农业科技大数据仓储建设与服务*

赵华[1,2] 赵瑞雪[1,2] 金慧敏[1,2] 郑建华[1,2] 鲜国建[1,2] 朱亮[1,2] 寇远涛[1,2]
(1. 中国农业科学院农业信息研究所，北京 100081；2. 农业农村部农业大数据重点实验室，北京 100081)

摘要：大数据时代，国家农业战略制定、科研创新、成果转化等都离不开农业科技数据资源的支撑。本文概述国内外农业领域大数据发展现状，分析农业科技大数据内容构成，提出农业科技大数据仓储框架，从数据来源、数据汇聚与整合、数据存储、数据管理与组织等方面分析农业科技大数据建设与管理问题，最后对农业科技大数据仓储应用与服务进行总结分析，以期为我国农业领域大数据研究与实践提供借鉴。

关键词：农业科技；大数据；数据仓储；数据服务
中图分类号：G250　　DOI：10.3772/j.issn.1673-2286.2020.08.008
引用格式：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.

在数据密集型科研范式时代，大数据引起政府、产业界、学术界的广泛关注，成为各行业各领域发展重点。2015年农业农村部印发《关于深化农业科技体制机制改革加快实施创新驱动发展战略的若干意见》，提出加强农业科技大数据平台建设，完善设施设备，统一数据接口和标准，强化数据积累，加强数据分析，为农业科技创新提出长期连续、全面翔实的基础数据[1]。同年底，农业农村部发布《关于推进农业农村大数据发展的实施意见》，全面部署农业农村大数据发展工作[2]。由此可见，农业领域对大数据的发展与应用有着迫切的需求。农业科技大数据作为农业大数据的一部分，与农业科技活动有着密切的关系，是农业科技活动中产生的、长期积累的各种信息的集合，是农业领域快速发现新知识、创造新价值、提升新能力的国家基础性战略资源，在农业领域科技研究、产业发展、管理决策等方面发挥重要的作用，其服务对象包含政府部门、科研机构、涉农企业、农业科技工作者、种植养殖户等。目前，农业科技信息快速增长、海量分散，“信息污染”和知识获取困难等问题不断凸显，用户对农业科技数据的需求难以得到满足，为此开展农业科技大数据的建设与应用方面的研究和实践显得尤为迫切，是我国农业信息服务机构迎接大数据时代必须要迈出的一步，也是为提升我国农业科技信息服务水平积累经验。

1 国内外农业领域大数据仓储建设现状

国际上农业领域大数据资源建设涉及科技文献、科学数据和统计数据等。国际农业和生物科学中心建设的CABI文摘数据库、美国农业图书馆组织建设的AGRICOLA数据库、联合国粮农组织建设的AGRIS数据库是典型的农业科技文献数据库，美国国家生物技术信息中心的GeneBank、澳大利亚的世界牧草属数据库和英国的世界草业数据库等都是农业专业领域典型的科学数据仓储库。联合国粮农组织建设的FAO统计数据库[3]、美国农业部数据中心[4]等都属于统计数据类仓储。除国际组织和政府部门建设的农业领域大数据仓储外，国外一些大型企业开展农业生产大数据建设，如孟山都公司（Monsanto）建设的Climate Pro或Field Scripts，主要服务于农业种植决策和精准生产。

国内在农业领域大数据仓储建设方面与国外类似，建设的资源包含科技文献、科学数据、市场数据等。前两者主要由农业科研机构主导建设，包括中国农

*本研究得到中国农业科学院科技创新工程项目（编号：CAAS-ASTIP-2016-AII）、中国工程科技知识中心建设项目“农业专业知识服务系统”（编号：CKCEST-2020-1-20）、国家科技基础条件平台项目“国家农业科学数据中心”（编号：NASDC2020XM12）资助。

业科技文献数据库、国家农业科学数据中心[5]、农业基础性长期性监测数据[6]等，以及专业领域的科学数据库（如全国农业资源区划基础数据库、外来入侵物种数据库[7]、中国饲料数据库等）。农业农村部联合各地政府部门主导建设的农产品市场价格数据库，收集了全国各地的大宗农作物重点农产品的生产情况、批发价格等方面的数据。除了政府和科研机构外，一些电商企业依托日常开展的业务，关注于农业产业链下游（流通、消费等）相关数据资源的建设。

