数据密集型农业科研服务平台架构设计*

张丹丹[1] 赵瑞雪[1,2] 王剑[1] 鲜国建[1,3] 黄永文[1,2]
（1. 中国农业科学院农业信息研究所，北京 100081；2. 国家新闻出版署农业融合出版知识挖掘与知识服务重点实验室，北京 100081；3. 农业农村部农业大数据重点实验室，北京 100081）

摘要：针对数据密集型科研范式下农业领域数据密集型计算服务能力不足的问题，研究设计集泛在智能知识服务与专业领域计算服务于一体的数字科研知识服务模式。调研分析国内外数字科研基础设施与数字科研平台建设经验，剖析数据密集型农业科研的新特征和知识服务需求，从基础设施层、数据层、关键技术层、核心功能层和服务层5个层面阐述数据密集型农业科研服务平台的架构方案，并给出面向作物计算育种知识服务的数字科研平台设计实例。以期为数据密集型农业科研平台的研发和应用场景的落地提供参考，为支撑数据密集型农业科研创新提供可借鉴的知识服务路径。

关键词：农业数字科研平台；数据密集型科研；知识服务；平台架构
中图分类号：G647.25　DOI：10.3772/j.issn.1673-2286.2023.10.008
引文格式：张丹丹，赵瑞雪，王剑，等. 数据密集型农业科研服务平台架构设计[J]. 数字图书馆论坛，2023（10）：71-78.

随着农业科研数据“井喷式”的增长和信息技术在农业科研中的深化应用，农业科学研究由假设驱动的被动探索转向数据驱动的主动知识发现。越来越多的科研工作对现有科研数据进行了重新分析、组织、解析和利用，科研人员的知识服务需求也由传统的检索服务转向对数据密集型计算服务的需求。数据、信息与知识转化并产生新知识，成为农业科技创新的引擎，这使得新型的数字科研平台对农业科学研究的支撑作用愈加凸显[1]。传统以信息检索为主的农业数字科研平台已经无法满足当前农业科技创新对知识服务的需求。因此，面向数据密集型农业科研的新特征与知识服务的新需求，构建技术自主、安全可控的数据密集型农业科研平台尤为重要。

在数据密集型科研范式不断推进的大背景下，数字科研基础设施和平台的建设已成为加速科技创新发展的引擎。在农业领域相关的数字科研基础设施建设方面，农业食品与环境研究基础设施agINFRA采用基于云的基础架构以实现数据存储和索引、算法执行、结果可视化等[2]。生物多样性基础设施LifeWatch-ERIC（European Research Infrastructure Consortium）作为一个分布式基础设施联盟，可提供数据共享、分析工具和虚拟实验室来支持生物多样性领域研究[3]。在农业领域相关的数字平台建设方面，各国相继开展了相关建设实践。其中，美国的DNA序列数据平台GenBank[4]、欧洲的核苷酸序列数据平台EMBL（European Molecular Biology Laboratory）[5]、日本的基因数据平台DDBJ（DNA Data Bank of Japan）[6]、美国国家生物技术信息中心的基因型与表型数据平台dbGaP（Database of Genotypes and Phenotypes）[7]等，为农作物表型分子机制的解析提供了基因组注释信息。水稻基因组注释平台Rice Galaxy[8]、玉米基因组和遗传分析平台MaizeGDB（Maize Genetics and Genomics Database）[9]、小麦基因组平台IWGSC（International Wheat Genome Sequencing Consortium）[10]、小麦蛋

收稿日期：2023-08-21
*本研究得到科技创新2030——“新一代人工智能”重大项目“农业智能知识服务平台研发与应用示范”（编号：2021ZD0113705）资助。

