DOI：10.19788/j.issn.2096-6369.190207

农业生物多样性大数据平台建设研究和展望

许哲平[1,3]　邵曾婷[1]　朱学军[1]　王　昉[1]　王媛媛[1]　肖　曼[1]　马克平[2*]

（1.中国科学院文献情报中心 资源建设与知识组织中心，北京 100190；
2.中国科学院植物研究所 植被与环境变化重点实验室，北京 100093；
3. 中国科学院文献情报中心 学术期刊动态语义出版与知识服务实验室，北京 100190）

摘　要：农业生物多样性是指直接或间接用于粮食和农业的动物、植物和微生物的多样性和变异性，包括作物、家畜、林业和渔业等。它位于整个农业系统的底层，是农业生产信息化的重要内容，也是国家战略资源和国家安全的重要基础。国内外相关组织机构和项目积累了很多的资源和建设经验，但是仍然存在资源零散分布、缺乏顶层设计、元数据标准规范应用不足、平台之间的数据交互不够、数据的快速实时响应比较困难、高端农业智库及其资源平台建设薄弱等问题。为了更好地推动我国农业生物多样性大数据平台的研究和开发工作，本文着重从科学数据平台（基础数据平台、作物数据平台、家畜数据平台、林业数据平台、渔业数据平台、传统文化知识平台、智库平台和评估指标）和相关信息化基础资源对象（术语库、主题词表、元数据标准规范、本体和科研工作流）来梳理农业生物多样性大数据的国内外研究和应用进展，并从基础层、资源层、组织层和应用服务层等四个层次来提出农业生物多样性大数据平台的顶层建设框架，还针对当前现状和问题提出了建议和展望，为我国农业生物多样性大数据平台建设和资源共享服务提供参考。

关键词：农业生物多样性；作物；科学数据；本体；科研工作流；生物多样性；数据平台；农业信息化
中图分类号：S-1　　**文献标识码：**A　　**文章编号：**2096-6369（2019）02-0076-12

本文引用格式：许哲平，邵曾婷，王昉,等. 农业生物多样性大数据平台建设研究和展望[J]. 农业大数据学报,2019,1(2): 76-87.
Xu Z P, Shao Z T,Wang F,et al. Big Data Portal Development in Agrobiodiversity: Current Research and Future Outlooks [J].Journal of Agricultural Big Data,2019,1(2): 76-87.

Big Data Portal Development in Agrobiodiversity: Current Research and Future Outlooks

Xu Zheping[1,3] Shao Zengting[1] Zhu XueJun[1] Wang Fang[1] Wang Yuanyuan[1] Xiao Man[1] Ma Keping[2*]

（1. Department of Collection & Knowledge Organization Center, National Science Library, Chinese Academy of Sciences, Beijing 100190
2. State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, Beijing 100093
3. Laboratory of Dynamic Semantic Publishing of Academic Journal and Knowledge Service, National Science Library, Chinese Academy of Sciences, Beijing 100190）

Abstract: Agrobiodiversity refers to the variety and variability of animals, plants, and microorganisms used directly or indirectly for food and agriculture, including crops, livestock, forestry, and fisheries. The underpinning of the entire

收稿日期：2019-04-06
基金项目：中国科学院 A 类战略性先导科技专项(XDA19050000)；中国科学院文献情报领域引进优秀人才计划
作者简介：许哲平，男，博士，研究方向：科学数据管理与应用、生物多样性信息学；E-mail: xuzp@mail.las.ac.cn
通讯作者：马克平，男，博士，研究方向：生物多样性与生态系统功能；E-mail: kpma@ibcas.ac.cn

agricultural system and an important part of agricultural production informatization, agrobiodiversity is also the basis of national resource strategies and national security. Although data and knowledge related to agrobiodiversity have been obtained from various research projects, several problems still exist; these include scattered data, lack of top-level designs, inadequate data standards, insufficiently interoperable systems, slow response, and few high-quality think tanks and data portals. To improve research and development on a big data portal for agrobiodiversity in China, we describe advances in agrobiodiversity big data in China and abroad in terms of research data platforms (basic, crop, livestock, forestry, and fishery data platforms; traditional cultural knowledge and think tank platforms; and assessment indicators) and basic resource objects (taxonomy, thesaurus, metadata standard, ontology, and scientific research workflow). In addition, we suggest a system architecture comprising four levels, namely, basic, resource, organizational, and application levels. Finally, we provide a future outlook on the construction and resource sharing of agrobiodiversity data platforms in China.