由此可见，农业领域大数据发展迅速，围绕科技文献、科学数据、统计数据、市场数据等建设了不同类型的数据资源仓储库或数据平台，但已有的数据仓储主题相对单一，资源类型、来源相对集中。与其他领域相比，农业领域多类型、多来源科技数据资源的整合还有待加强。在医学领域，针对海量的健康医疗数据的管理与利用问题建设了医疗大数据资源仓储系统，按照资源类型整合了不同来源的临床、管理和科研数据等资源，为保障医疗数据的互联互通，消除“信息孤岛”，提供基础支撑[8]。在农业领域，农业科技数据资源种类多、来源广，同样需要建设一个综合性的、可服务于农业科技的大数据仓储系统，以进一步全面整合与汇聚农业领域多来源的科技数据资源，为农业科技创新、管理决策提供有效的资源保障。

2 农业科技大数据内容框架

农业科技大数据与农业科技活动密切相关，一切服务于农业科技活动，以及因农业科技活动而产生的信息、数据资源都属于农业科技大数据的范畴。其中包括了种类丰富的数据、信息和知识，统一归纳，即形成多源异构的农业科技大数据。除传统的电子图书、电子期刊、专利数据、科学数据等外，随着互联网开放获取运动的开展，通过网络采集获取的各类科技信息资源使得农业科技大数据的种类更加丰富。

农业科技大数据中资源类型多样、结构复杂，包含文本数据、数值数据、图片数据、声音数据和图像数据等，按照资源的内容可以分为文献类、数据类、政策资讯类及事实工具类资源等。其中文献类资源通常包括农业领域图书、科技期刊文献、科技报告和专利等。数据类资源主要包括科学数据和统计数据，科学数据是指农业科技活动中产生的实验数据、监测数据、分析数据，以及形成的延续性的数据产品等；统计数据主要指来自国内外权威机构的对外公开的农业相关统计数据，该类数据资源是在已有各类统计资料的基础上对数据进行重新整合，从而形成具有时间序列属性的数据。科技政策和新闻资讯类资源通常是指通过网络化信息采集方式，采集互联网上开放的涉农政策和新闻资讯等信息，经过加工标引等环节，形成规范的科技政策和新闻资讯类资源。事实工具类资源主要是经过规范加工整理的农业领域专家库、农业科研机构库和农业科研项目库等事实类资源，以及农业叙词表、农业百科等工具类资源。农业科技大数据内容框架见图1。

3 农业科技大数据仓储建设

数据仓储是面向主题的、集成的、稳定的、随时间变化的数据集合。数据仓储可实现对数字资源的提交和收集、描述、发布、索引、检索、互操作等，同时兼具数据存储、管理与保存等功能。随着农业科技数据资源的急剧增长，需要建设一个开放、统一的数据仓储对数据资源进行规范管理，将会对数据资源的整合与揭示发挥重要作用，并为最终的数据共享与应用提供支撑。

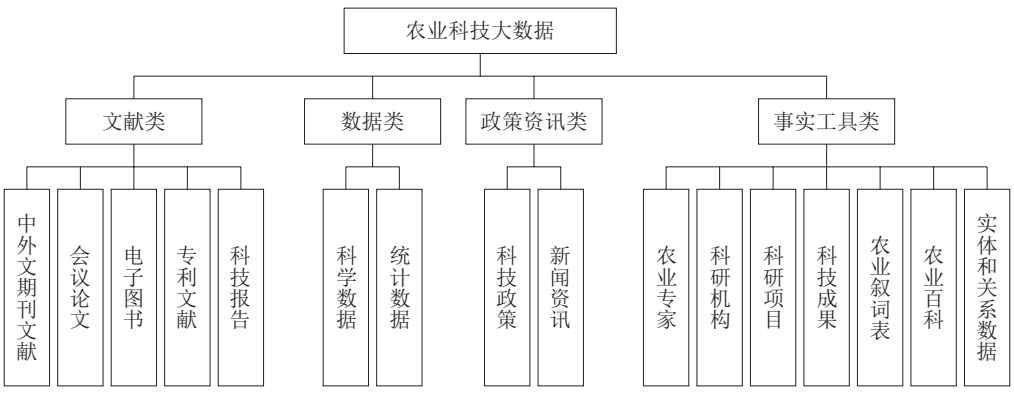

图1 农业科技大数据内容框架

农业科技大数据仓储的建设将有效解决在网络环境下对数字资源保存、访问和管理等方面的问题，除实现对数据资源保存与管理外，还面向各种基于仓储资源的数据应用与服务提供数据接口。

