DOI:10.13530/j.cnki.jlis.140018

# 国外数据管护(Data Curation)研究与实践进展

王　芳　慎金花

**摘　要**　数据管护(Data curation)可以促进科学数据共享,提高科学研究的质量。随着 e'science 的发展,数据管护正在引起科学家、大学与研究机构以及图书馆、档案馆等信息资源管理机构的重视。本文在总结已有成果的基础上,提出一个细化的数据管护生命周期概念模型,包括六个阶段 14 个关键活动。以此为分析框架,对国外数据管护的研究、实践与教育进展进行全面梳理。研究发现:目前欧美国家在数据管护的投资立项、软件系统开发、新技术应用、数据质量评价以及教育与培训等方面取得了一定的进展,值得借鉴;同时也存在一些问题,如管护模型缺乏质量控制措施,大学与科研机构忽视数据政策的制定,人文学科与"小"学科对数据管护的认知不足,以及专业教育和职业培训严重欠缺等。图 1。参考文献 68。

**关键词**　数据管护　数字管护　数字保存　数字知识库　E-science　机构知识库

**分类号**　G273

# Advances in Data Curation Abroad: Research and Practice

Wang Fang & Shen Jinhua

**ABSTRACT**　Data curation can promote the sharing of scientific data and increase the quality of research. With the development of e-science, data curation has received more and more attention from scientists, institutions and organizations for information resources management, such as libraries and archives. On the basis of summing up existing studies, this paper formulated a detailed data curation lifecycle model that comprises 6 stages with 14 key actions. Under this model, global advances in research, practice and education of data curation were combed. It is worth learning that European countries and the United States have made progress in project approval and investment on data curation, development of curating systems and application of new technology, data quality evaluation, education and training. Meanwhile, problems existing in data curation were also identified, such as lack of measures for quality control in curation models, ignorance of policy-making in universities and research institutions, shortage of related knowledge in humanities and "small" disciplines and insufficient education and training, etc. 1 fig. 68 refs.

**KEY WORDS**　Data curation. Digital curation. Digital preservation. Digital repository. E-science. Institutional repository.

## 1　引言

人类已经进入前所未有的大数据时代,精密仪器和大规模计算的应用,使科学研究数据呈指数级增长态势。截至 2008 年,天文学家共有 40TB 数据,地震学家有 60TB 数据,人类基因组有 80TB 数据,美国大气研究中心(NCAR)有 4PB 数据[1]。在工程制造领域,随着计算机辅助设计(CAE)的发展,产生了大量二维或三维数据,运用传统的静态保存方法已经很难保证其信息质量[2]。而随着网络归档步伐的加快,

通讯作者:王芳,Email: wangfangnk@nankai.edu.cn

大量的音频、视频、动画、游戏数据也在不断积累[3]。但是，由于认知局限以及管理策略不当，大量数据正面临着丢失、不可读、信息损失、共享和复用困难等风险，尤其是不可重复的观测数据更是如此。1995 年美国国家科学院的一项研究指出，“大量联邦基金项目支持的珍贵科学数据从不归档，或除原始调查者以外，其他任何人无权访问……数据集最终可能会丢失。”美国航空航天局丢失了第一次登月的录音，最后在一个标着“坏磁带”的盒子里找到了，而第一个登月录像的正本，已经 30 年没有再出现过[1]。

为了使数字资源能够被不同领域的研究人员所理解和共享利用，需要对数据进行全生命周期管理[4]。对研究数据进行保存、管理、增值、复用以及跨学科共享，将会大大减少科学数据的重复采集，提高未来研究的质量。目前，关于科学数据管护（curation）的实践与理论研究已经取得了重要进展，成为近十年来备受 e-science、数字图书馆与数字档案馆学者关注的热点领域。为了能够对我国的实践和理论研究有所启示，本文拟从概念内涵、管护模型、管护过程、教育与人才培养、存在问题及未来展望等几个方面，对世界数据管护的最新研究与实践动态进行全面梳理。

## 2 数据管护的理论研究回顾

### 2.1 数据管护与数字管护的内涵

“Data curation”成为图书馆学、档案学与 e-science 核心研究领域的标志是出现了专门的数据管理机构，以及相关主题的国际会议和学术期刊。2004 年，英国国家级综合数据管理机构 DCC（Digital Curation Center）成立。2005 年 9 月，第一届国际“Digital curation”会议在英国巴斯大学召开。2006 年，由 DCC 和英国爱丁堡大学联合主办的开放存取专业刊物《国际数字管护期刊》（*International Journal of Digital Curation*）问世。目前，国内学者对 Data curation 起源与内涵的辨析已有很多[5-7]，本文不再赘述。Curation 有资源看护（take care of）之意[8]，Data curation 指贯穿数字资源生命周期的管理与维护，强调资料增值与主动管理[9]。为了与已有的概念如“管理”、“保存”、“维护”等相区别，本文将它译为“数据管护”。

从国外文献来看，目前使用最多的是“数字管护”（digital curation）与“数据管护”（data curation）两个概念。在生物、化学、气象、物理等具体的科学研究领域，多使用“数据管护”一词。美国伊利诺伊州立大学平原分校（UIUC）的图书情报学院在描述其“数据管护原理”课程时，将数据管护定义为：在学术研究、科学与教育活动中，主动、持续地贯穿数据生命周期的管理活动。认为管护活动与政策有助于数据的发现与检索，可以维护数据的质量，增加数据的价值并提供长期复用[10]。而在档案馆与图书馆领域，则多使用“数字管护”一词。英国数字管护中心 DCC 认为，数字管护是指贯穿数字化研究数据整个生命周期的维护、保存与增值活动，通过主动管理来降低科研数据过时与研究价值降低的风险[11]。从概念内涵来看，“数据管护”与“数字管护”并无太大差异。二者都指对数字化数据进行收集、注解、整理、保存以备当前或未来使用的实践活动[12]，主要包括五个概念区域：生命周期管理、创建者和管护者主动参与、鉴定与选择、获取、持久保存[13]。在北美地区开设相关课程的 16 所大学中，7 所学校的课程名称使用“数据管护”，其余的使用“数字管护”[14]。由于本文述及的文献大多来自具体的科学领域，因此主要采用“数据管护”这一概念。