Keywords: agrobiodiversity; crop; research data; ontology; scientific workflow; biodiversity; data platform; agro-information

1 引言

农业一直以来都是我国国计民生的基础大业。随着大数据时代的到来，农业也开始驶入信息化发展的快车道。农业大数据的涉及面非常广，包括农业生产、农业管理和农业经营等多个领域，本文主要聚焦农业生产和管理环节中的农业生物多样性领域，面向“生产信息化”。按照世界粮农组织（FAO）的定义[1]，农业生物多样性（Agrobiodiversity）指的是直接或间接用于粮食和农业的动物、植物和微生物的多样性和变异性，包括作物、家畜、林业和渔业等。它包括遗传资源多样性（品种）和用于粮食、饲料、纤维、燃料和药物的物种多样性。FAO（粮农组织）估计，全球有 5 万多种可食用的植物，但是其中仅 15 种作物就提供了全球 90%的能源需求，我们三分之二的卡路里热量由水稻、玉米和小麦三种粮食提供[2]。

农业生物多样性位于整个农业系统的起始端，是农业生产信息化的重要内容。美国农业部的数据中心（https://www.usda.gov/topics/data）在线提供了农业市场服务(AMS)、经济研究服务 (ERS)、国外农业服务(FAS)、国家农业统计服务（NASS）、自然资源保护服务（NRCS）、乡村发展（RD）和世界农业展望理事会（WAOB）等多个基础和专题数据资源，全面覆盖了农业生物多样性的各个方面。英国、法国和德国等欧盟国家也在农业大数据方面进行了诸多实践，取得了良好的效果[3]。在我国，农业生物多样性的理论和应用研究非常广泛，涉及入侵种、粮食生产、遗传资源、农业生态系统、农业生态安全等诸多领域[4-9]。而大数据研究和应用则涉及农作物资源平台、生物多样性大数据平台、平台的开放共享政策、农业大数据平台整体架构以及一系列代表性的国家级科技基础条件平台（包括国家农业科学数据共享中心、林业科学数据中心等农业领域专业数据平台等）等[10-14]。而且，我国农业部 2016 年发布的《“十三五”全国农业农村信息化发展规划》中明确提到“生产信息化是农业农村信息化的短板，亟需加快补齐”。因此，本文着重通过基础资源对象和相关科学数据平台来梳理农业生物多样性大数据的国内外研究和应用进展，并提出顶层的建设框架和未来展望，为我国农业生物多样性大数据资源建设和共享利用提供参考。

2 数据资源平台

农业生物多样性的研究工作离不开各类数据资源的支持，除了 CABI Abstracts（国际农业和生物学中心文摘数据库 http://gateway.ovid.com/autologin.html）、AGRICOLA（美国农业文献联机存取书目型数据库，https://agricola.nal.usda.gov/）和 AGRIS（国际农业科技信息系统，http://agris.fao.org/a-

gris-search/index.do ）等以文献为主的三大农林数据库之外，还包括专业领域的科学数据资源。常规的数据资源平台主要面向科研工作，为其在生物名录、病虫害管理、外来入侵物种监测、土壤管理、作物监测、家畜利用、林业资源利用、渔业资源利用、传统文化知识了解和传播等方面提供支撑。而科学数据的另外一个重要的出口就是支撑政府决策，需要将一系列的数据转换成简单明了的量化指标，为国家和全球的农业可持续发展提供支撑，这一类数据的生产者和消费者主要是专业性的智库机构及其建设的智库平台。下面是农业生物多样性相关数据资源平台的介绍。