3.1 农业科技大数据仓储框架

农业科技大数据仓储总体框架包含数据来源、数据汇聚与整合、数据存储、数据管理、应用与服务5个层次（见图2）。数据来源层反映了农业科技大数据多来源、结构复杂的特性。数据汇聚与整合层体现了农业科技大数据各类资源采集、加工、处理和整合，运用ETL技术实现不同来源数据的清洗、抽取、转换和融合，完成数据从数据源向数据仓储转化的过程。在数据存储方面，采用SOL数据库（关系数据库技术）和NoSQL数据库相结合的方式，实现了非结构化和半结构化数据的海量存储，并提供数据检索。在数据管理方面，主要围绕元数据和主数据的管理与维护，针对不同层面的用户，提供相应的数据管理功能，还包括涵盖数据全生命周期的数据质量管理功能。在对数据资源进行分类整合、关联组织的基础上，数据仓储面向各类平台系统、用户提供不同形式的数据支撑服务，包含数据查询与检索、数据分析、数据接口服务等。

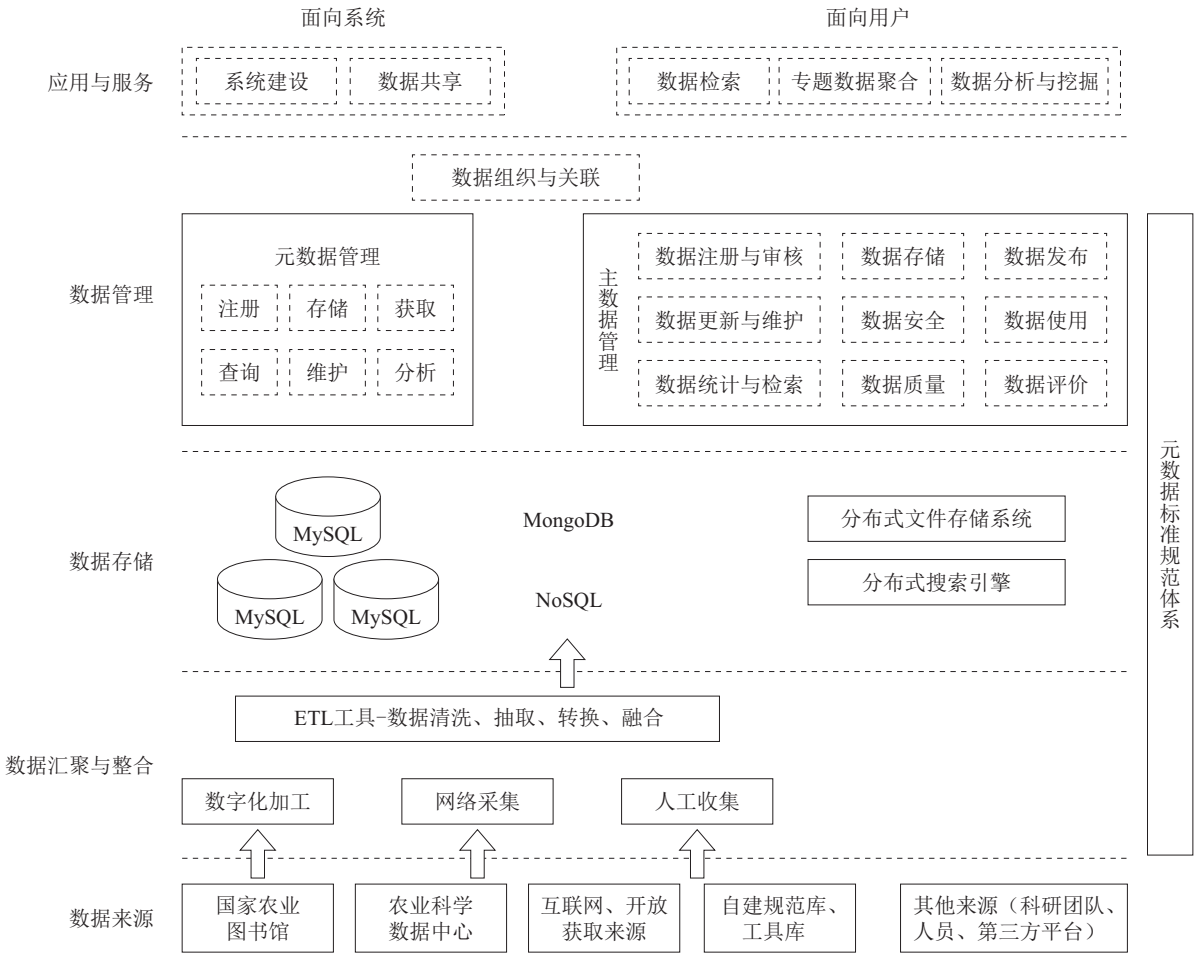

图2 农业科技大数据仓储框架

3.2 数据来源

农业科技大数据涉及多种类型的资源，建设方式主要包括自主加工、网络采集、开放获取资源收割、众包协同共建、采购等。自主加工主要是针对文献资源，通过对来源于国家农业图书馆馆藏文献的数字化加工，形成庞大的文献数据。此外，部分事实工具类资源，如专家规范库、机构规范库、叙词表、本体等，在长期

