长期保存和 存储, 访问、使用和重用, 转换	5	tDAR、ADS、GBIF、 ANDS、eCrystals
OAIS	6个功能实体: 数据收集、归档存储、数据管理、管理、保存规划和 数据访问	6	BODC、SeaDataNet、CCDC、 Protein Data Bank、NCEI、FishNet
DDI	8个阶段: 概念研究、数据采集、数据处理、数据存档、数据发布、数据发现、 数据分析和数据重用	4	UKDA、ICPSR、PURR、 Dataverse Network
UKDA	6个阶段: 数据创建、数据加工、数据分析、数据保存、数据访问、数据再利用	3	ANDS、UKDA、UKCCSRC

在对整个生命周期中的科学数据进行管理和监护。

(4) UKDA (UK Data Archive) 生命周期模型。

该模型由英国数据档案组织构建, 旨在保存高质量的研究数据以供分析和再利用^[31]。UKDA模型将数据生命周期划分为6个阶段: 数据创建、数据加工、数据分析、数据保存、数据访问和数据再利用^[32]。

根据上述调研可以看出, DCC模型、OAIS模型、DDI模型和UKDA模型虽然在具体结构和细节上有所区别, 但有关数据生命周期阶段的划分是有相通之处的。4个模型定义的数据生命周期大致包括数据管理规划、数据采集、数据处理、数据保存、数据访问、数据分析与利用这6个阶段。因此, 本文以科学数据的生命周期管理为切入点, 对数据监护平台在数据生命周期各个阶段采用的协议规范、技术实现方式, 以及提供的功能等方面进行调研分析。

2.2 数据管理规划

为保证科研项目的顺利进行和科学数据的可追溯、可重用, 越来越多的基金会和其他研究资助者要求无论是在投标准备阶段还是资金获得后, 科研人员和团队都应明确如何对科学数据进行收集、处理和生成。数据管理规划, 是概要介绍科研项目进行中以及结束后科学数据将如何被有效处理的正式文档^[33]。数据管理规划制定于科学数据生命周期管理的起始阶段 (即在科学数据产生前), 但它不是固定不变的, 通常在项目的生命周期中被不断充实和细化完善^[34]。数据管理规划对开展数据监护工作具有指导和驱动意义。

为满足科学数据管理的“FAIR”原则^[35], 即可寻找 (findable)、可访问 (accessible)、可交互 (interoperable)、可再用 (re-usable), 英国数据监护中心提出数据管理规划应包含5项信息, 分别是数据创建规范、数据组织

管理的标准、道德规范和知识产权问题、数据共享和访问计划,以及长期保存策略^[36]。

基金组织为数据监护平台开展或支持的科研活动提供资助,科研人员或团队为获得赞助,必须保证科研产出的质量满足相应的要求。不同的基金组织对数据管理规划的要求不一样,如表3所示。本文调研的6个基金组织均要求在数据管理规划中明确数据创建规范,部分基金组织未对道德规范与知识产权、数据归档和保存的相关内容做要求。数据管理规划一方面能够帮助研究人员明确资助者的具体要求,同时也为数据生命周期不同阶段的管控提供明确的指导,保证科研项目的产出质量。

表3 不同基金组织对数据管理规划内容的整体要求

基金组织	数据创建规范	数据组织管理	数据归档和保存	道德规范与知识产权	数据共享和访问
NSF	√	√	√		√
ICPSR	√	√	√	√	√
ESRC	√	√		√	√
AHRC	√	√			√
EPSRC	√	√	√	√	√
NERC	√	√		√	√

2.3 数据采集

数据采集是数据生命活动的起点,在该阶段,数据监护平台为采集到的数据集提供元数据描述,为后续的数据管理和利用提供基础^[37]。元数据在科学数据的组织、存储、检索、引用、共享等生命周期的各个环节都起到至关重要的作用。在不同的学科领域,根据其数据特点,通常使用不同的元数据规范。如ABCD (Access to Biological Collections Data) 元数据规范主要用于描述活生物体标本等初级生物多样性数据^[38]; DIF (Directory Interchange Format) 规范用于描述地球科学领域的相关项目中,数据捕获仪器、数据时间和空间特性等元素; CIF (Crystallographic Information Framework) 是物理科学领域主要用于描述晶体结构测定的数据。本研究所调研的20个数据监护平台涉及不同的学科,主要使用11种元数据规范,如表4所示。