与数据管护密切相关的概念还有数字保存（Digital Preservation）与信息管护（Information curation）。数字保存是指为保证数字对象在未来可被持续访问的主动管理过程[15]，是数字管护的一个具体环节。数字管护除了保存之意外，还强调资料增值与主动管理[9]3-4，以及研究者、出版者、管护者、资料管理者与使用者之间的紧密合作[4]。信息管护主要描述了在 web2.0 环境下社交网站用户的信息搜寻、偶遇、学习、

整合、包装、展示、共享与利用并实现增值的活动[16]。

### 2.2 国外数据管护的概念模型

为了有效地指导数据管护实践,一些数据管护机构、项目组或学者对数据管护的过程进行了概念化。其中应用最为广泛的是 DCC 的生命周期模型和国际标准 OAIS 参考模型。

(1)DCC 数据管护生命周期模型

DCC 旨在帮助英国高校研究社群发展数据管理能力,为他们提供专家咨询、管护培训和实际的合作帮助。在总结实践经验的基础上,DCC 提出了数据管护生命周期概念模型,包括概念化、创建或接收、鉴定与选择、采集、保存、存储、获取与利用、转化与迁移、保存规划、社区观察与参与、数据描述、信息表示等管护活动[11]。目前,采用 DCC 模型的项目有发育中的基因表达地图项目(DGEMap),英国的跨网格环境下移动环境传感系统项目(MESSAGE)等。

(2)开放档案信息系统(OAIS)参考模型

开放档案信息系统(OAIS)参考模型是国际性的数据管护标准,旨在为以长期保存为目的的信息系统建立参考模型和基本概念框架,包括环境模型、信息模型和功能模型[17]。OAIS 应用广泛,包括欧盟的电子资源保存与接入网络项目(ERPANET),美国佛罗里达数字档案项目(FDA)开发的数字资源库(DAITSS)[18],美国政府印刷办公室(GPO)开发的联邦数字系统(FDSYS)[19],欧盟的文化、艺术和科学知识的保存、获取和检索项目(CASPAR)等。

(3)牛津大学的机构数据管理基础设施模型

牛津大学的数据管理基础设施模型包括规划、数据创建、本地存储与检索、文件收集、机构存储、重新发现机制、检索机制以及贯穿管护活动始终的培训等环节[20]。

(4)美国数据保护项目(Data Conservancy)的概念框架

美国数据保护项目在实证研究的基础上提出了一个概念框架来表示科学数据实践(科学家的研究活动)、数据类型以及相关管护活动间的关系,具体包括知识表示、采集、系统管理、数据存储、政策、保存、链接、终端用户访问、可发现性等九项重要活动,涉及元数据标准的选择、数据标示符的应用、质量保证、完整性、知识产权、存储方法、迁移等内容[21]。

(5)美国加州大学嵌入式网络传感中心(CENS)的数据生命周期模型

美国加州大学洛杉矶分校的嵌入式网络传感中心(Center for Embedded Networked Sensing, CENS)针对收集的传感数据建立了一个数据生命周期模型,包括实验设计、(仪器)校准、数据捕获、清洗、计算与推导、集成、分析、成果发表、存储和保存等九个阶段[22]。

### 2.3 本文提出的一个细化的数据管护生命周期模型

上述五个模型虽然在数据对象和表述形式上各有不同,但所描述的大部分管护活动是相同的。部分现有管护框架缺乏数据质量标准和控制措施。一项针对美国七个预测毒理学数据源的研究表明,对数据来源标识和访问权限的规定还处于起步阶段[23]。

在总结上述模型的基础上,本文提出了一个细化的数据管护生命周期模型,以适用于多种类型数据的管护活动,包括六个阶段,共 14 个具体步骤:战略规划(成立管护工作小组、需求调查、制订管护战略规划),数据收集(数据采集、元数据管理、鉴定与选择),数据处理(数据表示与可视化、数据关联与集成、数据导入),数据保存(数据保存、数据存储),数据利用(数据挖掘与分析、数据获取与复用),服务质量评价。除此之外,还包括政策制定与人才培养两个外部支撑性因素。具体如图 1 所示。