2.1 基础数据平台

在生物名录方面，基础性名录影响力最大的是 Catalogue of Life 项目 （http://www.catalogueoflife.org/col/info/ac），其 2019 年的年度更新数据中包括 152.5 万个动物物种、38.2 万个植物物种、14 万个真菌物种以及其他门类的 3.6 万个物种，有效地支撑着现在全球生物多样性研究和应用工作的开展。在国内，由中科院生物多样性委员会主持和资助的中国生物物种名录（CoL China），自 2008 年以来每年发布更新数据，已经成为研究中国生物的权威在线平台（http://www.sp2000.org.cn）。其 2018 年的年度数据上包含 9.8 万个分类阶元（包括 8.6 万个物种和 1.2 万个种下单元）的同物异名、参考文献和空间分布信息。

在病虫害方面，EPPO Global Database（https://gd.eppo.int/）由欧洲及地中海植物保护组织（EPPO）管理和维护，目标是提供由 EPPO 收集和整理的所有害虫数据信息，包括 8.4 万个物种（植物和害虫）的基本信息（学名、异名、俗名、类群位置和 EPPO 代码），1650 个害虫的详细信息（空间分布、宿主植物和检疫状态）等。

在外来入侵物种平台方面，GRIIS（全球外来入侵种信息系统，网址：http://www.griis.org）是由欧盟和 CBD 共同资助的全球外来入侵种信息系统。中国外来入侵物种数据库（http://www.chinaias.cn/）由中国农业科学院植物保护研究所开发维护，包括 134 种植物病害、267 种动物和 352 种植物。

在土壤数据方面，合国粮农组织（FAO）、维也纳国际应用系统研究所(IIASA)和中科院南京土壤研究所等共同完成的 HWSD 世界土壤数据集（V1.2）包括 1.5 万种不同的地图单元（21600x43200 列的栅格文件）[15]。中国科学院南京土壤研究所建设的中科院南京土壤国家土壤信息服务平台（http://www.soilinfo.cn）建成了较为完整的土壤数据资源体系，整合的土壤数据涵盖土壤资源、土壤肥力、土壤环境、土壤生物等土壤学主要学科分支。

2.2 作物数据平台

国际上有许多单一作物的数据平台，如 CottonGen（棉花）、CassavaBase（木薯）、Citrus Genome Database（柑橘）、Alfalfa Breeder's Toolbox（苜蓿草基因组、基因和表型数据）。为了更好地对各类作物进行监测，全球很多国家都开展了作物监测，目前包括中国在内的主要作物监测平台包括以下 8 个[16]：

1)GIEWS（全球粮食和农业信息及预警系统，http://www.fao.org/giews，联合国项目）

2)FEWSNET（美国饥饿早期预警系统网络，http://fews.net/，美国项目）

3)MCYFS（遥感农业监测作物产量预测系统，http://agri4cast.jrc.ec.europa.eu/mars-explorer/， 欧盟项目）

4)CropWatch（作物监测， http://www.cropwatch.com.cn/，中国项目）

5)USDA-FAS（美国农业部国外农业服务，https://www.fas.usda.gov，美国项目）

6)GEOGLAM（全球作物长势信息服务，http://www.geoglam.org/ ，GEO 项目）

7)WFP Seasonal Monitor（联合国世界粮食计划署季节性监测平台，https://www.wfp.org/content/seasonal-monitor，联合国项目）

8) ASAP（农业生产异常热点监测，https://mars.jrc.ec.europa.eu/asap/，欧盟项目）

在国内，中国作物种质资源信息网 （CGRIS，http://www.cgris.net/）拥有粮食、纤维、油料、蔬菜、果树、糖、烟、茶、桑、牧草、绿肥、热作等

340 多种作物、47 万份种质的信息。中国西南野生生物种质资源库（http://www.genobank.org/）种子 10048 种（80 万份）；DNA 库 6154 种（5.5 万份），离体库 2003 种（2.3 万种），植物圃 437 种（4.6 万份），动物材料 1988 种（5.4 万份），微生物 2240 种（2.2 万份），包括野生近缘种和品种种质资源。此外，还包括独立建设维护的国家水稻数据中心（http://www.ricedata.cn/）和国家木薯产业技术体系信息平台（http://www.cassava.org.cn/）等单一作物数据平台。