在所调研的20个数据监护平台中,应用最广泛的是DC元数据规范,有7个平台使用,占比35%;其次是DDI规范,有3个平台使用,占比15%。DC属于通用元数

据规范,可扩展性强,广泛用于各学科领域^[45];而DDI是DC的延伸与扩展,在人文社会科学领域应用广泛^[46]。Darwin Core与DDI一脉相承,是DC元数据在生物学领域的扩展。FGDC/CSDGM由美国联邦地理数据委员会 (Federal Geographic Data Committee, FGDC) 颁布,是最广泛描述地理空间数据的标准。除此之外,为满足部分学科数据描述的特殊要求,很多数据监护平台融合采用多种规范或自行定义合适的元数据框架。譬如,除Darwin Core外,GBIF为描述纸质地图等非数字资源采用了EML规范;蛋白质数据银行开发了PDBx/mmCIF规范,用于准确描述蛋白质、核酸3D结构信息和大分子晶体信息。这些方式有利于提高元数据标准针对性,快速满足科研项目实践的需求。

除学科因素外,科学数据所处的生命周期阶段也会影响到元数据规范的选择。在不同阶段,通常选择全部用蛋白质数据银行不同的元数据规范描述数据以达到不同的目的。如美国普渡大学研究资料库 (PURR) 为兼顾数据描述、发现、可追踪和保存等各个阶段的数据管理需求,将几个元数据标准整合到一起,形成一整套PURR的元数据解决方案^[47]。在数据创建阶段,METS (Metadata Encoding and Transmission Standard) 标准表示数据集文件的结构和层次;在数据保存阶段,MODS描述了数据集的访问权限,PREMIS (Preservation Metadata Implementation Strategies) 标准记录每个数据集经历的保存事件;在数据利用阶段,采用DC元数据以支持数据的检索和发现^[48]。

为帮助科研人员准确描述资源信息,数据监护平台通过发布指导性文件和要求用户提交表格或制定元数据管理工具等方式提供元数据创建服务。ICPSR平台发布文档《元数据创建最佳实践》(Best Practice in Creating Metadata),推荐科研用户使用DDI标准来创建元数据,还补充设置了资金来源、项目描述、数据加权等十多个元数据项目;DataShare在《储户用户指南》(Edinburgh Datashare: Depositor's User Guide) 中描述了提供元数据的具体步骤,包括数据项名称、创建者信息、数据类型、项目基金等。Dyrad和Dataverse Network平台通过用户提交表格的形式创建元数据。Dyrad要求数据创建者上传短视频,阐述表格内容的填写原因和过程;Dataverse Network结合元数据标准体系和用户需求,设置下拉菜单为用户提供准确的选择。目前,应用较为广泛的元数据创建工具包括英国数据监护中心制定的DMPtools、美国联邦地理数据

表4 数据监护平台中使用的科学数据集元数据规范^[39]

元数据规范	全 称	简 介	学科领域	数据监护平台
ABCD	Access to Biological Collections Data	主要用于描述初级生物体的数据	生物科学	GBIF
CF	Climate and Forecast	用于描述地球科学数据	地球科学	BODC
CIF	Crystallographic Information Framework	主要用于描述晶体结构测定的数据	物理科学	CCDC
Darwin Core	Darwin Core	DC元数据在生物学领域的扩展和延伸 ^[40]	生物科学	GBIF、FishNet
DC	Dublin Core	简单、易懂的通用型元数据规范	通用领域	ADS、BODC、eCrystals、Dryad、Edinburgh DataShare、ANDS、Dataverse Network
DDI	Data Documentation Initiative	主要用于描述社会、行为、经济和健康科学中通过调查或其他观察方法产生的数据	人文社会科学	Dataverse Network、ICPSR、UKDA
DIF	Directory Interchange Format	起源于地球科学领域，旨在描述数据捕获仪器、数据时间和空间特性，以及其他相关联项目的元素 ^[41]	地球科学	BODC
EML	Ecological Metadata Language	可用于描述数字资源，也用于描述纸质地图等非数字资源	生物科学	GBIF
FGDC/CSDGM	Federal Geographic Data Committee/Content Standard for Digital Geospatial Metadata	描述美国联邦政府所要求的数字地理空间数据 ^[42]	地球科学、人文社会科学	BODC、Dataverse Network
MIDAS Heritage	MIDAS Heritage	用于记录建筑物、考古遗址、沉船、公园和花园、战场、文物等信息 ^[43]	地球科学、人文社会科学	ADS
PDBx/mmCIF	Protein Data Bank Exchange Dictionary and the Macromolecular Crystallographic Information Framework	关于蛋白质、核酸3D结构信息的全局档案库，在此基础上开发了大分子晶体信息的描述框架（mmCIF） ^[44]	生物科学	Protein Data Bank