## 3 贯穿数据生命周期的管护活动

依据图 1 所示的数据管护模型,本节从数据

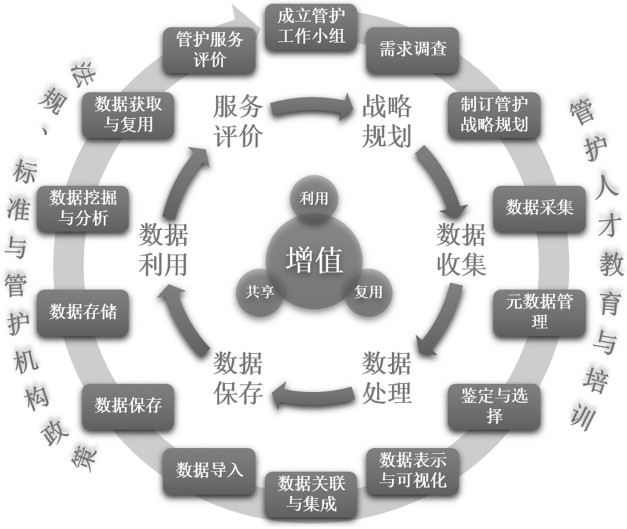


**图 1　细化的数据管护生命周期模型**

管护活动与教育培训两个方面展开述评。由于文献中关于数据管护政策的专门研究较少，部分讨论融入了具体的项目报告，本文暂不做专门评述。

### 3.1　数据管护过程

如图 1 所示，科学数据的具体管护过程可以划分为六个阶段 14 个具体步骤。

#### 3.1.1　战略规划阶段

（1）成立数据管护工作小组

一般来讲，国家级的管护计划由专业的数据管护机构如 DCC、国家或地区级的图书馆或档案馆实施。具体到机构层面，需要建立数据管护工作体系或工作小组。英国牛津大学的数据管护体系由牛津数字知识库领导小组（ODRSG）、研究服务办公室以及院系的研究服务小组、研究者本人、计算服务部、图书馆等部门组成。其中，ODRSG 由前任主管学术的副校长主持，负责校内管护项目的资助审批；研究服务办公室负责为研究项目的规划和申报提供支持；研究人员本人负责创建并描述数据；计算服务部为研究人员存储和检索数据提供集中的后台支持，负责维护元数据编辑系统以保证数据归档的标准化，为文件提供长期安全的存储，将 IT 主任办公室嵌入院系以保证一定层次的战略决策和沟通；图书馆基于 Fedora 数字资产管理系统开发了名为“数据银行”的数据知识库系统，提供元数据维护与资源发现和检索服务[20]。美国佐治亚理工学院（GT）图书馆于 2008 年夏天成立了数据管护工作组，由主管技术的副馆长、学术交流与数字服务部主任、数字服务人员、数字图书馆研发部主任和四名学科馆员组成[24]。

（2）进行需求调查

在进行数据管护活动之前，首先要进行需求调查。美国密歇根州立大学的数字管护项目团队为了解数据管护的现状和需求，于 2009 年对校内单位进行问卷调查和深度访谈[25]。Witt[26] 等人对高校教师进行访谈，发现他们希望数据管护工作能够提供数据的知识产权条款、元数据标准、规模、获取限制、数据挖掘或分析工具、互操作方式、数据的学术影响及保存政策等。Huang 与 Stvilia[27] 调查了 158 名基因科学家对数据质量的要求，发现数据的准确性、可获得性、完整性是最重要的维度，其次还有可信性、更新、可追踪性以及数据错误发现技术和数据挖掘技术等。

(3)制定战略规划

在需求调查的基础上,管护小组对数据管护的目标、策略与实施步骤进行规划,提出理论模型,以指导下一步的实践。欧盟委员会与瑞士政府合作的 ERPANET 项目提出了九大目标,包括识别、跟踪与过滤跨国、跨地区数据的来源,提供数字保存咨询服务,开展培训,研制指南与标准,鼓励软件开发等[28]。美国的长期生态研究项目(Long Term Ecological Research, LTER)的核心目标是理解多时空生态系统的长期演化模式[29-30]。加州数字图书馆(CDL)的数字保存项目(Digital Preservation Program of the CDL)[31]的目标是为利用、复用和改进数字资产提供高度可用的、响应的、全面的以及可持续的自动化管护服务。但是,在机构层面,还普遍忽视数字管护政策的制定。密歇根州立大学的调查发现,几乎所有校内单位都没有制定数字文件管护政策,很少有单位检查文件的完整性或创建元数据[25]。

3.1.2 数据收集阶段

(1)数据采集与抽取

数据收集是具体管护活动的起点,主要包括原始数据的采集与文献信息抽取两大类。实验、观测、调查等原始数据的采集工具有实验记录软件、网站系统、移动传感装置、观测仪器、问卷量表等。英国 CombeChem 项目支持的 SmartTea 课题开发出电子实验室笔记本软件(ELNs),用于记录实验数据[32]。CENS 中心利用嵌入式网络传感器收集生态感知数据[22]。LTER 项目以 26 个网站为基础,从不同生态系统收集大量分散的数据并从中抽取知识,除了观察数据,还收集与地方生态系统相关的实验数据[29]。而大量复杂、无序的远程医疗监测数据则是通过传感和移动技术,从大量患有中风,高血压,慢性阻塞如肺病、痴呆等疾病的病人那里收集起来的[33]。MESSAGE 利用多样、便宜、准泛在的传感器采集城市污染动态数据,并通过灵活的 e-science 基础设施对数据进行处理和传输[34]。

从文献中抽取特定信息建立专题数据库,可以节省研究人员检索和阅读文献的时间。泰国国家基因工程与生物技术中心(BIOTEC)开发了化学数据自动抽取软件 CHEMEX,通过四个主要模块(文件预处理、2D 化学结构图像识别、文本注释和信息查看器)抽取生物活性化合物、有机物和化验数据的文字和图像信息[35]。除此之外,还有免疫抗原决定基数据库与分析资源(Immune Epitope Database and Analysis Resource, IEDB)、发育中的基因表达地图项目(DGEMap)、蛋白质数据银行(Protein Data Bank)项目等。

(2)元数据捕获与管理

元数据管理是指将管理、描述、结构与技术存档的元数据进行科学配置,确定采用哪种级别的元数据,并提供互操作支持。对于实验科学而言,在元数据可以获得的时候进行采集远比靠记忆重新构建更为简单和便宜。英国 Southampton 的实验室知识库项目(R4L)在实验阶段就开始捕获元数据,除传统的 DC 数据项之外,还捕获分子名、InChI、光谱类型等元素,并用词表加以规范[34]。牛津大学的"研究项目的嵌入式机构数据管护服务"(The Embedding Institutional Data Curation Services in Research Project, Eidcsr)项目,其 3D 心脏项目采用的元数据由研究人员放入数据目录中,大部分可以映射到 DC,具有可扩展的用户定义和可产生 XML Read-me 文件的网页形式。元数据在归档过程中以及导入图书馆的数据银行系统时被解释,重新归档时会被更新[20]。元数据的互操作可以支持跨学科的数据共享[36]。目前,蛋白质数据银行(Protein Data Bank)、fMRIDC 和 Array Express 等项目提供多种工具支持元数据的互操作[37]。英国的网格嵌入式职业数据环境项目(Grid Enabled Occupational Data Environment, GEODE)采用基于 DDI 标准子集的 GEODE-M 元数据,允许原始数据提供者进行管护,支持跨案卷检索[38]。