白质组平台Wheat Proteome[11]以及棉花遗传育种数据平台CottonGen（Cotton Database Resources）[12]等数字平台的建设为作物育种科学研究提供了领域数据计算服务。我国在农业数字科研平台建设方面也取得了积极进展，涌现出了水稻相关物种基因组数据库Rice-RelativesGD（Rice Relatives Genomic Database）[13]、水稻基因组变异及功能注释平台RiceVarMap（Rice Variation Map）[14]、水稻表观组学注释平台eRice（Rice Epigenetic and Epigenomic Database）[15]、水稻泛基因组注释平台RPAN（Rice Pan-Genome Browser）[16]、水稻功能基因组育种平台RFGB（Rice Functional Genomics and Breeding）[17]、玉米多组学综合数据平台ZEAMAP[18]、玉米蛋白质-蛋白质互作平台PPIM（Protein-Protein Interaction Database for Maize）[19]、玉米多组学基因网络分析平台MCENet （Maize Conditional Co-Expression Network）[20]、小麦族同源基因数据平台TGT（Triticeae-GeneTribe）[21]、小麦基因定位与基因组功能研究平台WheatGmap（Wheat Gene Mapping）[22]等覆盖农业各类研究领域的数字科研平台。这些平台基于农业学科领域发展需求，明确支持目标与愿景，采用适合的技术基础设施、适用的工具模型和系统框架，面向平台使用者提供数据管理、科研协作、教育培训等方面的服务支撑，并在功能和结构上具有一定的可扩展性和灵活性，以方便适应新的科研需求以及与其他平台的协作，也为后续面向数据密集型科研范式的农业科研数字化设施架构研究提供了良好的借鉴。

综观发达国家取得的成果和我国的建设经验，农业数字科研平台在应对数据密集型科研范式的转变，推动农业科研创新，加速涉农领域内科研产出和数据收集、整合、共享等方面均起着关键的支撑作用。然而在实践中，我国农业数字科研平台建设依然面临着农业科研数据管理体系不健全、高性能计算能力不足、领域算法模型化程度不高与学科知识发现能力不足等问题。为此，本文系统梳理国外数字科研平台建设的成功经验，按照整体统筹和长期可持续的原则，充分借助人工智能、云计算和大数据等关键新兴技术的优势，从基础设施层、数据层、关键技术层、核心功能层和服务层5个层面阐述数据密集型农业科研平台的架构，并以此提出面向作物计算育种的数字科研平台设计思路与建设方案，为数据密集型农业科研平台的研发和应用场景的落地提供参考。

1 数据密集型农业科研的新特征

1.1 科学研究协同化

在数据密集型科研生态环境下，农业科学研究的跨学科、交叉性、综合性、复杂性等特征更加突出，学科交叉融合日益深化。农业科技重大问题的突破越来越依赖于生物科学、信息科学等不同学科之间的协同研究，多学科交叉所形成的综合性、系统性知识可以触发新的科学发现，学科交叉研究成为取得原创性科研成果的重要途径。

从解决农业科学研究问题的现实需求来看，需要开展多学科领域的协同研究。围绕共同的研究目标，充分发挥不同研究主体的优势与特色，形成协同高效的农业科学研究体系。同时，强化信息技术和领域技术的深度融合，发挥新兴技术在数据密集型科研范式中的前瞻性、引领性和战略性作用。加快各主体、各方面和各环节的有机互动，有效汇聚创新资源和创新要素，实现优势互补、资源共享和合作攻关，建立起完整的科研协同链条，不断提升协同效率和促进科技成果产出。

1.2 研究范式数据化

在数据密集型科研范式的时代背景下，越来越多的农业科研工作趋向于对现有科研数据进行重新分析、组织、关联、解析与利用，科研数据成为了科学研究的知识基础和有力工具。数据、信息与知识转化并产生新知识，成为农业科研创新的引擎，科研人员的工作重点转变为通过分析与挖掘科研数据提出科学假设以加快科学研究的进程。

聚焦农业科研创新中数据驱动的关键作用，基于数据资源共建共享的建设目标，需要构建嵌入科学研究过程的科研数据智能管理体系，实现阶段性试验数据的实时采集与存储。制定融汇治理多源异构数据的标准体系，实现科研数据的规范化融合与科学应用。研发数据计算分析工具集，实现在线计算和科学分析预测。最终形成集数据存储与管理、数据分析与计算以及基于数据决策于一体的科研数据智能分析体系，实现科研数据的全面整合、提炼加工和价值增值。

1.3 数据计算模型化

在数据密集型农业科学研究中，科学发现的模式

由传统的因果逻辑关系探求转向对多维度数据间相关关系的探索，农业科学研究更加依赖于数据密集型计算分析服务与知识挖掘服务。各类数据分析工具和领域数据计算模型在整个科研活动中的支撑作用凸显，并成为数据密集型农业科研创新过程必不可少的组成部分。基于海量数据的计算分析已成为攻克复杂农业科学研究问题的重要手段。