委员会提供的地理空间科学元数据编辑器EME（EPA Metadata Editor）等。

2.4 数据处理

在数据处理阶段，数据监护平台主要负责数据审核和格式管理的工作。

数据审核主要是指通过审核数据价值、元数据、数据格式等来保障科研数据的质量、可获得性和兼容性。UKDA平台在从数据的摄入到保存、再到数据发布过程中提供格式转换和转化质量检查的功能^[49]；ICPSR平台对科研项目是否处于核心研究领域、是否优

化现有工具、科研数据定量或定性等方面都有评估要求。格式管理是指转化并规范科学数据格式。有的监护平台还支持将存储的数据转换为更加易读的格式供用户下载。

调研发现，数据监护平台对数据上传格式的要求通常有3种，分别是上传格式无限制、上传指定格式的数据和上传格式符合特定标准。上传格式无限制的数据监护平台通常提供格式转换工具，能够对上传的数据进行处理，将其转换为平台支持的存储格式。如OCTOPUS是SeaDataNet平台提供的一种多重样式检测、转化和拆分工具，它将给定SeaDataNet格式的文件转换为另一种SeaDataNet格式。SeaDataNet平台还提

供NEMO工具,可以将数据在ASCII和自有格式ODV4之间相互转化。上传指定格式的数据是指数据监护平台只支持规定格式数据的上传。如GBIF平台只支持上传文本、关系型数据表和XML格式的数据^[50];有的数据监护平台虽然没有规定上传数据的具体格式,但是要求数据格式必须要满足特定标准,如Protein Data Bank平台规定上传数据需符合PDBx/mmCIF标准^[51-52]。

大部分数据监护平台的数据存储格式和下载格式基本相同,主要为文本、图像、关系型数据表等格式,且经过数据处理和格式转换流程之后,比数据最初的上格式更加规范。为了满足管理特定类型数据的需求且达到资源共享的目的,有些数据监护平台支持视频、音频、关系型数据库、元数据(XML、RDF)、程序代码、GIS等小众格式。如ADS平台能够通过VR(Virtual Reality)格式存储考古文物的信息,通过GIS格式定位考察地点;tDAR平台还支持遥感和3D扫描格式的数据;有些数据平台则要求按照ASCII、netCDF、HDF等标准存储数据。

2.5 数据保存

在数据保存阶段,数据监护平台的工作主要涉及数据资源的永久性标识和数据仓储软件。

永久标识符是对数字资源进行唯一标识的工具,在数字资源的可获得性和重复利用方面发挥重要作用。目前,具有代表性的数字资源标识符有数字对象标识符(Digital Object Identifier, DOI)、永久统一资源定位符(Persistent Uniform Resource Locators, PURL)和句柄系统(Handle System)3种。在所调研的20个数据监护平台中,15个平台采用DOI标识符,5个平台采用PURL标识符,仅有Dataverse Network平台同时采用了DOI标识符和句柄系统。PURL基于标准的Web资源地址URL,具有易推广的优势;而DOI则具有较强的兼容性和语义互操作性,管理机制发展完善,解析系统也较为成熟,同时不断推出CrossRef Search等增值服务^[53]。目前,DOI在数字资源标识中的应用最为广泛。除这3种通用的数字资源标识符,数据监护平台在积极探求更好的数据标引方式。如Dataverse Network平台提供的数据引用功能,在永久性标识符的基础上附加了基于每项数据集内容的通用数字指纹,以应对数字资源格式变化带来的问题。