2.3 家畜数据平台

国际上，联合国粮农组织（FAO）维护和开发的家畜多样性信息系统(DAD-IS，网址：http://www.fao.org/dad-is/zh/）可以查询全世界 182 个国家的 38 个不同种类的 8800 多个家畜品种的信息和图片，以及相关的其他在线资源链接，并提供了监测工具。美国的国家动物基因组研究计划（http://www.animalgenome.org），包括马、牛、羊、鸡、猪、鱼等多种动物基因组数据仓储，并提供了处理工具开发。美国农业部的经济研究服务的数据产品栏目（Economic Research Service，https://www.ers.usda.gov/data-products/）中包含了动物产品、农业经济、粮食营养、粮食安全等多种专题资源数据。此外，还有一些研究单个家畜类型的平台，如 Bovine Genome Database（牛的基因组数据，http://bovinegenome.org/）等。

在国内，国家家养动物资源实验平台（http://www.cdad-is.org.cn/）是中国农业科学院北京畜牧兽医研究所承担的科技部基础条件平台之一，覆盖猪、鸡、鸭、牛、马等多种家畜品种和鹿、狐、貂等特种动物资源信息。其子平台中国饲料数据库（http://www.chinafeeddata.org.cn/）是在中国农业科学院畜牧研究所主持下的数据共享项目，包括饲料成分表、饲料样本数据、饲料实体数据、国际饲料数据、动物需要量等专题资源。中国畜牧业信息网（http://www.caaa.cn/）由中国畜牧业协会负责建设运维，包括全国各地猪、牛、羊、鸡、鸭、鹅等家畜以及产品和饲料等市场信息。此外，各行业和地方也都建立了一批特色资源库，如云南省畜牧业协会公共服务平台（http://www.ynaaa.org.cn/）就聚焦云南本土的家畜产品和市场情况。由中国科学院武汉病毒研究所开发的病毒资源数据库（http://virus.microcsdb.cn/vri.jsp）能够检索包括人类医学病毒、动物病毒、人畜共患病毒、野生动物病毒、自然疫源性病毒、新发传染病病原、昆虫病毒、植物病毒、细菌病毒在内的多种病毒信息。

2.4 林业数据平台

联合国粮农组织（FAO）的森林业务范围包括森林管理、林产品服务、森林环境、森林管理、政策法规、评估监测、专题研究等，是一个在线知识库（http://www.fao.org/forestry/）。在基础数据方面，全球森林监测（Global Forest Watch，http://data.globalforestwatch.org/）是一个动态在线森林监测和预警系统，旨在帮助世界各地提高森林管理水平。主要包括森林变化、土地利用、保护、森林覆盖等多种数据集。

在国内，国家基础条件平台支持的国家林木种质资源平台（http://www.nfgrp.cn/）提供了种质资源、植物新品种和法律法规数据。而国家林业科学数据共享服务平台（http://www.cfsdc.org/）则整合了森林资源、湿地资源、荒漠化资源、林业生态环境、森林保护、森林培育、木材科学与技术、林业科技文献、林业科学研究专题和林业行业发展等 12 大类别的林业科学数据。中国林科院林业科技信息研究所负责建设维护的中国林业信息网（http://www.lknet.ac.cn）完善和建成了中国林业科技文献库、中国林业科技成果库、中国林业专利技术库、中国林业实用技术库、世界林业动态信息库等 80 多个拥有自主知识产权的林业科技信息数据库群。并在此基础上建成了中国工程科技知识中心中的林业专业知识服务系统（http://forest.ckcest.cn/）。除了这些大型平台之外，专业的林业信息还分布在一些特定的平台中。如中国植物主题数据库上的植物与昆虫相互关系基础信息数据库（http://www.plant.csdb.cn/hostinsects），植物与昆虫相互关系基础信息数据库以植物与昆虫之间的关系为主导，从 323 本文献

志书、著作和部分学术论文中收集数据 1.2 万条，包括 425 种植物的寄生昆虫 3961 种。

2.5 渔业数据平台

世界各国建立了很多特色的专题数据库，其中最有影响力的是世界渔业中心（World Fish Center）建立的全球最大的鱼类数据库 FishBase（https://www.fishbase.cn）。该数据包括 3 万种鱼类物种信息，包括名称、地理分布、生态系统地位和参考文献等。美国 NOAA（国家海洋和大气管理局）的渔业中心（https://www.fisheries.noaa.gov/）建立了海洋生物、海洋食品、海洋生物保护等多个专题数据。