针对数据密集型农业科研创新对领域数据密集型计算模型的依赖性，亟需加强数据密集型计算分析体系建设，重视数据密集型计算模型的构建。以农业领域中最具数据密集型计算典型特征的作物育种科学研究为例，针对基因型计算模型，运用大数据框架来管理并提供遗传距离计算、聚类分析和群体结构分析服务，为研究不同作物材料间亲缘关系与不同亚群间的分化提供科学有效的数据计算服务，旨在实现优异功能基因的挖掘，助力作物优质新品种的培育。面向农业领域各学科知识服务的需求，明晰科学研究活动的一般规律，梳理掌握科学研究工作的基本特征和需求关键点，强化算力、算法，构建拟合学科特点的数据密集型计算模型，形成支撑科研协同创新的新型计算分析体系。同时，需要嵌入科学研究过程，建立融入科研活动全流程的智能辅助技术体系和工具体系，满足科研机构和科研人员的数据密集型计算服务需求，助力农业科学研究机理的智能探索和科学知识的新发现。释放科研人员的最大效能，推动农业科学研究的智能化发展。

1.4 研究工具智能化

在数据密集型科研范式下，随着机器学习、知识表示、语义推理等技术的发展与深度应用，科学研究所依托各类工具的智能性不断提升，即各类研究工具已突破传统的统计分析的范畴，转向更高层次的计算建模方向。具体而言，在特定的科研场景下，由于传统的统计分析方法很难在海量数据中抽取出其所蕴含的知识和规律，科研人员只有运用智能化的计算与分析工具才能够在不同的数据环境下以自动化的方式发现潜在的知识、关系和规律，从而解决更复杂、更前沿的科学问题。

从农业科学研究对工具的应用需求来看，传统农业科研模式向数据密集型科研模式转变的实质是研究工具从简单的统计计算形式转变为高阶的建模计算应用形式，这说明在大数据时代，农业各领域的科学研究已不能单一地应用人脑判断和简单的统计分析来完整地发现科研对象的内部关系。因此，只有依托智能化的研究工具，农业领域研究者才能更快、更有效地揭示多源涉农大数据所蕴含的各类关系和特征，实现对研究对象的深度分析和理解，从而满足科研活动不断扩展深度与广度的需求。由此可见，在数据密集型农业科研范式下，研究工具必须深度融合大数据、人工智能等核心关键技术，形成一系列支撑智能化研究方法的产品，推动智慧化研究模式构建，充分满足研究者对各类涉农对象深度洞察的需求。

相对于其他领域的数据密集型科研活动的通用特征而言，由于本质和目标的独特性，农业科研具有多学科、多数据类型、实时性和可持续性强等个性化特征，因而农业领域的数据密集型科研更加强调多源异构数据的实时传输和归一化处理与应用，以期能够最大限度地汇聚和分析多领域的数据资源，发现影响农业领域科学问题的多种变量的运作规律。以华大基因为例，其就多源异构农业基因数据引发的存储、分析、共享瓶颈，开发了云平台系统BGI Online，通过高性能计算、大规模存储、安全互联网等强大的基础设施建设，初步开发并构建了农业基因数据实时化异构消除与归一化处理解决方案，实现了高效的基因数据流“存、算、传、管”各个维度的服务功能，有效支撑了基因组学领域的数据密集型科研。

2 数据密集型农业科研平台架构设计

构建的数据密集型农业科研平台需要基于上述数据密集型农业科研的新特征，面向科研数据科学管理、数据密集型计算分析以及知识挖掘等服务需求，聚焦数据密集型计算服务能力不足的问题，重点实现文本挖掘、深度学习、云计算与认知计算等数据密集型计算关键技术在农业领域各学科知识服务中的应用。具体而言，要实现以下功能：提供一流的数据基础设施和基于云端的知识服务，支持数据存储、数据发现、数据管理服务；提供虚拟研究环境和服务，支持跨领域的科研人员获取和共享数据、分析工具、计算模型；提供自主可控、开放协作、无缝访问、泛在可持续的新型数字农业科研协同创新环境。

从支持现代农业研究和农业决策的角度来看，数据密集型农业科研平台应该遵循实时性（实时的数据处理能力）、伸缩性（具有良好的功能扩展性）、开放性（具备开放的标准与协议）与互操作性（支持不同平台

间的数据与功能共享）的设计理念。数据密集型农业科研平台架构主要包括5层，如图1所示：基础设施层、数据层、关键技术层、核心功能层和服务层。为保证平台架构具有良好的互操作性，还需要制定数据标准、语义标准等标准规范体系，制定有关开放共享以及安全防护等的政策机制，保证数据的可访问和可重复使用。重点建设农业科研大数据中心，融合数据与计算分析，支持研究成果共享、数据访问和数据密集型计算，为不同类型用户提供场景化服务，以支撑数据驱动的农业科学研究。