数字资源仓储系统是构建数据监护平台的基础,

负责对数据监护平台中的数据进行仓储管理。主流的仓储系统有Fedora、Eprints和Dataverse 3款,皆为免费开源软件。在所调研的20个数据监护平台中,绝大部分平台都采用自主研发软件,因为定制软件能提供具有针对性的数据服务并具有专业化的管理能力。如BODC平台利用Linux、Microsoft Access、C++、Oracle SQL等自行开发仓储平台。仅有4个平台选择了开源软件,其中ADS和FishNet平台采用Fedora软件, Dataverse Network平台采用Dataverse软件, UKDA平台采用Eprints系统来管理机构的数据资源。采用免费开源软件的优势是简单易行,对数据机构的经费需求和技术要求都比较低。

2.6 数据访问

数据监护平台的一个重要功能是提供对所有存储的数据集的检索与访问,为实现跨平台检索与访问,数据监护平台往往通过数字资源互操作协议来实现不同平台间的互操作。

数字资源互操作协议是为了实现不同机构与系统之间数字资源共享所制定的协议标准,为数据监护平台之间的互操作提供了支持。目前,具有代表性的数字资源互操作协议有Z39.50和OAI-PMH协议两种。在所调研的20个数据监护平台中,19个平台遵循OAI-PMH协议,仅有ADS平台同时遵循OAI-PMH和Z39.50两种协议。由此可见,OAI-PMH协议在数据监护平台中的应用远超Z39.50协议,其原因是:一方面,Z39.50起源于图书馆界,主要应用于书目数据共享领域,虽然具有丰富的语义和强大的数据处理功能,但是结构复杂、数据处理效率低,不适于描述数字资源;而OAI-PMH协议起源于数字出版界,用于电子文档共享,基于跨学科的DC元数据规范,更加灵活通用^[54];另一方面,2001年美国数字图书馆联盟(Digital Library Federation)实施了采用OAI-PMH协议的通用网关计划,并获得美国安德鲁·梅隆基金会的大量经费支持,从而促进了该协议在美国和欧洲的推广。目前看来,OAI-PMH协议正获得越来越广泛的应用,同时图书馆领域也在开发基于XML编码规则的新一代Z39.50协议Zing(Z39.50 International: Next Generation),以期拓宽其适用范围。

数据监护平台检索项的设置关系到利用数据资源时的查全率和查准率。在20个数据监护平台中有6个平

台只提供一个通用的文本检索框,不区分检索内容的类型。我们一共总结出14个数字资源检索项,其中10个是单一检索项,4个是复合检索项。20个调研平台中,各检索项的出现频次如图1所示。其中,70%提供标题和人物检索项,近50%提供关键词和时间检索项,部分平台还提供机构/国家、描述/摘要/注释、主题、数据类型、全文、来源、地点、数据库/数据集合、项目、仪器/设备等检索项。

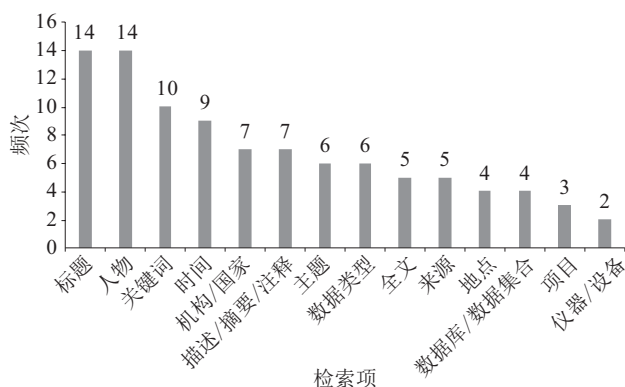


图1 数据监护平台检索项的频次统计

2.7 数据分析与利用

数据分析是数据监护平台实现数据价值增值的重要手段。在调研的20个平台中,主流的数据分析软件包括Nesstar、SDA和Dataverse。Nesstar统计功能较简单,易用性强,能够处理调查数据、列联表及文本资源,对数据进行统计描述、列联表、相关分析和回归分析,数据结果能够以表格、直方图、条形图等方式展示。SDA是一款功能全面、统计专业性强的在线分析软件,主要包括频数与交互列表、均值比较、相关矩阵、相关性检验、多元回归、Logit/Probit回归等分析功能,结果以表格、直方图、条形图等形式展现。Dataverse的在线分析功能提供描述性分析和高级统计分析,科研人员在浏览器窗口即可进行分析工作,无须下载整个数据。Dataverse还提供数据格式自动转化功能,将数据资源转化为.tab格式再进行分析处理,分析结果显示为各类图表、时间序列分析、GraphML等。