(3)鉴定与选择

鉴定是判断数据的保存价值以确定数据保

管期限的活动。归档数据的选择一般由研究人员或领域专家来承担。英国爱丁堡大学主要由研究人员负责归档鉴定[39]。在免疫学领域，为了建立免疫抗原决定基数据库与分析资源，Vita等人使用61个关键词和逻辑运算符检索了1,600万条PubMed的引文数据，从中选择重要的文章，并根据管护手册和本体数据库评价所选文献是否符合要求[40]。在微生物领域，为帮助科研人员从不断扩增的酵母属微生物基因数据库中快速查找各种突变基因表型，美国斯坦福大学开发了一个由特定词汇库控制的突变基因表型信息获取管理系统，定期与所有酵母属微生物基因数据库相连接，利用由专家确定的词汇库自动或人工检索与基因表型相关的论文，然后分类存储供研究人员使用[41]。在工程领域，英国的不朽信息与终生知识管理（KIM）项目，针对工程中只存储正式的产品模型却忽略大量中间或背景性文件的缺陷，提出对中间或背景性文本文件进行结构化与形式化处理，并加以保存，以满足未来工程推理与决策审计的需要[2]。

3.1.3 数据处理阶段

（1）数据表示与可视化

一些特殊的科学领域会产生复杂的数据或图像，为了使复杂的数据或图像更易于理解，需要对它们进行可视化表示。2012年美国NSF投入1,000万美元给加州大学伯克利分校的算法、机器与人实验室（Algorithms, Machines and People Laboratory, AMP），并邀请艺术家协助创建可视化数据，探索如何以人类容易理解的形式表达从数据中推导出的结论[42]。在欧盟资助的发育中的基因表达地图项目（DGEMap）中，与胚胎相关的信息要被导入HDBR项目管理数据库中，其中大部分实验结果是来自显微镜的原始数字图像。为保证数据质量，要先对这些图像进行Photoshop处理，去除“脏”的颗粒。经过元数据捕获之后，图像会被映射到一个三维模型中，图像的信号数据（如图像着色）也被划分为高、中、低三个层次叠加到模型里。然后，采用模型表达域和相关的坐标信息创建一个本地数据库条目，来映射图像数据[43]。牛津大学的Eidcsr项目针对3D心脏项目产生的大型图像数据集，开发了一种可视化工具，只有在下载高分辨率图像时才可以根据要求放大观看，以节省访问时间[20]。

（2）数据关联与集成

在相互影响的数据之间建立关联，有助于分散数据的发现和获取。沈志宏和张晓林等发现关联数据机制能够很好地满足科学数据库对开放访问机制在包容性、适应性、语义支持以及易推广方面的要求[44]，其中实体RDF化是关联数据发布的六个关键步骤之一[45]。科学数据的关联一般发生在原始数据与元数据、数据与其它来源信息之间。目前建立数据关联的主要方法包括通用资源标识符URI、语义网技术如RDF记录、本体[46]、综合性的数据库条目以及集中的网络信息系统等。英国的实验室知识库项目（R4L）将原始数据与过程数据和用于快速识别的小型图像同时存储，其中，识别分子的关键元数据不仅包括化合物全称，还包括国际化学识别符IUPAC InCHI，这样就可以通过通用资源标识符（URI）将数据与相关信息联系起来[32]。ELNs项目对“关联”以及语义网资料的捕获有独到之处。比如，在化合物的捕获过程中，将原料、过程以及二者之间的关联用RDF清晰记录，从而加强了对实验推理和反应进程的记录[32]。LTER运用集中的网络信息系统促进相互独立的站点间的数据发现、集成与综合[30]。GEODE项目通过标准化的目录声明将不同索引变量链接起来，以弥补格式不一致产生的缝隙[38]。DGEMap项目的HDBR数据库通过条目内容来集成胚胎数据信息，包括执行实验的个人和实验室信息，实验和标本条件，使用的探针或抗体信息，空间映射信息，表达模式及其分布细节，3D模型的影片，以及任何相关的出版物和与其它数据库的链接信息[43]。

（3）数据导入

根据DCC的定义，数据导入（Ingest）是指依

据规章条例将需要长期保存的数字对象及其元数据传送到数据库中。在数据导入之前,应当采取必要的质量控制措施,包括去重、句法检查、词汇控制、格式转换、交叉注释、格式认证等。国际分子交换联盟中心(IMEx Central)在导入数据时,给每篇收录论文一个收录号,如果出现相同数据系统会发出警告。IMEx 实施交互质量控制措施,包括 PSI validator(检查句法和语义以及受控词汇的使用)、交叉管护(Cross-curation,所有参与数据库共同注释有争议的论文)等[47]。IEDB 项目组为了将文献中所包含的信息准确转换为数据库要求的结构化格式,专门设计了管护手册、数据字段以及相应的技术工具[40]。在结晶学领域,新的结晶结构在发表之前要求必须要用结晶信息文件格式(Crystallographic Information File,CIF)编码,由剑桥结晶数据中心(CCDC)存放并认证,然后导入剑桥结构数据库(CSD)[32]。

3.1.4 数据保存阶段

(1)数据保存

数字保存是指为保证数字对象在未来可被持续访问的主动管理过程[15]。出于长期利用的目的,需要对有价值的数据进行归档保存。美国 NCAR 要求所有社区地理系统模型(CESM)的观测数据需在档案库中保存一段时间。其中,研发数据不少于三年,综合产出数据不少于七年[48]。目前,越来越多的自然科学领域已经开始对研究数据进行长期保存,但是在人文与社会科学领域,情况很不乐观。牛津大学的调查发现,人文学科的结构化数据较少,数据格式多样,许多来自文献的汇编数据存在不完整、不一致、不可靠等问题,给数据管护带来困难[20]。

