数据监护现状分析及对我国农业科学数据监护的启示*

赵华，朱亮，鲜国建，赵瑞雪
（中国农业科学院农业信息研究所，北京 100081）

摘要：在数据密集型科学研究新形势下，科学数据作为重要的科技资源需要得到有效地管理，数据监护应运而生。本文在分析国内外科学数据监护发展现状的基础上，针对我国农业科学数据管理与共享现状，从政策制度、实施策略、流程管理、平台建设和培训教育等方面为我国农业领域科学数据监护提出5点建议，以期为农业科研机构开展数据监护实践提供借鉴。

关键词：科学数据；数据监护；农业；生命周期
中图分类号：G250
DOI：10.3772/j.issn.1673-2286.2017.11.002

在E-Sicence环境下，数据密集型科学发现成为科学研究的第四范式，这种研究范式的一个显著特征是以数据考察为基础，即从科学数据中发现理论与知识，使科学数据成为科学发现的核心，其重要性已被广大科研人员认同，科学数据与科技文献成为科研活动不可或缺的支撑材料，更是学术交流的基本单元。随着科研活动不断深入，科学数据量不断增长，数据的管理与共享在各学科领域备受关注。各国政府、各级科研机构均将科学数据作为一种重要资源，通过建立专门的数据共享中心或委托图书馆等部门进行存储与管理。国内外学术界围绕科学数据的管理与共享开展多项研究，其中基于数据的产生、收集、整理、发布和利用等全生命周期的数据监护问题成为重要的研究主题。国外多个科研机构开展了数据监护方面的研究与实践，我国也逐步开展了数据监护的探索。在农业领域，随着科研活动的持续深化与拓展，大量宝贵的数据资源不断产生，有部分数据伴随学术论文或数据论文的发表得到有效管理，但也有部分数据尘封在科研人员的电脑中，随时面临丢失的危险。为促进农业科学数据资源的保存和再利用，有必要开展数据监护实践。为此，本文在分析国内外数据监护发展现状的基础上，针对我国农业科学数据管理与共享现状，为我国农业领域科学数据监护的实施提出建议。

1 数据监护概念

数据监护最早由微软首席研究员、计算机图灵奖获得者Gray于2002年提出，是E-Science环境下科学数据共享和大规模科学计算的产物。英国数字监护中心认为，数据监护指贯穿整个数字化科学数据生命周期的维护、保存与增值活动，通过主动管理来减少科学数据过时与研究价值降低的风险，目标在于使这些数据能够满足当前和未来的研究需要，维持数字资源的长期可生存能力、可呈现能力和可理解能力[1]。数据监护可以使科学数据增值，数据监护活动包含评价、筛选、重现及组织，最终服务于数据获取和使用[2]。杨鹤林认为数据监护是为了确保数据当前使用目的，并能用于未来再发现及再利用，从数据产生时就对其进行管理和完善的活动。对动态数据集而言，数据监护意味着需要对科学数据进行持续性补充和更新，以使其符

* 本研究得到公益性科研院所基本科研业务费项目（编号：JBYW-AII-2017-06）、中国农业科学院科技创新工程项目（编号：CAAS-ASTIP-2017-AII）和国家科技基础条件平台专项“农业科学数据共享中心”（编号：2005DKA31800）资助。

合用户需求[3]。英国联合信息系统委员会对数据监护及其相关概念进行区分（见表1）[4]，由此可见，数据监护含义更广泛，涵盖数据监护、数据归档、数据保存等活动。数据监护是基于数据生命周期而开展的一系列数据管理活动的集合，涉及管理策略、方法和技术等，其目的是促进数据发现、提升数据检索效率、维护数据质量、增加数据价值，在不同的生命周期阶段数据监护侧重点不同。