除上述3种通用的数据分析工具外,数据监护平台根据学科属性和科研项目需求,提供特定的数据分析工具。譬如,SeaDataNet平台采用开源的ODV (Ocean Data View) 软件包作为数据分析和可视化软件。ODV具有非常丰富的交互功能,提供海洋学和其他地理参

考剖面或序列数据的交互式探索,并且支持各种各样的绘图类型;CCDC平台开发了Python的API接口,允许用户直接导入数据,并通过Python强大的工具进行数据分析;UKCCSRC平台服务于地球科学领域的科研项目,不仅采用Nesstar,还开发了GIS地理可视化服务。数据监护平台定制数据分析工具,学科和项目针对性更强,处理数据更加高效可行。

3 数据监护平台框架

基于对20个数据监护平台各组件及其支撑技术与规范协议的调研结果,我们提出一个整体的数据监护平台框架,如图2所示。该框架以科学数据的生命周期为切入点,围绕数据管理规划、数据采集、数据处理、数据保存、数据访问、数据分析与利用6个核心阶段,设计数据监护平台的各项功能及其相应的支撑组件。



图2 数据监护平台框架

数据管理规划是数据监护平台特有的阶段之一,对管理整个生命周期中的数据起到纲领性作用。通过制定数据管理规划,研究人员一方面能够明确外部政

策,即资助者的具体要求;另一方面也满足了科研需要,为数据生命周期不同阶段的管控提供指导。

在数据采集阶段,数据监护平台提供数据创建和数据发布两项基本功能。在数据创建时可根据所属学科选择合适的元数据规范来描述数据。创建元数据时不仅要考虑学科背景,还要结合科研项目的特殊要求以及科学数据所处的生命周期等因素,制定元数据管理工具;与此同时,可以通过发布指导性文件、要求用户提交表格或制定元数据管理工具等方式指导用户进行操作。相应的,数据监护平台应配备有元数据创建工具、数据发布软件等基础设施以支撑这两项功能的实现。

在数据处理阶段,数据监护平台应提供数据审核和数据格式转换的工具。数据审核是指对数据内容和格式、数据集大小、元数据质量等方面的评估。数据格式转换工具应用于科学数据生命周期的多个阶段,从数据上传到数据存储,再从数据保存到数据浏览和下载。在数据上传过程中,如果对用户上传的数据无格式要求,则平台应在数据处理阶段对资源进行格式转换;如果平台在用户上传数据时就做出要求,如上传指定格式的数据,则数据格式转换工具不是必需的。

在数据保存阶段,数据监护平台的主要工作是对科学数据进行标引和构建数据仓储系统。数字资源永久性标识符保证了科研数据的长期可获得、可应用,目前主流的标识符包括DOI、PURL和句柄系统,数据监护平台也在积极探求更好的数据标引方式,以应对数字对象格式或内容的变化。数字资源仓储系统对科研数据进行仓储管理,是构建数据监护平台的基础。主流的免费开源软件简单易行,包括Fedora、Eprints和Dataverse等,但大部分平台选择采用自主研发软件,能提供更具有针对性和专业性的数据服务。相应的,数据监护平台应配备数据库管理工具、数据备份工具、数据仓储软件等基础设施以支撑这两项功能的实现。

在数据访问阶段,数据监护平台的主要工作是提供数据检索和数据互操作。目前具有代表性的数字资源互操作协议有Z39.50和OAI-PMH协议两种,其中OAI-PMH协议起源于数字出版界且基于跨学科的DC元数据规范,更加灵活通用。数据监护平台检索项的设置关系到利用数据资源时的查全率和查准率。通过调研发现,数据监护平台的数据集检索项可分为3类,即基本检索项、推荐检索项和参考检索项。基本检索项包括标题和人物2项,是必须提供的检索内容项;推荐检索项包括关键词、时间、机构/国家、描述/摘要/注