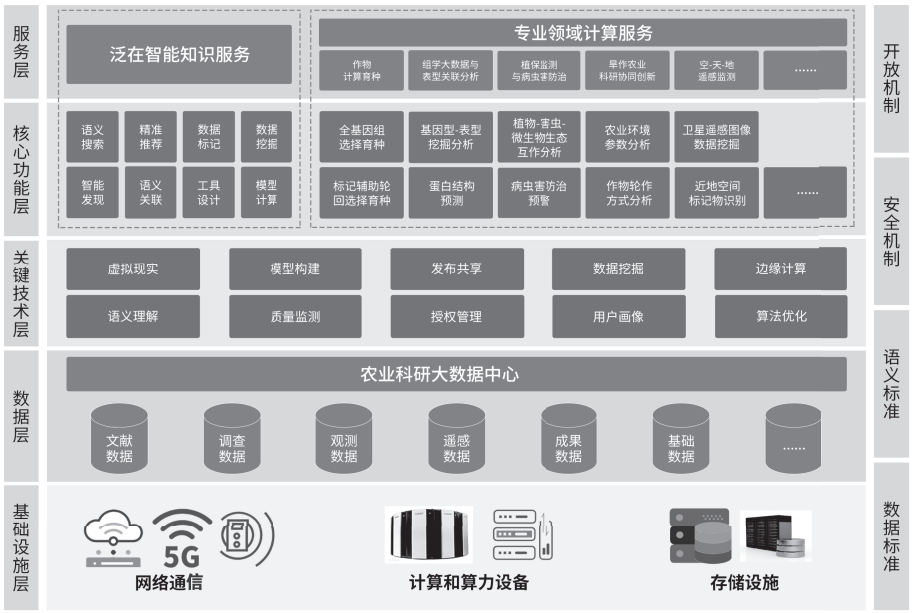

图1 数据密集型农业科研平台典型架构

2.1 基础设施层

基础设施层是整个架构的基础，能够为上层的应用提供网络、计算与存储等基础资源，并要具备一定的资源监控、安全防护等功能。具体而言，基础设施层能够联合已有且分散的各类基础设施，保障数据和设备可发现和可使用，以分布的方式为平台的其他层次提供联网、计算、存储服务。

基础设施层在功能上应用基础设施与数据、技术和应用相分离的形式，将网络资源等设施分离到一个相对独立的层级结构中。一方面，提高既有设备的复用水平，有效降低IT基础设施和资源方面的资金投入与配置难度，从而极大方便科研数据资源的接入与处理，显著增强数字科研平台的弹性化部署能力。另一方面，基础设施层能够有效支撑异构系统的集成，快速建立用于通信联络与协同工作的虚拟专用通路和共享设备，从而实现异构资源之间的无缝对接。例如，在农业科研大数据采集场景下，应用基础设施层的功能架构，能够有效地借助云端技术，将已有的设备纳入平台体系进行复用，从而有效克服传统集中式部署重复购买和配置基础设备的弊端，极大地提升数据密集型农业科研平台建设效能。

2.2 数据层

数据层集中管理、存储与应用核心资源，是平台构建的关键基石。在数据层的构建中，需要研究和制定融合多类型数据的元数据标准、语义标准及互操作标准，研发融汇治理农业科技文献数据、调查数据、观测数据、遥感数据、成果数据等的关键技术，形成科研大数据集市。提供大规模异构数据的存储服务，并支持对海量数据的分布式存储和管理，从而为打造集约高效、开放互联、泛在适用、自主可控、安全可靠的农业科研大数据仓储提供服务支持。

就功能与架构设计而言，数据层主要为整体平台提供管理、鉴权、审计、融合与分析数据等功能，其能够向下对接基础设施层中的存储设备进行数据读取和写入，向上则通过各类接口与各层级进行数据交换与服务应用。相对于传统科研平台架构，这种数据与设备、技术、应用和服务相分离的架构能够使平台建设工

作更加聚焦数据自身的需求，强调数据在新型科研平台运营中的作用。例如，在这种平台架构下，可以根据数据敏感程度，对所收集的科研数据进行分级分类，选择不同的存储方式、处理技术和应用领域。即当数据涉及个人信息、国土安全、生物信息等较为敏感的信息时，科研平台的数据层会选择私有云进行存储，科研平台也会根据相关规定设计相关处理技术和选择应用范围，赋予不同的访问权限，从而实现细致的权限管理，同时根据需求的变化，动态调整服务级别和数量，真正保障数据安全与技术完备。