表 1 数据监护相关概念

概 念	含 义
数据监护	从数据产生开始即对其进行管理，以促进数据利用，确保数据能在需要时被发现和再利用。对于动态数据集，需要持续性补充和更新，深度的数据监护还应包括对数据的标注及提供数据和相关发表资料的链接
数据归档	数据监护的子任务，指对数据合理的选择和存储，确保数据物理上和逻辑上的持续完整性，注重数据可获取性、安全性和可靠性等方面，从内容层面确保数据的可用性
数据保存	数据监护的子任务，对数据对象进行持续维护，确保随技术的更新与变革，数据对象仍能被读取和理解，从技术层面确保数据的可持续性

2 国内外数据监护发展现状

2.1 国外数据监护发展现状

国外高等院校及科研机构十分重视数据监护并开展多个方面研究。普渡大学图书馆、伊利诺伊大学图书馆与信息科学研究生院共同主持Data Curation Profiles项目，探讨科学数据的具体监护内容、方式，及培养数据监护方面的应用人才等问题[5]，并开发了Data Curation Profiles工具包。康奈尔大学图书馆针对本校学者的数据服务需求，以机构库为基础构建数据阶段型存储库[6]，为学术界共享“小科学”数据集提供暂时、过渡性的存储节点，力图构建一个能提供完善服务方案的数据监护平台。牛津大学建立数据监护模型（也称为数据管理基础设施模型），该模型包括规划、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，以及贯穿数据监护活动始终的培训等环节[7]。约翰霍普金斯大学在2009年启动了数据保护项目（Data Conservancy），对天文、地球、生物和人文社会科学等学科的数据进行一系列基于科研工作流程的数据监护活动，为科研人员提交数据、获取数据，以及图书馆馆员管理数据等提供技术支持，提高用户数据归档、元数据生成、数据存储等多个环节的效率[8]。数据监护最典型的案例是英国数字监护中心（Digital Curation Centre，DCC）[9]，DCC创建了数据监护生命周期模型，在数字科学数据全生命周期内对其价值进行维护、保存和附加，把数据监护过程划分为描述和表示信息、建立存储计划、对科研组织活动的观察和参与、数据监护和保存四个阶段。DCC在历时9年的探索中，在数据监护的模型构建、职业教育、环境调研、开发工具和政策引导方面取得很大成效，为数据监护理论与方法的发展奠定良好基础。

数据监护的实施离不开技术工具的支持，目前国际上普遍盛行的数据平台开源工具有Dspace[10]、Fedora[11]和Dataverse[12]。Dspace是麻省理工学院与惠普公司合作开发的数字资产管理系统，具有对知识资源的收集、保存、发布等功能。Fedora是由康奈尔大学提出设计方案，在美国国家科学基金会和美国国防部资助下开发的管理系统，主要解决内容管理、资产管理和资源保存等方面的问题。另外，哈佛-麻省理工数据中心于2007年构建了开源软件Dataverse，能够对科学数据进行发布、引用、存储、发现和在线分析，方便研究人员管理和传播。

综上所述，国外对数据监护的研究比较系统，理论与实践紧密结合，数据监护的发展相对成熟，无论在理论层面，还是在实践层面都取得一定成效。特别是针对数据监护模型的构建提出多种思路，给出数据监护活动所包含的内容，不同的模型对数据监护活动的解释存在差异（见表2）。这些数据监护模型有繁有简，为数据监护的实施提供丰富的理论依据。除数据监护模型外，国外科研机构还开发了数据监护平台及相关工具，对数据监护所涉及的各方面因素进行分析，认为数据监护是集政策环境、理论研究、技术工具和教育培训为一体的系统工程。

2.2 国内数据监护发展现状

近年来，我国开始逐步重视数据监护的发展，但

表 2 国外典型的数据监护模型

机 构	数据监护模型	监护活动内容
英国数字监护中心	数据监护生命周期模型	概念化、创建或接收、选择、采集、保存、存储、获取与利用、转化与迁移、保存与规划、社区观察与参与、数据描述、信息表示
牛津大学	数据管理基础设施模型	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、培训
康奈尔大学	数据阶段型存储库模型	数据收集、数据上传、元数据自动生成、元数据与数据集存储、数据发布
约翰霍普金斯大学	数据保护模型	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