释、主题、数据类型、全文7项,是科学数据集的通用属性,但是区分度不够高,因此笔者推荐数据监护平台提供这些检索项,但是不做必须要求;参考检索项包括来源、地点、数据库/数据集、项目、仪器/设备5项,属于科学数据集的非通用属性,只针对某些学科领域的部分科学数据,特定领域的数据监护平台可参照执行。

数据分析与利用主要提供数据的价值增值,是数据生命周期的核心阶段之一。目前主流的数据分析软件包括Nesstar、Dataverse和SDA。部分数据监护平台也会根据学科属性和科研项目需求,开发特定的数据分析工具。

4 结论

本文首先界定了数据监护和数据监护平台的相关概念,然后以数据生命周期为切入点,调研分析了国外20个有代表性的数据监护平台,得出如下结论。

目前国内外有关数据监护平台的研究差距较大。欧美国家开展数据监护工作较早,理论与实践结合紧密;国内关于数据监护平台的理论与实践研究尚处于起步阶段,主要停留在数据生命周期模型、数字资源仓储系统和国外代表性数据监护平台调研等方面。因此,主动学习国外成熟平台的构建和管理经验,对我国数据监护平台的发展具有重要价值。

目前,数据监护平台建设仍面临一些问题和挑战。数据管理规划的内容和结构尚未形成统一规范,数据评估审核工作尚未完全落实;由于学科、领域、科研项目等多方面原因,数据监护平台的元数据规范、永久性资源标识符、互操作协议等尚未达成一致标准,严重影响跨学科、跨平台的数据互操作;特殊格式的数字资源缺乏完善的存储机制;目前,数据管理工具开发尚未成熟,元数据无法自动抽取、数据无法自存档等问题也为科研工作的开展带来了不便。统一数据监护工作的相关标准,同时完善数据管理工具,才能保证数据监护平台更好地为科研人员提供服务,发挥科学数据更大的科研和社会价值。

参考文献

- [1] 杨鹤林. 英国数据监护研究成果及其在高校图书馆的应用——DCC建设回顾[J]. 图书馆杂志, 2014, 33(3): 84-90.
- [2] BEAGRIE N, POTHE P. Digital Curation: Digital Archives,