数字资源的长期保存在很大程度上依赖于格式的选择,支持数字对象长期存取的核心容器(container)的格式应包括七个属性[49]。目前适宜于数字资源长期保存的格式有 XML、PDF/A、ODF、MPEG、TIFF、JPEG2000 等。荷兰的数据归档与网络化服务项目(Data Archiving and Networked Services,DANS)将不同学科特定的数据格式转换为 XML 格式文件加以保存[50]。电子数据向中介 XML 迁移项目(Migration to Intermediate XML for Electronic Data,MIXED)开发了一个持久的文件格式转换存储库,将二进制文件格式转换为包含 XML 特征的标准数据保存格式 SDFP[50]。

(2)数据存储

数据存储(Storage)是指依据相关标准将数据安全存放。佐治亚理工学院从提供存储情况、额外存储服务和云保存服务三个方面检验研究项目对管护模型的实施情况[24]。英国爱丁堡大学的调查发现,80%的研究人员对文件存储的需求将达到 100 千兆字节,建议学校建立一个可集中存取的跨平台文件存储中心[39]。威尔士知识库网络项目(Welsh Repository Network,WRN)投资 1,400 万英镑用于高校存储库与数字内容基础设施的建设,其中包含 12 个机构知识库的物理实体和跨机构知识库的虚拟网络[51]。

目前,存储领域的前沿问题包括存储介质、存储软件、自动同步备份、迁移与仿真、云存储等。2009 年美国密歇根州立大学(MSU)的调查表明,大多数校内单位的数字内容存储在硬盘驱动器上,也有一些采用可移动介质与网络存储相结合的方式,有 23 个单位已经实施或计划实施内容管理系统或数字仓库软件[25]。DataUp 软件是一款针对数据云存储的中介软件,而 Rise4fun 则是用于从网络浏览器运行存储在云端的近 30 个软件工具[52]。

3.1.5 数据利用阶段

(1)数据挖掘与分析

数据挖掘与分析可以充分发挥数据的价值,实现数据增值。在生物学领域,从基因研究到流行病学研究,都有研究论文是通过分析现有数据而获得研究成果[1]。美国 NCAR 研究数据档案库的主要特色是对大气数据进行再分析,通过把多种类型的观测数据和最新的数据模拟系统结合起来,就可以获得一个 3D 时空的大气模拟状态[53]。

(2)数据获取与复用

为保证目标用户安全获取、共享所需数据，需要采取适当的数据发布方式。目前，最常使用的数据发布与共享平台是集成网站系统。GEODE 项目通过门户网站发布职业信息文件，非专业用户可以通过网站界面连接在线数据库，也可以使用 GridSphere 工具获得数据服务[38]。DGEMap 项目的 HDBR 数据库条目在经过审查许可之后，被上传到外部可见的数据库，供研究人员通过互联网浏览和搜索[43]。牛津大学 Sudamih 项目的“作为服务的数据库系统”（DaaS），也是通过网络界面来创建、编辑管理和查询数据集[20]。IMEx 的五个分子交互数据库通过共同的搜索网站实现数据共享[47]。

3.1.6 服务评价阶段

评价的目的是检查管护模型的实施效果，找出不足，加以改进。目前数据管护评价的对象包括数据质量、服务效果和管护工具与技术。一项美国预测毒理学的数据质量评价标准包括数据的准确性、完全性（completeness）、完整性（integrity）、元数据管理、可用性、授权六个维度[23]。DIRECT 是一款数字图书馆管理工具，用于对在搜索引擎系统的大规模评价活动中产生的科学数据进行管理，特别关注数据质量的以下方面：①数据文件的可执行性（DOI 标引，采用 XML、JAVA 及 HTTP 支持松散耦合的编辑系统）；②数据的兼容性（采用 HTTP、URI、XML、AJAX 等标准）；③数据的合法性；④版权/许可；⑤系统（开发云环境下基于组件的系统）；⑥数据的规模；⑦来源（通过系统日志功能跟踪来源事件）等[54]。加州数字图书馆的数字保存项目采取以数字对象为中心的管护和以服务为中心的管护两种策略。对象中心的评价标准包括标识符、生存能力、稳定性、真实性、本体、可查找性、效用、可移动性、鉴定与时效性。用户服务中心的评价标准包括可获得性、响应度、安全性、互操作性、可扩展性、可信赖性与可持续性[31]。英国的 CASPAR 数字保存项目提出了普适性的评价体系，对数字保存工具与技术的有效性进行检验，并通过所谓的“加速寿命”测试提供评价示例[55]。

### 3.2 数据管护专业教育与职业培训

如前所述，一些大学图书馆已经设立了专门的数据管护岗位，而像 DCC 这样的专门数据管护机构更需要专业人才。数据管护人员（又称数据科学家）需要承担比传统图书馆员更为复杂的角色，包括研究者、图书馆员、早期的技术使用者和政策制定者[56]。生物学领域的研究表明数据管护专家应当具备以下知识和能力：数据归档和保存，数据仓库及工具，元数据标准，生物本体，工作流捕获，数据合成，基于文献的发现以及版权和知识产权知识等[57]。然而，作为一个新兴领域，数据管护常常被等同于创建一个“档案备份”[12]，相应的人才培养还没有得到足够重视。“目前尚没有图书馆或者信息科学课程以及大学证书项目为这种复杂的工作准备好人才”[8]。Harris-Pierce 与 Liu 研究了 55 所北美地区的大学，发现到 2012 年只有 16 所学校在研究生层次开设了数据管护课程[14]。除此之外，英国[39,58]与新西兰[59]的一些大学或政府机构、美国航空航天局[60]、美国国家档案和文件管理局[61]等都开设了数据管护课程或培训项目，所涉层次包括硕士、研究生后、博士、博士后以及职业培训。课程名称各不相同，涉及以下内容或其中的一部分：数据管护基础、数字保存、鉴定与选择、归档、元数据、馆藏发展、信息组织与获取、信息系统/平台/软件、数据类型/标准/生命周期、数据库、数据质量、数据分析、开放存取、信息环境与信息社会、来源、项目管理、案例分析、实习、毕业设计等[56,62]。