多数停留在理论研究的层面，主要包括数据监护概念分析[13]、国外数据监护经验介绍与比较分析[14]，以及将数据监护引入机构库建设[15]。在分析机构库和数据监护关系的基础上[16]，对高校图书馆机构库理论层面和实践层面的数据监护进行阐述，将数据监护的管理思想与方法作为机构知识库提升服务水平与创新能力的捷径与突破口。王芳等在总结已有研究成果的基础上，对数据监护模型开展研究，提出数据监护生命周期模型[17]；徐坤通过数据监护理论和相关技术，聚焦医学科学数据，构建基于本体的高校科学数据监护平台[18]；复旦大学构建了基于Dataverse的社会科学数据共享平台，该平台主要面向复旦大学的研究机构、项目、课题组、研究者个人等数据管理的需求，具有数据提交、保存、共享、发现、交换、传播等功能，其中数据监护功能主要体现在用户权限、数据访问限制、数据更新和长期保存等方面[19]；北京大学图书馆在“开放数据”理念下，为满足校内科研团队的迫切需求，联合国家自然科学基金——北京大学管理科学数据中心，以Dataverse为基础开发了北京大学开放科学数据平台，为研究者提供数据管理、发布、存储和使用追踪服务，为用户提供数据浏览、检索和下载等服务[20]。

相较国外，我国数据监护以理论研究为主，集中于概念解析和对国外数据监护研究与实践的梳理层面，数据监护的实践还处于起步阶段；在研究机构方面，同样以高校图书馆为主体；在研究思路上，将数据监护与机构知识库的发展相联系，以数据监护方法与理论促进机构知识库建设，这些是我国数据监护研究与发展的特色。

3 农业科学数据管理与共享现状及问题

农业科学数据是农业科技信息数据的重要组成部分，不仅是重要的战略资源，也是农业科技创新的重要基础，需要进行有效地管理，以便更好地开发利用[21]。为促进农业科学数据的管理与共享，我国已经搭建多种层次的数据共享平台。在国家层面，作为科技部主导的国家科学数据共享工程试点领域之一的农业科学数据共享中心，对分散在全国各机构、各部门的农业数据资源进行整合，构建数据共享管理与服务体系，面向社会提供服务。由于农业学科的综合性特点，农业科学数据分散在不同的子领域或研究主题，因此除国家农业科学数据共享平台外，还建设有农业子领域的数据共享平台，如国家农作物种质资源平台、国家水产种质资源平台、国家水稻数据中心、家养动物种质资源平台和中国饲料数据库等。这些不同的数据管理与共享平台为我国农业科学数据共享提供基础条件。此外，除国家级平台外，还有地方农业科研单位建设的各类信息共享平台，以及从事与农业学科有交叉的相关研究部门建立的信息共享平台也提供农业科学数据的管理与共享服务。由此可见，我国在农业领域对科学数据资源的共享与再利用已经迈出一大步，国家农业科学数据中心构建了由“数据主中心-数据分中心-数据节点”三个层级组成的数据整合工作体系，在数据组织与管理方面做了大量工作。

基于上述现状可以发现，我国农业领域十分重视科学数据管理，数据共享服务也已初具规模。尽管如此，我国农业科学数据资源建设、数据管理与服务工作仍然存在问题。数据资源由于得不到有效地监护，导致数据流失的现象依然存在，且数据共享与再利用的程度与用户预期还存在差距。在总结国家农业科学数据中心多年数据管理与服务经验的基础上发现我国农业科学数据管理主要存在三方面问题。（1）数据收集方面。现有平台多关注国家计划和大型科研项目所产生的数据，而小型项目所产生的科学数据，由于缺乏政策和制度的约束没有进行有效地管理，仍分散在科研人