- Libraries and e-Science Seminar [EB/OL]. [2018-10-25]. <http://www.ariadne.ac.uk/issue30/digital-curation/>.
- [3] 王芳, 慎金花. 国外数据管护 (Data Curation) 研究与实践进展 [J]. 中国图书馆学报, 2014, 40 (4): 116-128.
- [4] 王晴. 国内Data Curation研究综述 [J]. 情报资料工作, 2014 (5): 33-38.
- [5] SHREEVES S L, CRAGIN M H. Introduction: Institutional repositories: current state and future [J]. Library Trends, 2008, 57 (2): 89-97.
- [6] How and why you should manage your research data: a guide for researchers | Jisc [EB/OL]. [2019-04-16]. <https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data>.
- [7] ABBOTT D. What is Digital Curation? [EB/OL]. [2018-10-20]. <http://www.dcc.ac.uk/resources/briefing-papers/introductioncuration/What-digital-curation>.
- [8] What's the cost of curating content in the digital age | Jisc [EB/OL]. [2019-04-16]. <https://www.jisc.ac.uk/blog/whats-the-cost-of-curating-content-in-the-digital-age-01-apr-2015>.
- [9] 杨鹤林. 数据监护: 美国高校图书馆的新探索 [J]. 大学图书馆学报, 2011 (2): 18-22.
- [10] 张智雄, 吴振新, 刘建华, 等. Digital Curation和Digital Preservation之概念辨析 [J]. 现代图书情报技术, 2014 (1): 4-13.
- [11] What is Digital Curation? Digital Curation Center [EB/OL]. [2018-10-25]. <http://www.dcc.ac.uk/digital-curation/what-digital-curation>.
- [12] BISHOFF C, JOHNSTON L. Approaches to data sharing: An analysis of NSF data management plans from a large research university [J]. Journal of Librarianship and Scholarly Communication, 2015, 3 (2): 1231.
- [13] PEJŠA S, DYKE S J, HACKER T J. Building infrastructure for preservation and publication of earthquake engineering research data [J]. International Journal of Digital Curation, 2014, 9 (2): 83-97.
- [14] KERRIEN S, ARANDA B, BREUZA L, et al. The IntAct molecular interaction database in 2012 [J]. Nucleic Acids Research, 2012, 38 (Database): 525-531.
- [15] MCMANAMON F, KINTIGH K, BRIN A. Digital Antiquity and the Digital Archaeological Record (tDAR): Broadening Access and Ensuring Long-Term Preservation for Digital Archaeological Data [EB/OL]. [2018-10-25]. <http://csanet.org/newsletter/fall10/nlf1002.html>.
- [16] Marine Environmental Data and Information Network [EB/OL]. [2018-10-25]. <http://www.oceannet.org/>.
- [17] 周晓燕, 宰冰欣. 澳大利亚高校科研数据管理政策制定研究 [J]. 图书馆建设, 2017 (2): 63-70.
- [18] 殷沈琴, 张计龙, 张莹, 等. 社会科学数据管理服务平台系统选型研究——以复旦大学社会科学数据平台为例 [J]. 图书情报工作, 2013, 57 (19): 92-96.
- [19] 杨鹤林. 从数据监护看美国高校图书馆的机构库建设新思路——来自DataStaR的启示 [J]. 大学图书馆学报, 2012 (2): 23-28, 73.
- [20] KING G. An introduction to the dataverse network as an infrastructure for data sharing [J]. Sociological Methods & Research, 2007, 36 (2): 173-199.
- [21] Home-dSPACE [EB/OL]. [2018-10-25]. <https://www.dspace.com/en/inc/home.cfm>.
- [22] HIGGINS S. The dcc curation lifecycle model [C] // Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, Pittsburgh, June 16- 20, 2008. New York: ACM, 2008: 453.
- [23] 周满英, 付禄. 数据策管生命周期模型比较研究 [J]. 图书馆研究与工作, 2018 (9): 34-37, 87.
- [24] Lifecycle Model FAQ [EB/OL]. [2018-10-25]. <http://www.dcc.ac.uk/resources/curation-lifecycle-model/lifecycle-model-faqs#>.
- [25] HIGGINS S. The DCC Curation Lifecycle Model [J]. The International Journal of Digital Curation, 2008, 1 (3): 134-140.
- [26] OAIS Reference Model (ISO 14721) | The fundamental standard for digital preservation [EB/OL]. [2018-10-25]. <http://www.oais.info/>.
- [27] 章燕华, 刘霞. OAIS参考模型: 数字资源长期保存的概念框架 [J]. 浙江档案, 2007 (3): 38-42.
- [28] 李明娟. OAIS参考模型与数字信息长期保存 [J]. 图书情报知识, 2007 (5): 65-69.
- [29] 钱毅. OAIS对数字档案馆系统技术路线和管理策略的启示 [J]. 档案学究, 2009 (4): 46-49.
- [30] LANCIA R. The DAF DDI profile, a metadata set to address digital curation and preservation issues in cultural heritage institutions [J]. Lecture Notes in Computer Science, 2010, 6436: 291-305.
- [31] 孟祥保, 钱鹏. 高校社会科学数据管理的国际经验及其借鉴——以UKDA和ICPSR为例 [J]. 情报资料工作, 2013 (2): 77-80.
- [32] 李杨, 裴丽. 基于UKDA研究数据生命周期的图书馆学科服务模式研究 [J]. 山西档案, 2018 (3): 99-101.
- [33] 邓珞华. 图书情报数学 [M]. 长春: 东北师范大学出版社,