国内的包括国家水产种质资源平台（http://zzzy.fishinfo.cn/），平台门户网站包含 33 家研究单位的 129 个数据库，标准化表达了 3.5 万条资源记录。农业科学数据共享中心的“渔业与水产科学数据分中心”（http://fishery.agridata.cn/）整合了渔业水域资源与生态特征数据、渔业物种资源与生物基础数据、渔业生物资源野外调查数据、渔业生态环境野外调查数据、水产养殖数据、捕捞渔业及管理数据、渔业装备与设施技术数据、渔业基础设施状况数据、渔业科技与经济管理数据类渔业科学数据。中国水产科学研究院主持的渔业专业知识服务系统（http://fishery.ckcest.cn），资源包括渔业种质资源包括海水和淡水共计 1377 条记录以及渔业灾害数据、渔业环境数据、水产品追溯数据。广西海洋药用资源名录（http://hyyy.gxtcmu.edu.cn/）收集了广西有分布且具有药用价值的海洋生物资源，共收录 697 种海洋药用生物，其中植物界资源 85 种，动物界资源 612 种，收录其科属、分布、采集及药用价值等信息。广东海洋大学建设的海洋生物特色数据库平台（http://210.38.136.66:3213/ocean/）收集和整理了中国鱼类、南海贝类、虾蟹类等海洋生物资源数据库。由农业农村部渔业渔政管理局承建的全国水生生物资源养护信息采集系统的公共基础数据库（http://zyyh.cnfm.com.cn/database.aspx）中包含有水产种质资源保护区数据库等多个水生和渔业数据库。此外，台湾鱼类资料库也是了解我国沿海鱼类资源的重要参考资料（http://fishdb.sinica.edu.tw/）[17]。

2.6 传统文化知识平台

农业生物多样性离不开人的活动，本地或土著居民在长期的生产生活当中与自然相处，形成了特色鲜明的农业活动和文化产物，特别是对植物的用途挖掘。如美国的 Dr. Duke's Phytochemical and Ethnobotanical Databases（杜克博士的植物化学和民族植物学数据库，https://phytochem.nal.usda.gov/）汇总了 1.3 万种全球各国的民族植物，能够利用学名或俗名查询植物的生物活性和民族植物学的用途。Native American Ethnobotany DB（美洲本土民族植物数据库，http://naeb.brit.org/）可以查询食物、药物、染料、纤维以及其他用途，共有 4029 个物种（一半为药用植物）的 4.4 万条记录。孟加拉民俗植物数据库是孟加拉的第一个民俗植物学数据库（http://www.ethnobotanybd.com/），能够从中了解植物学、民俗植物学、林业、生物化学、微生物、制药等信息。

在国内，相应的数据库资源包括民族药用植物数据库（http://www.plant.csdb.cn/herb，2.2 万条）包括各少数民族对植物药用用途、增强型药用植物数据库（6225 条，http://www.plant.csdb.cn/advherb）包括药材基源、采收和储藏、性味、功能主治等。SEADiv（东南亚植物多样性）平台上面收集整理了东南亚各国的 4213 个植物的 3.6 万条药用记录（http://www.seadiv.org/medicinal），对于了解整个东南亚的植物用途信息有很好的参考价值。

为了更好将本土文化与生物多样性的保护与利用结合起来，澳大利亚从 1997 年就开始启动 IPA（土著保护区）计划（https://www.environment.gov.au/land/indigenous-protected-areas/new-ipa-program）。截止 2018 年，澳大利亚的土著保护区计划已经在 75 个 IPA 保护区的 67 万平方公里实施，占整个保护区系统面积的 44.6%。2017-2021 年，澳大利亚政府计划拨款 1500 万美元来资助新的土著保护区计划（NIPA），资助土著居民建立新的土著保护区。在这项名为“土著土地和海洋管理项目”的平台中，用户能够了解这 75 个项目的总体情况，通过土著和传统文化的保护来促进传统文化和生物多样性的保护。