研究与实践中形成的数据资源丰富了农业科技大数据的工具类资源。网络采集和收割的资源主要来源于互联网，结合农业科技大数据建设需求，在对互联网资源进行分析的基础上，建设了农业科技大数据数据源库，构建了由国内外近500个农业领域的网络站点组成的监测源，主要涉及农业相关的国际重要组织、咨询机构、联盟、学会协会、科技管理部门、政府管理部门、科研机构、科研资助机构、大学、科技企业、期刊、新闻网站、重大项目、科技政策研究机构、研究计划、会议以及其他类信息源。协同共建主要是针对科学数据资源，通过与国家农业科学数据中心、部分科研团队、研究机构等开展合作，选择优质的科学数据资源进行加工处理，汇聚到农业科技大数据中。此外，随着数据开放获取的快速发展，农业科技大数据拥有更加丰富的数据来源，如DOAJ（开放存取期刊目录）和一些开放获取仓储等。综合来看，农业科技大数据的来源主要包括国家农业图书馆、农业科学数据中心、互联网和开放获取来源、研究机构、科研团队、第三方数据平台等。

3.3 元数据标准

为实现农业科技大数据多类型数据资源建设的规范化和标准化，研究团队在广泛借鉴和参考《都柏林核心元数据元素集》《NSTL文献资源加工规范》等现有国内外典型元数据标准规范的基础上，制定了农业科技大数据元数据标准规范体系[9]，根据农业科技大数据各类资源特点，提出了各类资源的元数据集合，为实现资源的描述、管理、检索、交互、关联组织提供工具。元数据标准规范体系内的元数据包含描述元数据和通用元数据两大类。描述元数据主要是确定描述各类数据资源的内容结构以及针对每项描述内容的描述细则。通用元数据指各类资源描述元数据中共性的内容，通用元数据的制定是为了简化描述元数据内容，可被描述元数据所复用，通用元数据通常包含主题、责任者、责任机构、国别（地区）等。元数据标准规范的制定不仅规范了农业科技大数据资源建设、提升数据质量，而且有效地解决了农业科技大数据资源描述问题，既向用户揭示大数据内容，也为各类数据资源有效组织打下了基础。

3.4 数据汇聚与整合

农业科技大数据汇聚与整合主要解决数据采集、数据整合两个层面的问题。数据采集以网络自动采集为主，人工采集为辅。针对网络采集的数据资源，需遵循准确性、全面性与连续性相结合的采集原则，在分析采集来源的特征基础上，制定科学的采集策略，通常采用完整性采集与选择性采集相结合的策略，尽可能保证数据采集的广度；同时还须兼顾数据采集的深度，对采集到的数据进行严格的质量审查，经过数据清洗、去噪等环节，确保数据的完整性、一致性、准确性、合规性等。通过分布式集群采集策略，依据各类数据源数据更新情况进行定时和实时数据采集，确保数据的时效性。

在数据整合层面，农业科技大数据中整合的数据资源包含自主加工的一手数字文献数据，也包含通过采集、整理等手段获得的来自其他来源的二手数据，还包含由科研人员提供的科学数据等。不同来源的数据结构存在很大的差异，有结构化数据、半结构化数据和完全非结构化数据。针对来源不同、结构不同的数据资源，需要对数据资源进行融合处理，离不开技术层面的支持。农业科技大数据在处理时，运用ETL技术对不同来源的异构数据（如XML、数据库文件、Excel、原始PDF等）进行处理，完成数据从源头向目标数据仓库的转化。主要通过典型的ETL工具Kettle对采集自各个来源的各类数据进行去重、字段映射、拆分、标引等加工处理操作，完成数据进入仓储系统存储前的所有准备工作。在数据融合方面，主要解决数据内容层面的融合，针对自主加工的文献数据和采集自开放共享获取平台的文献数据开展数据融合实践，实现期刊论文元数据的精准匹配、查重去重和内容的查漏补缺，使文献资源的规模和质量得到提升。