员手中。此外，科研人员数据管理经验存在不足，仅关注于分析结果数据，对原始数据、中间环节数据重视不够。（2）数据加工与质量控制方面。数据因缺乏深加工导致利用率低下，加之已有数据资源质量不高，数据后期维护与更新力度不足，导致数据缺乏鲜活性，不能满足用户需求。（3）数据汇交方面[22]。数据开放共享的程度不足，数据管理规范化程度有待提升，数据汇交机制有待进一步完善，缺乏有效便捷的数据管理与提交工具。这些问题的解决有赖于数据监护的实施，需要将数据监护理论与方法引入我国农业科学数据管理实践，使数据资源得到有效保存的同时提升数据管理效率和数据再利用价值。

4 我国农业领域实施科学数据监护的建议

数据监护是一项复杂的系统工程，在借鉴国外数据监护先进经验的基础上，结合我国农业领域科学数据管理与共享存在的问题，本文对实施数据监护提出以下建议。

4.1 完善数据监护政策与制度，鼓励科研人员参与数据监护

政策制度是数据监护系统的第一要素，是数据监护的有力保障[23]。尽管农业领域科研人员在数据意识和数据素养上得到很大提升，但主动参与数据监护工作的意愿还不够强烈，政策制度方面的约束和规范，成为数据监护的前提和保障。农业科研管理部门或项目资助部门在制定科研管理政策时，可以考虑纳入数据监护方面的规定，可以在对数据的访问限制、知识产权、数据共享等方面作出相应的规定，同时鼓励科研工作者积极参与数据监护活动。承担数据监护任务的部门（如研究型图书馆）需要全面充分地了解各科研机构的意愿和诉求，制定农业科学数据监护的相关规范和政策，对数据组织规范与策略、数据存储与安全管理、数据重用、数据归档与长期保存等进行详细规定，为各机构制定数据管理计划、参与数据监护活动提供具体指南。

4.2 确定数据监护发展策略，完善数据汇交机制

农业科学数据的监护工作仅靠单独机构难以完成，需要各层级的农业科研机构开展跨机构、跨部门的协同合作。数据监护的发展策略可以采取自下而上的方式，各农业科研机构应循序渐进，稳步推进数据监护工作，在强化科学数据资源管理的同时，对数据监护的启动与发展做好规划。除对大型科研项目产生的科学数据进行监护外，对一些小规模科研项目产生的数据更应给予重视，注重对小学科、小项目所产生的科学数据的监护工作。在具体实施层面，基层农业科研机构通过构建机构数据阶段性存储库的方式实施数据监护，在明确数据可以对外发布共享时，可以向比较成熟的数据仓储汇交数据；国家级农业科研机构则以构建大型数据仓储为主要目标，汇聚农业领域全面的数据资源，着眼于我国农业科学数据的长期保存服务。只有不断完善各级机构间的数据汇交机制，提升数据管理规范化水平，营造数据开放共享的良好氛围，才能全面推进我国农业科学数据监护向前发展。

4.3 规范数据监护流程，提升数据资源质量

数据监护的流程源自数据生命周期理论，数据生命周期的实质是在科研过程中管理数据，包括数据产生、加工、发布、再利用等循环过程，其中数据产生、收集、处理、发布与利用属于核心阶段[24]。数据监护活动需要围绕数据生命周期进行，农业科学数据的监护也不例外，需要密切关注农业科研活动的进展，在明确数据监护任务的基础上，规范数据监护流程，才能保证数据监护的成功实施。数据监护通常包括数据收集、数据评价、数据组织与处理、数据存储和数据再利用等阶段[25]。数据收集和数据评价是数据监护工作的基础环节，组织数据、处理数据和描述数据是数据监护过程中使数据增值的必要环节，数据存储和数据再利用是数据监护的终极目的。数据收集阶段通过与科研人员沟通筛选需要保存与管理的数据，确定数据收集的范围以及所收集数据所包含的附件；评价数据阶段主要是按照一定的标准对数据文件的完整性作出评价；数据组织与处理阶段需要对数据进行加工处理和必要的描述，提升数据与其他资源的关联度，提高数据发现和再利用效率；数据存储和利用阶段属于数据监护的尾声，在数据进入仓储后，需建立数据长期访问机制，提升数据的检索能力，跟踪数据影响力和用户反馈，对原始数据和各版本的再利用数据建立链接，实现对数据的动态管理。