2.7 监测评估：从数据到指标

农业生物多样性大数据一个重要的应用方向是对农业可持续发展的贡献。当前，最有影响力的全球性可持续发展的指标当属联合国的 SDGs（Sustainable Development Goals，可持续发展目标，https://unstats.un.org/sdgs/）。它通过 17 个目标和 169 个指标从经济增长、社会包容和环境保护等三个目标来促进社会的可持续发展。其中，与农业密切相关的目标（http://www.fao.org/3/I9900EN/i9900en.pdf）包括 SDG 2（零饥饿）、SDG 5（性别平等）、SDG 6 (清洁饮水和卫生设施)、SDG 12（负责任的消费和生产)、SDG 14（水下生物）和 SDG15（陆地生物)。

随着世界各国不断推动 SDG 目标的实现，以前或现有的一些行动目标和框架也在逐步向 SDG 靠拢和对照。以《全球植物保护战略（2011-2020)》(GSPC，http://www.plants2020.net）为例，该行动将在 2020 年结束，为了考虑后续工作的延续性，2016 年，BGCI（国际植物园联盟）开展了如何将 GSPC 目标贡献到 SDG 的工作中，其中 SDG15 的关系最为密切，同时也能够在 SDG1、SDG2、SDG3、SDG 6、SDG 7、SDG11、SDG12 和 SDG13 中发挥重要作用。

在农业生物多样性领域，国际生物多样性中心推出的 Agrobiodiversity Index（农业生物多样性指数）是一种度量农业生物多样性和实际行动来达成多样化和可持续粮食系统的工具[18]。该指标能够帮助政府决策部门、投资商和公司，确保其在膳食和市场、生产系统和遗传资源粮食系统三个方面的多样化和可持续发展。与该指标相关的 SDG 目标包括 SDG1、SDG2、SDG3、SDG5、SDG12、SDG13、SDG14、SDG15 和 SDG16，相关的目标包括 7（可持续农业、水产养殖和林业）和 13（遗传多样性维持)。

2.8 农业科技智库与智库平台

智库是政府决策的专业智囊团。根据最新发布的《2018 年全球智库索引报告》中的说明和定义[19]，科技智库（Science and Technology Think Tanks）指在全球范围内领先的科学技术研究机构，提供前沿专业的创新研究和政策分析，研究范围包括创新研究、电信、能源、气候和生命科学领域等。《2018 年全球智库索引报告》中的粮食安全智库（Food Security Think Tanks）包括来自 57 个国家的 133 个智库，其中美国独占 29 个名额，且有不少是与农业生物多样性研究相关的智库。中国农业科学院和中国农业科学院农业资源与农业区划研究所 2 家单位入选其中。我国的智库总体上与发达国家的差距较大，但是国内有不少人开始进行了相关研究[20]。

在农业生物多样性的智库方面，不能不提的是国际农业研究磋商组织（CGIAR)。该组织创立于 1971 年，目的是通过在农业、畜牧业、林业、渔业、政策及自然资源管理等领域，开展科学研究以及与研究相关的活动，帮助发展中国家实现可持续粮食保障和减少贫困人口。CGIAR 为以非洲水稻中心(Africa Rice) 为代表的 15 个国际农业研究中心提供经费，通过高质量的科学研究促进农业的可持续发展。这 15 个中心也全部入选《2018 年全球智库索引报告》，在全球的农业生物多样性领域有着极高的专业权威性。

在数据驱动的背景下，智库往往会加大对智库平台的建设，通过多种资料和技术手段来分享专业领域的知识和专家经验。CGIAR 提出的农业大数据平台五年计划（2017-2021)，旨在从组织机构联盟、资源汇聚（涉及作物模型、时空分析、家畜专题、本体和社会经济等 6 个领域）和应用案例三个角度来进行大数据平台建设和资源共享（https://bigdata.cgiar.org/)。农业生物多样性研究平台（PAR）是 2006 年在罗马成立的公益性平台，由国际生物多样性中心等多个国际机构共同轮值管理。该平台除了提供了全球的农业生物多样性研究者和研究机构数据的注册和查询功能，还负责维护作物的生物多样性（减少病虫害损失）数据库、农业生物多样性和粮食主权的土著伙伴关系数据库和 REFARM（农业和风险管理适应框架数据库）等三个数据库。