3.5 数据存储

在数据存储方面，农业科技大数据中的结构化、关系型数据采用传统的关系数据库SQL进行存储，建立了MySQL结构化关系数据存储集群。在非结构化数据和半结构化数据存储方面，鉴于MongoDB在所有非关系型数据库中功能最丰富、最接近关系数据库，用于存储类型复杂的数据时具有明显优势[10]，而且在开发和运维上成本低、实用性强，因此农业科技大数据仓储采用MongoDB存储半结构化及非结构化数据。通过对农业科技大数据中的非结构化数据资源进行分组，按照集合存储，却不用对集合的模式进行定义，存储

在MongoDB中的数据库文件无需定义文件结构，直接存储XML格式文件，不仅有利于计算机读取数据，更方便了非结构化数据资源的存储，较好地解决了海量数据的存储和访问，并且极大地满足了高并发的读写请求。此外，仓储系统针对大量的图数据采用Virtuoso数据库存储。由此农业科技大数据仓储形成MySQL集群结构化存储、MongoDB非结构化存储和Virtuoso图存储相结合的多模态混合存储模式。除了解决海量数据的存储问题外，数据仓储采用开源搜索引擎Solar与分布式搜索引擎Elastic Search相结合的检索技术，实现各类数据的结构化搜索和全文搜索，通过采用RESTful的架构风格，向用户提供数据查询和数据共享接口。

3.6 数据管理与组织

3.6.1 数据管理

数据仓储属于一种在网络环境下提供对数字对象保存、访问和管理的系统，数据存储仅是数据仓储最基本的功能，仓储的其他功能还包括数字资源的提交和收集、描述、发布、索引、检索、互操作等，在数据资源管理方面发挥着重要作用。农业科技大数据仓储系统提供的数据管理功能包括元数据管理和主数据管理两大部分，管理的内容因管理对象不同而不同，针对元数据的管理主要涉及元数据注册登记、存储与维护、查询与分析等方面；针对主数据的管理涉及范围更广，包括数据采集、数据加工与维护、数据存储、数据安全与备份、数据检索、数据发布、数据评价、数据分析及应用等方面，基本涵盖了数据生命周期各个阶段，还包含了贯穿数据生命周期全过程的数据质量管理，并提出数据权益管理方案，设置了数据权益管理功能模块。此外，仓储系统还具备向用户提供数据开放接口服务、资源调度接口等功能，同时对仓储各类数据资源的加工量、更新量、使用情况等进行动态跟踪与监控，提供关于数据资源的各种统计与分析功能。

仓储的数据质量管理重点解决产生于数据源和ETL过程的质量问题。针对由数据源产生的数据质量问题，通过优化数据来源、定期清洗历史数据、及时补充缺失数据、修正错误数据、清除冗余数据等手段减少源于数据源的数据质量问题。针对在数据抽取、数据转换、数据加载等ETL过程中产生的数据质量问题，运用脏数据预处理、排序邻居方法、优先排队算法、多次遍历数据清理方法等多个集成数据清理算法来解决脏数据清洗问题[11]，通过设计重复记录识别算法实现重复数据的自动识别与判定，对数据重复程度进行比对后，进行字段映射与补充，实现数据内容层面的互补与融合，完成数据去重。

农业科技大数据仓储为了实现对不同数量、类型、来源、深度的数据使用和再利用，提升数据仓储的资源服务能力，提出针对多源异构数据的专业数字集成、深度知识标引与检索利用过程中的数据权益管理方案，专门设置了数据权益管理功能模块，在数据收集阶段，启动对各类数据源发布的数据政策、许可协议或者与数据提供方签订的数据建设合同等登记，并对数据权益进行分类，做好权益信息管理[12]，针对不同来源的数据做好权益标记，在数据应用与服务时审核权益信息，确保数据的合理、合规使用。

3.6.2 数据组织

信息资源的组织需结合资源的使用和用户的需求，传统的信息组织方式主要包括分类法和主题法，主要处理结构化数据，为用户提供资源导航，方便查找和发现数据。随着数据类型的多样化，半结构化和非结构化数据的出现，数据组织方式扩展到了多种形式，包括自动分类、语义网络、本体、知识图谱等[13]，以能够向用户揭示资源更丰富、更深层次的信息为目标，除了资源自身的信息，更多地关注于资源间的关系、联系等，满足用户多个层次的信息需求。在大数据时代，需要更高效的数据组织方式来应对处理多样数据格式、支持数据实时动态更新和挖掘分析、便于信息整合等新形势下的数据组织需求。农业科技大数据面对多样的数据类型、复杂的数据结构，简单的分类法和主题法已经远远不能满足其数据组织的需求。为了使大数据成为一个有机的整体，而不是资源的散乱堆砌，农业科技大数据仓储在实现对资源进行分类整合的基础上，对每类资源进行主题概念和学科分类的自动标引，主题标引依据的是经过了网络化适应性改造的《农业科学叙词表》，改造后的词表中的叙词以及词间的语义关系的规范化描述可实现向本体批量转化[14]，不仅提升了农业科技大数据各类资源的标引工作的规范性，更为实现数据的关联提供了支撑。学科分类标引是基于已经构建且经过实践检验的农业科技资源分类导航体系中