2.3 关键技术层

关键技术层是数据密集型科研范式下知识服务的重要支撑，也是新兴信息技术重构传统知识生态与科研创新模式的重要依托。具体而言，关键技术层应充分利用人工智能、机器学习等技术大幅度提升算法和模型的效率和性能，包括语义理解、数据挖掘、边缘计算、用户画像等关键技术，以及支撑特定领域研究的算法和模型、认证授权和质量监测服务。提供对海量数据的计算分析服务，支持对大数据进行高可靠、高性能的并行计算，支持云计算和高通量计算。基于FAIR原则，确保科研数据的可重复性与互操作性，支持跨学科的数据服务，有效解决数据孤岛的问题。

在实践中，关键技术层作为承接数据资源与核心功能的重要保障环节，其功能主要体现在2个方面：一是能够对数据层的各类资源进行加工处理，提升数据资源弹性和融合性，将各类多源异构的数据资源集成为统一的资源形式，显著增强数据的可用性；二是根据知识服务场景的应用需求，能够挖掘各类数据资源并进行可视化展示，实现数据与知识服务的双向弹性互动。例如，在环境生态数据采集过程中，关键技术层能够为知识服务场景提供运算、数据库和网络通路选择等相关采集与传输技术，以及使用Hadoop等工具进行数据挖掘分析。因此，关键技术层是数据密集型农业科研平台最大化数据价值的抓手，也是赋能知识服务应用场景的重要保障。

2.4 核心功能层

核心功能层作为平台的业务支撑层，应支持用户以简单的方式访问开放数据和跨学科科研数据，并支持无缝嵌入科研工作流程的服务、工具和模型。主要包括语义搜索、智能发现、精准推荐、模型计算等服务功能，同时用户可以发现、访问、重用、合并和分析研究数据，利用相关的计算模型和工具实现数据挖掘、在线分析、可视化展示以及模型计算。此外，还应提供面向专业学科领域数据的各类计算服务，以支撑数据密集型科研的转型升级。例如全基因组选择育种、基因型-表型挖掘分析、植物-害虫-微生物生态互作分析、作物轮作方式分析等。

就功能定位而言，核心功能层基于关键技术与数据接口，面向不同用户需求和应用场景，提供相应的服务方式，如内部办公、邮件通信、数据加工、资源检索等。一般来说，核心功能层中功能的分布与服务方式是由科研业务活动需求决定的，其目标是解决具体科研场景中的各项问题，重点支撑相应的主营业务流程和专业领域的学科知识服务。由此可见，面向差异化的科研服务场景，核心功能层将提供面向场景应用需求的知识服务。在泛在的智能知识服务方面，可提供语义检索、精准推荐、模型计算和智能发现等服务。在专业领域知识服务方面，可提供蛋白结构预测、作物轮作方式分析和近地空间标记物识别等服务。尽管核心功能层在上述不同学科场景下功能服务方式各异，但其最终目标是提升科学研究的效率，从而推动整体科研协作水平和创新能力提升。

2.5 服务层

针对农业科学研究中数据密集型计算服务的关键需求，结合各学科科研工作的基本特征与知识服务的重要需求点，服务层为用户提供数据发现、访问、使用和再利用等泛在智能知识服务，还可以为用户提供满足各学科多样化场景需求的数据密集型计算服务。例如，为用户提供动态按需分析服务，支撑作物计算育种、植物监测与病虫害防治、组学大数据与表型关联分析以及空-天-地遥感监测等学科体系下的计算型科学研究。此外，科研人员还可以利用虚拟社区中的计算资源开展合作研究。

就服务层的实现方式而言，其本质是封装一些核心功能并以服务的形式提供给用户。一般来说，根据服务和功能所应用的数据性质，服务层所涉及的各项应用服务也具备不同的等级。普惠式的知识服务可以部署在公有云平台上，而个性化的知识服务则应部署在私有云上，如政务云、行业内网等，供具备一定访问权限

的用户使用，从而在服务层面上实现分级分类与个性化应用。此外，服务标准体系也是服务层的重要内容之一。在实践中，农业数字科研平台内统一、规范的服务标准是科研协同、信息共享、知识传播的重要保障。由此可见，基于统一规范的服务应用，“融入环境、嵌入过程”的科研平台特征凸显，平台真正地同科研活动协同演进，实现服务层面的创新驱动，进而有效提升整体的服务能力。