4.4 开发数据监护软件工具，提升数据监护效率

数据监护的实施离不开强大的技术支撑，目前国外已经出现多个数据管理平台及开源软件（如Dspace、Fedora、Dataverse和HUBzero等），这些工具在数据监护过程中发挥重要作用。我国农业领域开展数据监护时，同样需要软件工具的支持。为减少重复开发，可针对数据管理软件进行调研，在已有的开源软件基础上，结合农业科学数据监护平台的功能需求进行二次开发，从而构建集数据收集、保存、发布、标记、注释、引用跟踪、发现和在线分析等功能为一体的软件工具，尽量减少数据监护活动中的人工因素，简化工作流程，从而提升数据监护效率。

4.5 构建数据监护社群，开展数据监护培训

数据监护是一项复杂的工作，需要来自图书馆、科研机构和科研人员的多方合作，尤其是科研人员间的合作。以图书情报相关专业为背景的数据馆员在信息管理与存储、信息增值等方面具有丰富经验，在理论知识和实践经验上已具备做好数据监护工作的条件。但对科研人员而言，数据监护是新鲜事物。因此在农业科学数据监护实施过程中，承担数据监护任务的图书馆馆员（或数据馆员）需要主动贴近用户，调研用户需求，开展各种讲座，宣传和推广数据监护的益处和相关知识，以构建数据监护社群，为数据监护的开展营造良好氛围。另外，关于农业科研人员欠缺数据监护必需的数据计划、技术标准、数据编目、元数据标准和数据保存管理等方面的基础知识，以及需要不断跟进的最新数据科学理论知识，则需靠系统的数据监护培训来解决。培训形式包括参加高级研究进修课程、交流论坛、网络交流等，还可以通过与国内外高等院校合作的途径，邀请高校名师到机构进行授课，为数据监护人员输送最新的数据科学前沿知识。

5 结语

数据监护为科学数据的有效管理与共享提供了理论支撑。国外在数据监护理论研究与实践方面已经走在前列，我国也开始逐渐重视数据监护的研究与发展。本文针对我国科学数据管理与共享现状，在借鉴国外数据监护先进经验的基础上，为我国农业领域实施科学数据监护，提出完善数据监护政策与制度，鼓励科研人员参与数据监护；确定数据监护发展策略，完善数据汇交机制；规范数据监护流程，提升数据资源质量；开发数据监护软件工具、提升数据监护效率，构建数据监护社群，开展数据监护培训等建议，希望能为农业科研机构开展数据监护研究与实践提供借鉴。相信未来随着数据监护研究和实践的不断深入，我国农业科学数据的管理与再利用定会迈上一个新台阶。

参考文献

[1] DCC.What is digital curation?[EB/OL].[2017-03-11].http://www.dcc.ac.uk/digital-curation/what-digital-curation.

[2] SHREEVESS L,CRAGIN M H.Introduction:institutional repositories: current state and future[J].Library Trends,2008,57(2):89-97.

[3] 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.

[4] LORD P,MACDONALD A.Data Curation for e-science in the UK:an audit to establish requirements for future curation and provision[EB/OL].[2017-10-08].http://www.jisc.ac.uk/media/documents/programmes/preservation/e-sciencereportfinal.pdf.

[5] WITT M,CARLSON J R,CRAGIN M H,et al.Constructing data curation profiles[J].International Journal of Digital Curation,2009,4(3):1-11.

[6] DataStar[EB/OL].[2017-05-10].http://DataStar.mannlib.cornell.edu/.

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

[8] 杨志伟,卫军朝.基于Data Curation的机构库建设研究——以约翰霍普金斯大学Data Conservancy项目为例[J].图书馆学研究,2016(7):55-61.

[9] HIGGINSS.The DCC curation lifecycle model[J].2008,3(1):453-453.

[10] SMITH M K,BARTON M,BASS M,et al.DSpace:an open source dynamic digital repository[J].D-Lib Magazine,2003,9(1):1-10.