的基础语料对资源进行标引。农业科技大数据还尝试了不同类型资源的关联组织，在自建规范库的基础上，如机构规范库、专家规范库、期刊规范库，对农业科技大数据中各类资源进行实体标注，并通过构建科研本体描述框架和科研本体实例库[15]，实现各类数据资源的关联整合，使类型多样、内容丰富的农业科技大数据成为一个有机的整体。

3.7 数据仓储建设成效

经过多年实践，农业科技大数据仓储确立了稳定、权威的数据来源，建立了多种类、多路径的数据采集、加工整合模式，形成了多源异构农业科技大数据汇聚标准规范体系和协同工作流程。在仓储管理系统建设方面，选择开源软件Fedora作为底层技术支撑，基于Fedora灵活、可扩展、模块化的架构，且可支持数据资产的长期保存与管理等优点，建设了具有良好通用性和可扩展性的农业科技大数据仓储管理系统，设计并开发了数据提交、数据描述、数据审核与发布、数据质量控制、数据安全管理、数据权益管理、数据动态跟踪与统计等功能模块。数据仓储的建设过程始终秉承着以数据为中心的原则，数据是永久的，系统是暂时的，仓储系统功能随着需求的不断演变进行完善与扩展[16]。鉴于农业科技大数据资源类型多、来源广，仓储系统增加了数据收集、数据互操作模块，方便仓储系统能有效地与其他系统或平台之间交换数据，还设计了多种格式资源的元数据转换器来支持各种元数据对象转换，进一步优化大数据仓储管理系统的功能。

目前农业科技大数据仓储整合集成了文献类、数据类、政策资讯类、事实工具等数据资源，涵盖科技文献、专利、法规、资讯、项目、成果、专家、机构、报告、科学数据、统计数据、专题、百科及农业基础知识库等，资源总量超过2亿条。其中，文献类数据整合了农业领域中外文期刊论文、会议论文、科技报告、专利文献等资源近亿条；数据类资源整合了来自国家农业科学数据中心和部分科研团队的科学数据资源，还包含了采集自国内外权威机构发布的农业相关统计数据；政策资讯类资源收集整理了采集自政府机构、高等院校、国际组织和学会/协会等官网发布的农业相关政策以及科技前沿动态信息；事实工具类数据资源主要收集了涉及农业科研活动本身的事实型数据（如专家学者、科研项目、科技成果、科研机构等）和经过长期研究实践整理出的知识组织工具类资源（如农业百科、农业专业术语、《农业科学叙词表》、学科分类体系、资源代码等）。由此可见，农业科技大数据仓储已经初具规模，可为农业领域科技信息服务和知识服务提供基础数据支撑。

4 农业科技大数据仓储服务

4.1 数据仓储服务内容

农业科技大数据仓储的建设实现了对农业领域各类科技信息资源的汇聚、整合、存储和管理，为农业领域的科技创新、产业发展、成果转化、管理决策等提供了数据资源保障。按照服务对象划分，农业科技大数据仓储包含面向系统和面向用户两大类服务。面向系统的服务包括：①基于农业科技大数据仓储搭建不同规模的农业领域数据服务平台或系统，向用户提供一站式公共集成服务，用户无需在多个数据平台切换，一个数据服务集成平台即可向用户提供数据检索、浏览和下载等服务，为用户查找或获取自己所需的资源提供便利；②面向其他数据平台或第三方平台提供数据接口服务，在保证大数据仓储系统稳定与安全的前提下，可供相关平台访问、检索、调用农业科技大数据仓储资源，实现数据共享，提升资源利用效率。