- 1983: 654-668.
- [34] 刘峰, 张晓林. 数据管理规划构成规范及其可操作数据监护模型研究[J]. 现代图书情报技术, 2016 (1): 11-16.
- [35] WILKINSON M D, DUMONTIER M, AALBERSBERG I J, et al. The FAIR guiding principles for scientific data management and stewardship [J]. Scientific Data, 2016, 3 (1): 160018.
- [36] FAQ on Data Management Plans | Digital Curation Center [EB/OL]. [2018-10-25]. <http://www.dcc.ac.uk/resources/data-management-plans/faq-data-management-plans>.
- [37] 毕强, 朱亚玲. 元数据标准及其互操作研究[J]. 情报理论与实践, 2007 (5): 666-670.
- [38] HOLETSCHEK J, BERENDSOHN W G. The ABCD of primary biodiversity data access [J]. Giornale Botanico Italiano, 2012, 146 (4): 771-779.
- [39] Disciplinary Metadata [EB/OL]. [2018-09-28]. <http://www.dcc.ac.uk/drupal/resources/metadata-standards>.
- [40] WIECZOREK J, BLOOM D, GURALNICK R, et al. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard [J]. PloS One, 2012, 7 (1): e29715.
- [41] MARY P. File conversion market grows [J]. InfoWorld, 1985, 7 (31): 36-37.
- [42] GREENBERG J. Understanding metadata and metadata schemes [J]. Cataloging & Classification Quarterly, 2005, 40 (3/4): 17-36.
- [43] BERK F M. The role of mythology as a cultural identity and a cultural heritage: The case of phrygian myhtology [J]. Procedia Social and Behavioral Sciences, 2016, 225: 67-73.
- [44] JAISWAL D, SEHNAL D, LONESCU C M, et al. Motive validator: interactive web-based validation of ligand and residue structure in biomolecular complexes [J]. Nucleic Acids Research, 2014 (42): 227-233.
- [45] 王益兵. 都柏林核心元数据集的特点[J]. 兰台世界, 2005 (3): 12-13.
- [46] 杨波, 胡立耘. 用于社会科学信息组织的元数据标准——DDI [J]. 现代图书情报技术, 2005 (8): 7-11.
- [47] DEARBORN C C, BARTON A J, HARMEYER N A. The Purdue University research repository: HUBzero customization for dataset publication and digital preservation [J]. OCLC Systems & Services, 2014, 30 (1): 15-27.
- [48] 黄鑫, 邓仲华. 国外高校图书馆科学数据的元数据服务研究[J]. 图书与情报, 2017 (2): 84-90.
- [49] 殷沈琴, 张计龙, 窦方. 欧洲科学数据监护的标准与实践——UKDA案例研究[J]. 图书馆杂志, 2013, 32 (6): 76-80.
- [50] WHEELER Q D. What if GBIF? [J]. Bioscience, 2004, 54 (8): 718-718.
- [51] PDBx/mmCIF Dictionary Resources [EB/OL]. [2019-04-17]. <http://mmcif.wwpdb.org/>.
- [52] SUSSMAN J L, BERNSTEIN F C, JIANG J, et al. Protein Data Bank (PDB): A Database of 3D Structural Information of Biological Macromolecules [M]. Hoboken: John Wiley & Sons, Ltd, 2002.
- [53] 张旭. DOI在学术出版产业链中的应用研究[J]. 数字图书馆论坛, 2007 (10): 5-9.
- [54] 张海涛, 郑小惠, 张成昱. 数字图书馆的互操作性研究: Z39.50和OAI协议的比较[J]. 现代图书情报技术, 2003 (2): 13-15.

作者简介

汤子钰, 1994年生, 女, 硕士研究生, 研究方向: 数字图书馆。

欧石燕, 1971年生, 女, 博士, 教授, 博士生导师, 通信作者, 研究方向: 语义网技术与数字图书馆、自然语言处理与文本挖掘, E-mail: oushiyan@nju.edu.cn。

A Survey and Analysis of the Foreign Data Curation Platforms

TANG ZiYu OU ShiYan

(School of Information Management, Nanjing University, Nanjing 210023, China)

Abstract: Currently, various data curation platforms have been developed home and abroad so as to provide data services for scientific research. However, in comparison with some advanced countries, the development and construction of data curation platforms in our country starts relatively late with only a small number of platforms having been constructed, and thus lacks mature construction experience. Therefore, we selected 20 representative data curation platforms and carried out a thorough survey to investigate, compare and analyze various aspects of these platforms including data lifecycle models, technical specifications, components, software tools, functions, etc. On the basis of the survey results, we proposed a general framework for building data curation platforms, aiming to provide a reference for the development and construction of domestic platforms.

Keywords: Data Curation; Data Curation Platform; Survey; Data Curation Platform Framework

(收稿日期: 2019-03-25)