3 作物计算育种数字科研平台功能架构

随着高通量基因测序技术在作物育种领域的快速应用，作物育种已进入智能计算育种时代[23]。作物计算育种以人工智能为依托，整合现有的基因、环境和表型等多模态多维度海量数据集，建立基于深度学习的精准预测模型，推动作物育种从“试验选优”向“计算选优”转变。这意味着作物计算育种领域是当前数据密集型农业科研中最具典型意义的应用场景。为此，基于所提出的数据密集型农业科研平台典型架构，结合作物计算育种领域研究对优异基因挖掘和性状精准预测知识服务的关键需求，提出作物计算育种数字科研平台功能架构（见图2），并以此为数据密集型科研平台架构落地应用的切入点。

围绕作物育种的科学研究流程，以基因组测序、基因编辑、全基因组选择等前沿生物技术驱动，并深度融合人工智能、大数据等信息技术，遵循“管理—分析—决策”数字化科研平台建设思路，基于数据密集型农业科研平台典型架构理念，设计作物计算育种数字科研平台，该平台涵盖多维育种数据存储、数据分析模型选择、育种数据分析计算与育种方案智能决策四大模块。其中：多维育种数据存储模块主要运行于数据密集型农业科研平台典型架构的基础设施层与数据层，数据分析模型选择模块依托关键技术层，育种数据分析计算与育种方案智能决策模块则分别与核心功能层与服务层关联。该平台旨在促进育种科学研究范式的变革，为农作物新品种的研发决策提供信息和数据支撑，缩短作物育种周期并提高育种精度，最终实现高效、智能、定向培育新品种。

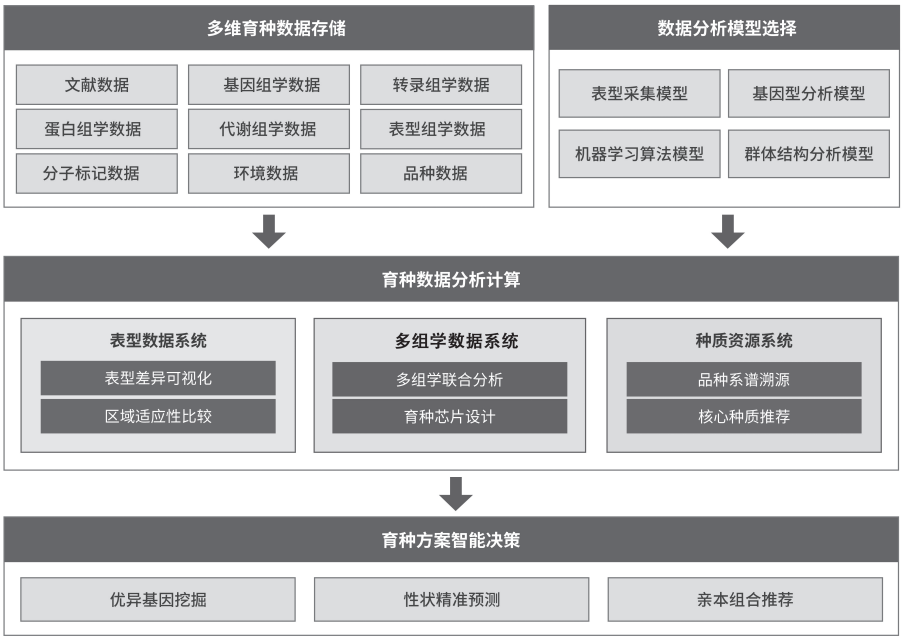

图2 作物计算育种数字科研平台功能架构

多维育种数据存储模块主要利用基础设施层的存储设备与数据层的相关功能来提供面向文献数据、基因组学数据、转录组学数据、蛋白组学数据、代谢组学数据、表型组学数据、分子标记数据、环境数据和品种数据的分布式存储服务，实现对多源异构育种数据的汇聚治理。在科学管理多维育种数据的基础上，根据应用的需求，平台的数据分析模型选择模块基于关键技术层的工具，提供自研的表型采集模型、基因型分析模型、机器学习算法模型和群体结构分析模型，可按需为用户提供专业的数据采集和分析工具，支持田间数据汇总、聚类分析、群体结构分析以及育种材料之间的基因型比对等知识服务。