除了面向系统的数据支撑与共享服务外，农业科技大数据仓储还可直接面向用户群体或个人提供数据服务，用户类型包括科研个体或团队、情报分析专家、企业研发人员、政府决策人员、一般的数据用户等。数据服务的内容紧密围绕数据用户的需求，按照用户需求层次可划分为以下方面。①数据管理服务。面向科研团队或个人提供数据代管服务，尤其是针对科研人员在科研过程中产生的研究数据，为其提供数据管理计划、数据归档与存储、数据发布与共享等方面的服务。②数据查找、收集、整合服务。围绕数据用户的具体需求，针对相关研究领域热点研究问题，基于数据仓储中的资源为用户提供数据抽取服务，如果仓储中的数据资源不能满足用户需求，还可为用户提供数据定向采集、整合等相关数据服务，数据仓储作为资源供给方，在其应用过程中，随着用户需求的变迁，在为用户提供数据服务的同时，也为数据仓储资源的不断完善与扩充提供了有效的途径[17]。③数据挖掘与分析服务。基于数据仓储的资源，利用数据分析工具、模型和算法，面向研

究机构和学科团队、企业、管理部门等，提供个性化数据分析服务、专题知识服务，服务内容包括向用户提供情报产品、数据分析报告等[18]。此时大数据仓储可为用户提供的服务，不再直接向用户呈现数据资源，更倾向于把隐藏在数据背后的信息和知识挖掘出来提供给用户，同时也可以为用户提供数据挖掘与分析工具、模型算法等方面的服务。

4.2 数据仓储服务案例

目前基于农业科技大数据仓储构建了农业专业知识服务系统、农业科技创新联盟平台、中国农业科学院机构知识库等平台系统，用户通过这些平台可以浏览、检索和下载数据。在面向各类用户提供数据服务方面，农业科技大数据仓储开展了面向院士团队、其他科研团队提供数据聚合、抽取和数据分析等方面的服务，如遇到仓储资源不能满足需求的情形，还会根据用户的具体需求，开展数据定向采集、加工处理等服务。农业科技大数据仓储也开展了以内容管理为主的专题数据服务，结合领域用户需求，围绕国家重大战略、农业领域关注焦点，开展专题数据收集、汇聚、管理服务，并设计了相应的专题数据服务产品，如乡村振兴专题服务，收集了国家乡村振兴战略及政策、中国“三农”十年数据、精准扶贫、乡村振兴经典案例集等资源，并提供实时更新与维护。在数据挖掘与分析方面，针对数据用户的关注点，对农业科技大数据仓储中的国际农业统计数据进行分析，涵盖了世界作物产量、世界畜牧产量、世界肉类产量、世界各国农业产值等方面的统计分析，并提供了相关统计指标波动分析功能，实现了各类指标变化趋势的动态展示。此外，还基于农业科技大数据仓储中丰富的文献、专利、项目、获奖成果等数据资源，面向数据用户提供文献计量分析、数值分析、聚类分析以及模型算法等方面的数据分析服务，提供领域战略情报，内容涵盖学科发展态势分析、研究热点分析、机构科研竞争力评价、国家科研情况对比分析等。

5 结语

农业科技大数据的建设已初具规模，通过数字化加工、网络采集、人工搜集等多种形式，采集与农业科技相关的信息与数据，经过规范化加工处理，目前形成了包含文献类、数据类、政策资讯类、事实工具类等的数据资源体系。为规范资源建设、保障数据质量，针对各类资源制定了可扩展性较强的元数据标准规范，方便了各类资源的集成与整合，并确立了包含数据来源、数据汇聚与整合、数据存储、数据管理、应用与服务等层面的大数据仓储框架体系，建成了我国农业领域资源类型丰富的农业科技大数据仓储，整合集成了包含科技文献、科学数据、专利、政策法规、科技项目、科技成果、专家、机构、行业报告等类型的数据资源，资源总量近2亿条，且仍在源源不断地补充中。数据仓储一直秉承边建设、边应用的原则，以用户需求为导向，在实际应用中不断完善数据资源，目前已经为多个服务平台或系统提供资源支撑，包括农业专业知识服务系统、农业科技联盟信息资源共建共享平台等，此外还向部分机构知识库提供数据服务，也面向不同的用户提供各类数据服务。由此可见，农业科技大数据仓储的应用取得了一定进展，形成了一些产品，但在多源异构数据汇聚融合、知识挖掘与分析方面的研究还不够深入，因此如何实现多源异构农业科技数据资源的全面汇聚、深度融合仍然是今后努力的方向之一。此外，大数据时代数据挖掘与分析非常重要，面对庞大的数据资源，需要引入神经网络等人工智能新兴技术进行数据挖掘分析，并开展数据挖掘模型与算法方面的研究，才能使农业科技大数据充分发挥其价值。

参考文献

[1] 农业部：加快农业现代化 建设农业科技大数据平台 [J]. 农村实用技术，2015（10）：61.

[2] 农业部市场与经济信息司. 农业部关于推进农业农村大数据发展的实施意见 [J]. 农业工程技术，2016（3）：15-20.