除了专业的数据管护人员之外，科研人员也需要掌握一定的数据管护知识。因此，有必要对科研人员进行数据管理意识、知识和技能培训[63]。但是对武汉 11 所高校的问卷调查结果显示，国内科研人员的科学数据管理意识薄弱，很多科研人员没有接受过数据管理培训，缺乏相关知识[64]。尤其是一些人文学科对数据备份和存储、版本控制、文件同步保存等认识不

足,对于可共享的数据服务器利用率很低[20]。英国数据档案馆专门为社会科学领域的学者提供培训服务[59]。密歇根州立大学的数字资产管护小组则建议通过网上论坛和会议促进"实践社区"的形成,帮助校园单位和其他机构分享数字管护经验,促进相互合作[25]。

## 4 未来展望以及对我国的启示

从各国的研究与实践情况来看,目前数据管护已经取得了一定的进展,但是也存在如下问题:①现有的管护框架缺乏数据质量标准和控制措施;②科研机构普遍忽视数据管护政策的制定;③人文学科及"小"学科的数据管护存在困难;④专业教育和职业培训严重不足。展望未来,科学数据管护呈现出如下发展趋势:

(1)分级管护体系逐渐形成

从各国实践来看,目前数字管护项目或机构可以分为国际性、国家级、地区级及机构级几个层次。国家级或跨地区的管护项目或机构通过政策引导、标准制定、多方合作、软件开发、教育培训、学术交流等方式,促进数据管护的发展。如美国的可持续数据保存与获取网络伙伴计划(Sustainable Digital Data Preservation and Access Network Partners, DataNet),它不资助具体学科的数据库建设,而是投资 1 亿美元建立了一批示范性的新型数据保存与共享伙伴[65]。

与国际性或国家级的"高"级别管护项目相比,机构知识库的数据管护级别较"低",主要关注机构内部的数据管护。相对来讲,高级别的管护投入成本更大,专家在数据导入阶段就开始介入元数据的设置;而低级别的管护则意味着更大程度的自动化和最少的人工干预。例如,牛津大学的机构数据管护采用比较简单的自动化程序,主要由研究者自己负责归档数据的鉴定和特殊元数据的设置[20]。

(2)存储外包与云技术应用渐成趋势

机构知识库的倡导者认为数字化保存功能应当由第三方外包合作机构承担[66]。目前爱丁堡大学图书馆[39]、佐治亚理工学院图书馆[24]都在考虑将存储外包给商业云运营商。云存储可能带来信任问题,还需要进一步研究。

(3)数据科学家将成为新兴职业

随着大学与研究机构数据管护的发展,对专业数据管护人才的需求将大幅上升,设立数据科学家和高层次的数据管护职位成为必须。这也意味着,数据管护的专业教育和职业培训将有一个新的发展空间。

目前,我国机构层面的数据管护取得了一定的进展。中科院已经建成了机构知识库服务网络平台,集成 100 多家所级机构知识库系统,对各研究所的知识资产进行统一采集、集中展示和长期保存,并提供全院科研成果的一站式检索和发现服务。除此之外,截至 2012 年 12 月 18 日,中国的 CALIS 机构知识库登记成员达 29 家,元数据总量达 79,039 条[67]。截至 2013 年 5 月 10 日,中国的机构知识库联盟成员数达 92 家,数据总量达 478,975 条[68]。但是总体看来,目前,我国国家级的数字管护中心与跨行业、跨机构的数字管护项目发展还比较缓慢,尤其是专业领域的数据知识库建设与共享相对滞后,在数字管护的理论探索、软件开发、政策制定、标准研制、质量评价、人才培养等方面,与国际相比还有较大差距。为此,我国需要借鉴国际的先进经验,重视政策引导与人才培养,在国家、地区与机构层面建立起多层级的数据管护中心或知识库平台,支持跨地区、跨行业的项目合作,积极利用移动传感、数据挖掘与云存储技术促进科研数据的管护、共享与利用,进而推动我国 e-science 的发展。

致谢:

南开大学商学院信息资源管理系 2012 级硕士研究生闫红霄、吴格、杨鸿芳、许璐、白艳华、富靓参与了文献资料的收集与初步翻译工作,博士生翟羽佳承担了图 1 的制作工作,在此致谢!

## 参考文献

[1] Lesk M. Recycling information: science through data mining [J].International Journal of Digital Curation, 2008, 3(1):154-157.

[2] Ball A, Patel M, McMahon C,et al. A grand challenge: immortal information and through-life knowledge management (KIM) [J].International Journal of Digital Curation, 2006, 1(1): 53-59.

[3] 王芳,史海燕. 国外 Web Archive 研究与实践进展[J].中国图书馆学报,2013,39(2):36-45. (Wang Fang, Shi Haiyan. Progress of foreign research and practices in web archive[J].Journal of Library Science in China, 2013, 39 (2): 36-45.)

[4] Beagrie N. Digital curation for science, digital libraries, and individuals [J].International Journal of Digital Curation, 2006, 1(1):3-16.

[5] 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, 29(2):18-21. (Yang Heli. Data curation: a new development of university libraries in the U. S. [J].Journal of Academic Libraries, 2011, 29(2): 18-21.)

[6] 崔宇红. E-Science 环境中研究图书馆的新角色:科学数据管理 [J].图书馆杂志,2012,31(10):20-23. (Cui Yuhong. The new role of academic libraries in e-science:scientific data curation [J].Library Journal, 2012, 31(10):20-23.)

[7] 张计龙,朱勤,殷沈琴. 美国社会科学数据的共享与服务[J].大学图书馆学报,2013(5):13-17. (Zhang Jilong, Zhu Qin, Yin Shenqin. Social scientific data sharing and services in the United States [J].Journal of Academic Libraries, 2013(5): 13-17.)

[8] Kouper I. CLIR/dLF Digital curation postdoctoral fellowship:the hybrid role of data curator [J].Bulletin of the American Society for Information Science and Technology, 2013, 39(2): 46-47.