育种数据分析计算模块能够嵌入整个作物育种科学研究流程，从而充分发挥核心功能层的价值，其主要包括表型数据系统、多组学数据系统和种质资源系统3个子系统。表型数据系统可整合分析杂交组合品种多年多点的试验数据，可视化展示不同品种间表型的差异以准确判断品种的特点，可帮助育种学家筛选出各方面表型最佳的新品种以及适合推广的区域。多组学数据系统可支持强大的组学数据在线计算和科学分析预测服务，旨在实现育种材料多组学数据与杂交后代田间表型的关联分析。系统将提供多组学联合分析工具，对尚未配制的杂交组合品种的田间表型进行预测，以提高育种的效率。此外，还可以提供聚类分析、主成分分析等数据分析服务，可应用于目的基因检测、定向改良、优异基因聚合以及回交育种等。种质资源系统结合表型数据系统、多组学数据系统，为育种学家提供品种系谱溯源和目的性状的核心种质资源推荐服务。能够可视化展示品种间亲缘关系，实现对野生近缘种或野生种中优异功能基因的溯源挖掘。基于优异基因挖掘、性状精准预测和亲本组合推荐等知识服务，作物计算育种数字科研平台能够有效支撑数据密集型科研范式下育种方案的智能决策。

作物计算育种数字科研平台围绕作物育种“数字化—信息化—智能化”的发展路线，贯穿育种数据采集，育种数据的信息化管理、统计分析，机器学习建模和育种精准预测，旨在从多个组学水平对性状进行由内而外、相互关联的全面深入揭示，系统深入地挖掘基因与性状之间的内在关系，在不同层次上揭示生命活动的规律。为作物计算育种提供结合多维组学数据的数据密集型计算服务，支撑作物育种由经验主导向定位化和精准化方向转变。

4 结语

在数据密集型科研范式的时代背景下，数据密集型农业科研平台已成为加速农业科学研究进程的引擎和农业科研创新的重要支撑力量。针对农业科研创新对数据密集型计算服务的关键需求，本研究提出了数据密集型农业科研平台的架构模型，并以此设计出嵌入作物计算育种科学研究全流程的数字科研平台功能架构。本研究为数据密集型农业科研平台的研发和应用场景的落地提供参考，为作物计算育种科学研究领域提供了一个可借鉴的知识服务方法。数据密集型农业科研平台的构建构筑在科研数据深度挖掘和人工智能基础之上，以科研数据规范管理为基础保障，以新兴信息技术为驱动力，以虚拟科研环境为支撑，以开放共享为机制，实现对数据密集型科研环境下资源、技术和服务场景的高效融合，并不断推进智能驱动的新一代科研范式的产生。

参考文献

[1] NA L，YAN Z. Promote data-intensive scientific discovery，enhance scientific and technological innovation capability：new model，new method，and new challenges comments on “the fourth paradigm：data-intensive scientific discovery”[J]. Bulletin of Chinese Academy of Sciences，2013，28（1）：115-121.

[2] DRAKOS A，PROTONOTARIOS V，MANOUSELIS N. agINFRA：a research data hub for agriculture，food and the environment[J]. F1000Research，2015，4：127.

[3] GÄRDENFORS U，JÖNSSON M，OBST M，et al. Swedish LifeWatch：a biodiversity infrastructure integrating and reusing data from citizen science，monitoring and research[J]. Human Computation，2014，1（2）：147-161.

[4] SAYERS E W，CAVANAUGH M，CLARK K，et al. GenBank[J]. Nucleic Acids Research，2020，48（D1）：D84-D86.

[5] CANTELLI G，BATEMAN A，BROOKSBANK C，et al. The European bioinformatics institute （EMBL-EBI） in 2021[J]. Nucleic Acids Research，2022，50（D1）：D11-D19.

[6] OGASAWARA O，KODAMA Y，MASHIMA J，et al. DDBJ Database updates and computational infrastructure enhancement[J]. Nucleic Acids Research，2020，48（D1）：D45-D50.

[7] MAILMAN M D，FEOLO M，JIN Y M，et al. The NCBI dbGaP database of genotypes and phenotypes[J]. Nature Genetics，2007，39（10）：1181-1186.

[8] JUANILLAS V，DEREEPER A，BEAUME N，et al. Rice Galaxy：an open resource for plant science[J]. GigaScience，2019，8（5）：giz028.

[9] PORTWOOD J L，WOODHOUSE M R，CANNON E K，et al. MaizeGDB 2018：the maize multi-genome genetics and genomics database[J]. Nucleic Acids Research，2019，47（D1）：D1146-D1154.

[10] The International Wheat Genome Sequencing Consortium （IWGSC），APPELS R，EVERSOLE K，et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome[J]. Science，2018，361（6403）：eaar7191.