[3] FAO. databases [EB/OL]. [2020-01-15]. http://www.fao.org/statistics/databases/en/.

[4] U.S. Department of Agriculture. Data [EB/OL]. [2020-01-20]. https://www.usda.gov/topics/data.

[5] 朱亮，孟宪学，赵瑞雪，等. 国家农业科学数据共享中心资源建设探析 [J]. 数字图书馆论坛，2017（11）：15-20.

[6] 赵红伟，崔运鹏. 农业科技工作数据汇交系统建设 [J]. 内蒙古农业大学学报（自然科学版），2019，40（4）：85-93.

[7] 冼晓青，陈宏，赵健，等. 中国外来入侵物种数据库简介 [J]. 植物保护，2013，39（5）：103-109.

[8] 王觅也，郑涛，李楠，等. 医疗大数据集成及应用平台体系构建 [J]. 医学信息学杂志，2019，40（8）：37-42.

[9] 赵瑞雪，鲜国建，罗婷婷，等. 中国工程科技知识中心元数据规范（II）[M]. 北京：中国农业科学技术出版社，2019：2-18.
[10] 宋瑜辉. 基于MongoDB存储和分析辅助决策系统中的海量日志 [J]. 科技创新与应用，2019（33）：5-8.
[11] 谢福成. 面向金融行业数据仓库的数据质量控管的研究与实现 [D]. 厦门：厦门大学，2009.
[12] 刘静羽，黄金霞，王昉，等. 数字资源权益状况描述框架研究 [J]. 数字图书馆论坛，2019（9）：9-15.
[13] 刘栢嵩. 学术型大数据知识组织与服务模式研究 [M]. 杭州：浙江大学出版社，2018：22-23.
[14] 鲜国建，赵瑞雪，寇远涛，等. 农业科学叙词表关联数据构建研究与实践 [J]. 现代图书情报技术，2013，29（11）：8-14.
[15] 鲜国建，赵瑞雪，孟宪学，等. 基于知识组织体系的多维语义关联数据构建研究 [J]. 数字图书馆论坛，2014（3）：11-18.
[16] 曾婷，董丽，邹荣，等. 开源仓储软件在清华大学图书馆的研究应用与思考 [J]. 图书馆杂志，2012，31（5）：58-64.
[17] 赵瑞雪，鲜国建，寇远涛，等. 大数据环境下的农业知识发现服务探索 [J]. 数字图书馆论坛，2016（9）：28-33.
[18] 钱力，谢靖，常志军，等. 基于科技大数据的智能知识服务体系研究设计 [J]. 现代图书情报技术，2019，3（1）：4-14.

作者简介

赵华，女，1980年生，博士研究生，助理研究员，研究方向：信息资源管理。
赵瑞雪，女，1968年生，博士，研究员，通信作者，研究方向：信息管理与信息系统、信息资源管理、知识组织及数字图书馆，E-mail：zhaoruixue@caas.cn。
金慧敏，女，1978年生，硕士，助理研究员，研究方向：计算机科学与技术。
郑建华，女，1986年生，博士，高级工程师，研究方向：农业信息管理。
鲜国建，男，1982年生，博士，研究员，研究方向：知识组织、关联数据、语义出版。
朱亮，男，1981年生，博士，副研究员，研究方向：科学数据管理。
寇远涛，男，1982年生，博士，研究员，研究方向：数字图书馆理论与技术、信息管理与信息系统。

Construction and Service of Big Data Warehouse of Agricultural Science and Technology

ZHAO Hua[1,2] ZHAO RuiXue[1,2] JIN HuiMin[1,2] ZHENG JianHua[1,2] XIAN GuoJian[1,2] ZHU Liang[1,2] KOU YuanTao[1,2]
(1. Institute of Agricultural Information, China Academy of Agricultural Sciences, Beijing 100081, China; 2. Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing100081,China)

Abstract: In the era of big data, agricultural science and technology resources will play an important role in national agricultural strategy formulation, scientific research and innovation, and achievements transformation. This paper summarized the current situation of big data development in agriculture at home and abroad, put forward the conception and content system of big data of agricultural science and technology. The paper analyzed the construction of big data warehouse of agricultural science and technology knowledge from the aspects of data source, data collection, data storage and management, data organization, etc., established the data warehouse framework, finally expounded the application and service of big data warehouse of agricultural science and technology in order to provide reference for the research and practice of big data in the field of agriculture in China.

Keyword: Agriculture Scientific and Technological; Big Data; Data Warehouse; Data Service

（收稿日期：2020-06-20）