[9] Harvey R. Digital curation: a how-to-do-it manual [M].New York: Neal-Schuman Publishers, 2010.

[10] University of Illinois: full catalog [EB/OL]. 2012 [2013-11-24]. http: www. lis. illinois. edu/academics/courses/catalog#500level.

[11] DCC. What is digital curation? [EB/OL].[2013-08-11].http://www. dcc. ac. uk/digital-curation/what-digital-curation.

[12] Heidorn P B, Tobbo H R, Choudhury G S, et al. Identifying best practices and skills for workforce development in data curation [C]//Proceedings of the American Society for Information Science and Technology, 2007, 44(1): 1-3.

[13] Elizabeth Y. Digital curation [J].OCLC Systems & Services, 2007, 23(4):335-340.

[14] Harris-Pierce R L,Liu Y Q. Is data curation education at library and information science schools in North America adequate?[J].New Library World, 2012, 113(11/12):598-613.

[15] Semple N, Beagrie N, Williams P, et al. Digital preservation policies study: part 1(final report)[R/OL].2008 [2013-11-24]. http://www. jisc. ac. uk/media/documents/programmes/preservation/jiscpolicy_p1finalreport. pdf.

[16] Jacobson J. Information curation[C]//Proceedings of the American Society for Information Science and Technology, 2012, 49(1):1-2.

[17] Laughton P. OAIS functional model conformance test: a proposed measurement [J].Program: electronic library and information systems, 2012, 46(3):308-320.

[18] Caplan P. DAITSS, an OAIS-based preservation repository[C]//Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop, ACM, 2010,17.

[19] LaPlant L, Zwaard K. A holistic approach for establishing content authenticity and maintaining content integrity in a large OAIS repository [C]//Archiving Conference on Society for Imaging Science and Technology, 2008, 1:

109-113.

[20] Wilson J A J, Martinez-Uribe L, Fraser M A, et al. An institutional approach to developing research data management infrastructure [J].International Journal of Digital Curation, 2011, 6(2):274-287.

[21] Cragin M H, Palmer C L, Chao T C. Relating data practices, types, and curation functions: an empirically derived framework [C]//ASIST 2010, October 22 - 27, PA, USA.

[22] Wallis J C. Moving archival practices upstream: an exploration of the life cycle of ecological sensing data in collaborative field research [J].International Journal of Digital Curation, 2008, 3(1):114-126.

[23] Fu X, Wojak A, Neagu D, et al. Data governance in predictive toxicology: a review [J].Journal of Cheminformatics, 2011, 3(1):24. doi: 10.1186/1758-2946-3-24.

[24] Walters T O. Data curation program development in U.S. universities: the Georgia Institute of Technology example [J].International Journal of Digital Curation, 2009, 3(4):83-92.

[25] Schmidt L, Ghering C, Nicholson S. Notes on operations digital curation planning at Michigan State University [J].LRTS, 2010, 55(2):101-115.

[26] Witt M, Carlson D, Brandt D S, et al. Constructing data curation profiles [J].International Journal of Digital Curation, 2009,4(3):93-103.

[27] Huang H, Stvilia B. Roles and perceived priorities for data quality dimensions and skills in genome curation work [C]//Proceedings of ASIST, 2012, OCT, 26~30, MD USA.

[28] Ross S. ERPANET: a European platform for enabling digital preservation[J].Journal of Information and Knowledge Management Systems, 2004 ,34(2 ):77-83.

[29] Karasti H. Digital data practices and the long term ecological research program growing global [J].International Journal of Digital Curation, 2008, 3(2):42-58.

[30] Michener W K, Porter J, Servilla M, et al. Long term ecological research and information management [J].Ecological Informatics, 2011(6):13-24.

[31] Abrams S, Cruse P, Kunze J. Preservation is not a place [J].International Journal of Digital Curation, 2009, 4(1):8-21.

[32] Frey J. Curation of laboratory experimental data as part of the overall data lifecycle [J].International Journal of Digital Curation, 2008, 3(1):44-62.

[33] Ure J, Hanley J. Curating complex, dynamic and distributed data: telehealth as a laboratory for strategy[J].International Journal of Digital Curation, 2011, 6(2):128-145.

[34] Donnelly M. The milieu and the MESSAGE: talking to researchers about data curation issues in a large and diverse e-science project [J].International Journal of Digital Curation, 2011,6(1):32-44.

[35] Tharatipyakul A, Numnark S, Wichadakul D, et al. ChemEx: information extraction system forchemical data curation [C]//Eleventh International Conference on Bioinformatics, Bangkok, Thailand. 3-5 October, 2012[2013-11-16].http://link.springer.com/article/10.1186%2F1471-2105-13-S17-S9#page-1.

[36] 王芳,王小丽. 基于 OAI 协议的数字档案馆元数据互操作问题研究[J].现代图书情报技术,2007(3):18-24.(Wang Fang, Wang Xiaoli. On OAI-PMH based interoperation of digital archival metadata [J].New Technology of Library and Information Service, 2007(3):18-24.)

[37] Macdonald A. Digital archiving, curation and corporate objectives in pharmaceuticals [J].Journal of Medical Marketing, 2006, 6(2): 115 - 118.

[38] Lambert P, Gayle V, Tan L, et al. Data curation standards and social science occupational information resources [J].International Journal of Digital Curation, 2007, 2(1):73-91.

[39] Rice R, Haywood J. Research data management initiatives at University of Edinburgh [J].International Journal of

Digital Curation, 2011, 2(6):232-244.

[40] Vita R, Vaughan K, Zarebski1 L, et al. Curation of complex, context-dependent immunological data [EB/OL]. BMC Bioinformatics, July, 2006[2013-12-25]. http://www. biomedcentral. com/1471-2105/7/341.

[41] Costanzo1 M C, Skrzypek M S, Nash R, et al. New mutant phenotype data curation systemin the Saccharomyces Genome Database[J]. The Journal of Biological Database and Curation, 2009. DOI: 10. 1093/database/bap001.