[11] Fedora.About Fedora[EB/OL].[2017-10-11].http://Fedora-commons.org/about.

[12] Dataverse Project.The history[EB/OL].[2017-10-15].https://dataverse.org/about.

[13] 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,2012,30(2):23-28.

[14] 李文文,陈雅.国内外Data Curation研究综述[J].情报资料工作,2013(5):

35-38.

[15] 王文联.嵌入数据监护的图书馆机构库高效运行模式[J].图书馆学刊,2013(12):39-41.

[16] 宋秀芬,邓仲华.基于数据监护的机构知识库研究[J].图书馆学研究,2016(2):44-48.

[17] 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.

[18] 徐坤.基于本体的科学数据监护平台研究[D].长春:吉林大学,2014.

[19] 张计龙,殷沈琴,张用,等.社会科学数据的共享与服务——以复旦大学社会科学数据共享平台为例[J].大学图书馆学报,2015,33(1):74-79.

[20] 朱玲,聂华,崔海媛,等.北京大学开放研究数据平台建设:探索与实践[J].图书情报工作,2016,60(4):44-51.

[21] 孟宪学.国家农业科学数据中心的设计与建设研究[J].农业图书情报学刊,2004,16(12):5-8.

[22] 赵瑞雪.农业科学数据共享中数据汇交与管理研究[J].科技管理研究,2009(8):284-286.

[23] 高芹,钟晓莉.高校图书馆科学数据监护平台构建研究[J].图书馆学刊,2015(8):113-116.

[24] GOLD A.Data curationand libraries:short-term developments,long-term prospects[EB/OL].[2010-04-04].[2017-10-10].http://works.bepress.com/agold01/9/.

[25] 宋秀芬,邓仲华,金勇.高校图书馆数据监护的流程管理研究[J].图书馆学研究,2015(5):35-40.

作者简介

赵华，女，1980年生，博士研究生，助理研究员，研究方向：科学数据管理，E-mail：zhaohua02@caas.cn。
朱亮，男，1981年生，博士，副研究馆员，研究方向：文献计量、情报分析、科学数据建设与管理研究，E-mail：zhuliang@caas.net.cn。
鲜国建，男，1982年生，博士，副研究馆员，研究方向：知识组织、关联数据，E-mail：xianguojian@caas.cn。
赵瑞雪，女，1968年生，博士，研究员，博士生导师，研究方向：信息管理与信息系统、信息资源管理、知识组织与数字图书馆，E-mail：zhaoruixue@caas.cn。

Status Analysis of Scientific Data Curation and Its Inspiration to Chinese Agricultural Scientific Data Curation

ZHAO Hua, ZHU Liang, XIAN GuoJian, ZHAO RuiXue
(Agricultural Information Institute of CAAS, Beijing 100081, China)

Abstract: Due to data-intensive scientific research, scientific data as an important scientific and technological resource need to be managed effectively, so data curation emerges as the times require. Based on the status analysis of scientific data curation at home and abroad, according to the present situation of agricultural scientific data sharing and management, the paper puts forward some suggestions about the policy system, implementation strategy, process management, platform construction and personnel training for agricultural scientific data curation, in order to provide reference for data curation practice of agricultural scientific research institutions.

Keywords: Scientific Data; Data Curation; Agriculture; Life Cycle

（收稿日期：2017-10-31）

《数字图书馆论坛》在2016年度复印报刊资料转载指数排名中喜获佳绩

由中国人民大学人文社会科学学术成果评价研究中心联合书报资料中心研制的2016年度复印报刊资料转载指数排名于2017年3月28日正式发布。

在“图书馆、情报与档案管理学科期刊”全文转载排名中，《数字图书馆论坛》转载率位列第15名，综合指数位列第20名。

该排名根据人大复印报刊资料近100种学术系列期刊在2015年度转载的学术论文数据，从转载量、转载率、综合指数三个维度对中国人文社科期刊和教学科研机构进行统计形成。