[11] DUNCAN O，TRÖSCH J，FENSKE R，et al. Resource：map-

ping the triticum aestivum proteome[J]. The Plant Journal，2017，89（3）：601-616.

[12] YU J，JUNG S，CHENG C H，et al. CottonGen：the community database for cotton genomics，genetics，and breeding research[J]. Plants，2021，10（12）：2805.

[13] MAO L F，CHEN M H，CHU Q J，et al. RiceRelativesGD：a genomic database of rice relatives for rice research[J]. Database，2019，2019：baz110.

[14] ZHAO H，LI J C，YANG L，et al. An inferred functional impact map of genetic variants in rice[J]. Molecular Plant，2021，14（9）：1584-1599.

[15] ZHANG P X，WANG Y F，CHACHAR S，et al. eRice：a refined epigenomic platform for japonica and indica rice[J]. Plant Biotechnology Journal，2020，18：1642-1644.

[16] SUN C，HU Z Q，ZHENG T Q，et al. RPAN：rice pan-genome browser for ~3000 rice genomes[J]. Nucleic Acids Research，2017，45（2）：597-605.

[17] WANG C C，YU H，HUANG J，et al. Towards a deeper haplotype mining of complex traits in rice with RFGB v2.0[J]. Plant Biotechnology Journal，2020，18（1）：14-16.

[18] GUI S T，YANG L F，LI J B，et al. ZEAMAP，a comprehensive database adapted to the maize multi-omics era[J]. iScience，2020，23（6）：101241.

[19] ZHU G H，WU A B，XU X J，et al. PPIM：a protein-protein interaction database for maize[J]. Plant Physiology，2016，170（2）：618-626.

[20] TIAN T，YOU Q，YAN H Y，et al. MCENet：a database for maize conditional co-expression network and network characterization collaborated with multi-dimensional omics levels[J]. Journal of Genetics and Genomics，2018，45（7）：351-360.

[21] CHEN Y M，SONG W J，XIE X M，et al. A collinearity-incorporating homology inference strategy for connecting emerging assemblies in the triticeae tribe as a pilot practice in the plant pangenomic era[J]. Molecular Plant，2020，13（12）：1694-1708.

[22] ZHANG L C，DONG C H，CHEN Z X，et al. WheatGmap：a comprehensive platform for wheat gene mapping and genomic studies[J]. Molecular Plant，2021，14（2）：187-190.

[23] WALLACE J G，RODGERS-MELNICK E，BUCKLER E S. On the road to breeding 4.0：unraveling the good，the bad，and the boring of crop quantitative genomics[J]. Annual Review of Genetics，2018，52：421-444.

作者简介

张丹丹，女，博士研究生，研究方向：功能基因组、学科知识服务、数字科研平台。
赵瑞雪，女，博士，研究员，通信作者，研究方向：农业信息管理，E-mail：zhaoruixue@caas.cn。
王剑，男，博士，副研究员，研究方向：农业信息技术。
鲜国建，男，博士，研究员，研究方向：知识组织与知识服务。
黄永文，女，博士，副研究馆员，研究方向：知识组织与知识服务。

·

Architecture Design of Data-Intensive Agricultural Research Service Platform

ZHANG DanDan[1] ZHAO RuiXue[1,2] WANG Jian[1] XIAN GuoJian[1,3] HUANG YongWen[1,2]
(1. Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081, P. R. China; 2. Key Laboratory of Knowledge Mining and Knowledge Services in Agricultural Converging Publishing, National Press and Publication Administration, Beijing 100081, P. R. China; 3. Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081, P. R. China)

Abstract: Aiming at the insufficiency of data-intensive computing service capability in the agricultural field under the data-intensive scientific research paradigm, a digital scientific research knowledge service model integrating universal intelligent knowledge service and professional computing service is researched and designed. We investigate and analyze the construction experience of digital scientific research infrastructure and digital scientific research platform at home and abroad, and analyze the new characteristics of data-intensive agricultural scientific research and knowledge service requirements. From the five levels of infrastructure layer, data layer, key technology layer, core function layer, and service layer, an architecture scheme of data-intensive agricultural scientific research service platform is expounded, and a design example of crop computational breeding knowledge service oriented digital scientific research platform is given. This study provides references for the research and development of data-intensive agricultural scientific research platform and the landing of application scenarios, and provides a referable knowledge service path for supporting data-intensive agricultural scientific research innovation.

Keywords: Agricultural Digital Scientific Research Platform; Data-Intensive Research; Knowledge Service; Platform Architecture

（责任编辑：王玮）