[42] Yang S. Big grant for big data: NSF awards $10 million to harness vast quantities of data. UC Berkeley News Center [EB/OL]. [2014-01-01]. https://newscenter. berkeley. edu/2012/03/29/nsf-big-data-grant.

[43] O'Donoghue J, van Hemert J I. Using the DCC lifecycle model to curate a gene expression database: a case study [J]. International Journal of Digital Curation, 2009, 3(4):57-70.

[44] 沈志宏,张晓林,黎建辉. OpenCSDB:关联数据在科学数据库中的应用研究[J].中国图书馆学报, 2012, 38(5):17-26. (Shen Zhihong, Zhang Xiaolin, Li Jianhui. Open CSDB: Application of linked data in scientific database[J]. Journal of Library Science in China, 2012, 38(5):17-26.)

[45] 沈志宏,刘筱敏,郭学兵,等. 关联数据发布流程与关键问题研究——以科技文献、科学数据发布为例[J]. 中国图书馆学报, 2013, 39(2):53-62. (Shen Zhihong, Liu Xiaomin, Guo Xuebing, et al. A research on publishing workflow and key issues of linked data: experience with publishing scientific literature and scientific data as linked data[J]. Journal of Library Science in China, 2013, 39(2):53-62.)

[46] Gelernter J. Use of ontologies for data integration and curation [J]. International Journal of Digital Curation, 2011, 6(1):70-78.

[47] Orchard S, Kerrien S, Abbani S, et al. Protein interaction data curation: the international molecular exchange (IMEx) consortium [J]. Nature Methods, 2012, 9(4): 345-350.

[48] Strand G. Community earth system model data management policies and challenges [J]. ICCS, volume 4 of Procedia Computer Science:558-566.

[49] Kim Y. Digital forensics formats: seeking a digital preservation storage container format for web archiving[J]. International Journal of Digital Curation, 2012, 7(2):21-39.

[50] van Horik R, Roorda D. Migration to intermediate XML for electronic Data (MIXED): repository of durable file format conversions [J]. International Journal of Digital Curation, 2011, 6(2):245-252.

[51] Knowles J. Collaboration nation: the building of the Welsh Repository Network Program [J]. Electronic Library and Information Systems, 2010, 44(2): 98-108.

[52] Bishop J. Industry's role in data and software curation in the cloud [J]. Journal of Systems and Software, 2013, 86(9): 2327-2329.

[53] Jacobs C A, Worley S J. Data curation in climate and weather: transforming our ability to improve predictions through global knowledge sharing [J]. International Journal of Digital Curation, 2009, 4(2):68-79.

[54] Ferroa N, Hanburyb A, M¨ullerc H, et al. Harnessing the scientific data produced by the experimental evaluation of search engines and information access systems[J]. Procedia Computer Science, 2011(4): 740-749.

[55] Giaretta D. The CASPAR approach to digital preservation. International Journal of Digital Curation, 2007, 2(1): 112-121.

[56] Joint N. Data preservation, the new science and the practitioner librarian [J]. Library Review, 2007, 56(6): 451-455.

[57] Palmer C L, Heidorn P B, Wright D, et al. Graduate curriculum for biological information specialists: a key to integration of scale in biology [J]. International Journal of Digital Curation, 2007, 2(2):31-39.

[58] Day M. Report from the DigCCurr 2007 International Symposium on Digital Curation [C]//Chapel Hill, NC, April 18-20, 2007. International Journal of Digital Curation, 2007, 2(1): 102-111.

[59] Franks P C, Oliver G C. Experiential learning and international collaboration opportunities: virtual internships [J].Emerald Group Publishing, 2012, 6(4): 272-285.

[60] Renea A H, Dolan M, Trainor K, et al. Towards a cross-disciplinary notion of data level in data curation [C]// Proceedings of the American Society for Information Science and Technology, 2009, 46(1):1-8.

[61] Ball A, Day M. Report from the Digital Curation Curriculum Symposium (DigCCurr) 2009 [C]//International Journal of Digital Curation, 2009, 4(1):138-150.

[62] Harvey R, Bastian J A. Out of the classroom and into the laboratory: teaching digital curation virtually and experientially [J].International Federation of Library Associations and Institutions, 2012, 38(1): 25-34.

[63] 李晓辉. 图书馆科研数据管理与服务模式探讨[J].中国图书馆学报,2011, 37(5):46-52. (Li Xiaohui. Research data management and service pattern in libraries [J].Journal of Library Science in China, 2011, 37(5): 46-52.)

[64] 胡永生,刘颖. 基于用户调查的高校科学数据管理需求分析[J].图书情报工作,2013, 57(6):28-32,78. (Hu Yongsheng,Liu Ying. Demand analysis on scientific data management in universities based on user survey [J].Library and Information Service, 2013,57(6):28-32,78.)

[65] Sandusky R J, Palmer C L, Allard S, et al. The DataNet partners: sharing science, linking domains, curating data [C]//Proceedings of the American Society for Information Science and Technology, 2009, 46(1): 1-8.

[66] Day M. Toward distributed infrastructures for digital preservation: the roles of collaboration and trust [J].International Journal of Digital Curation, 2008, 1(3):15-28.

[67] CALIS 机构知识库[EB/OL].[2014-02-10].http://ir.calis.edu.cn/index. (CALIS Institutional Repository [EB/OL].[2014-02-10].http://ir.calis.edu.cn/index.)

[68] 中国机构知识库联盟[EB/OL].[2014-02-10].http://www.cspace.org.cn. (China IR Alliance[EB/OL].[2014-02-10].http://www.cspace.org.cn.)

**王 芳** 南开大学商学院信息资源管理系教授,博士生导师。
通讯地址:天津市南开区卫津路94号。邮编:300071。

**慎金花** 同济大学图书馆馆长,教授。通讯地址:上海市同济大学图书馆。邮编:200292。

(收稿日期:2014-02-19;修回日期:2014-04-01)