3 信息化基础设施资源

随着共享意识的不断提高和先进采集设备的不

断部署和升级，农业生物多样性领域的数据资源越来越丰富。由于该领域的交叉性和复杂性，要想更好地对不同平台的数据发生交互、关联和揭示，除了数据资源本身之外，还需要在术语库、主题词表、知识组织体系、元数据标准规范、本体、工具和工作流平台上等基础设施资源的建设上进行更多投入，以下为相关的基础设施资源的介绍。

3.1 术语库

专业术语表是一个学科的基础。在国际上，FAOTERM（粮农组织词汇，网址：http://www.fao.org/faoterm/）是农业领域一个代表性的术语库，包括相关科学领域的专业术语，如农学、营养、生物技术、渔业、林业、食品安全等。在语种方面包含阿拉伯文、英文、西班牙文、法文、俄文和中文等多语言。

国内并没有大规模的农学术语库，只是分散在其他大型术语库中。术语在线（http://www.termonline.cn）由全国科学技术名词审定委员会主办，聚合了全国名词委会权威发布的审定公布名词数据库、海峡两岸名词数据库和审定预公布数据库累计 45 万余条规范术语，覆盖基础科学、工程与技术科学、农业科学、医学、人文社会科学、军事科学等各个领域的 100 余个学科。农业生物多样性的术语分散在土壤学、微生物学、水产、植物学、园林学、动物学等相关学科中。

还有一类术语则是以科学数据专题库的方式存在，如 Species 2000 中国节点（http://www.sp2000.org.cn）、中国植物数据库（http://www.plant.csdb.cn）和中国动物主题数据库（http://www.zoology.csdb.cn），通过以物种名称的方式来关联文献、分布地和资源用途等资料。

3.2 主题词表和知识组织体系

与术语库不同的是，主题词表能够对术语进行概念定义，并建立不同术语之间的关系，有更大的应用价值。在国际上，与该领域相关的主要叙词表包括：

（1）AGROVOC（多语种农业主题词表）：这是一部多语种结构的叙词表（http://aims.fao.org/zh-hans/agrovoc），它涵盖了农业、林业、渔业、食物安全及其他相关学科领域（例如：可持续发展、营养学等等）。AGROVOC 包含 FAO 使用的 5 种官方语言（英语、法语、西班牙语、汉语和阿拉伯语），并通过关联数据技术与其它合作词汇表保持一致，实现与外部数据源的关联。这些外部数据包括中文农业主题词表、水科学和渔业文摘（ASFA）和 DBpedia 等在内的 16 家外部数据源的 5 万余条关联记录。AGROVOC 还是农业本体服务（AOS）项目发展的基础。

（2）NALT（美国国家农业图书馆叙词表）：这是一个英语和西班牙语的在线双语叙词表（https://agclass.nal.usda.gov/），主要收录了农业、生物及相关领域的 25.5 万个术语（英语 13.9 万个，西班牙语 11.6 万个），可通过 17 个主题分类进行浏览。其中，将生物类群的名称库作为一个主题是其特色。NALT 还与 GBIF 开展了词表的空缺分析[21]，目标是希望从 GBIF（全球生物多样性信息机构）中获取有关农业生物多样性数据相关的统计信息、可视化变化趋势和为农业研究者处理 GBIF 数据提供案例和代码。

（3）EUROVOC（欧盟农业主题词表）：是由欧盟管理维护的多语主题词表表，涉及 21 个一级领域（法律、经济、贸易、环境、农林渔业等）和 127 个二级领域，包括德语、法语、英语、爱沙尼亚语、希腊语、保加利亚语、西班牙语、捷克语、丹麦语、意大利语等。该词表也提供了关联数据版本的浏览和 SKOS 文件下载（https://eur-lex.europa.eu/browse/eurovoc.html）。

（4）MeSH（医学主题词表）：这是美国国立医学图书馆编制的权威性主题词表，是一套生物医学领域的主题词表，包含生物类群名称在内的 19 个主题（与农业生物多样性相关的包括疾病、健康、粮食、农业等次级主题，https://www.ncbi.nlm.nih.gov/mesh/1000048）。提供 XML、TEXT、MARC21 和 RDF 等多种下载格式。国内的万方数据在 MeSH 2011 版本的基础上，通过人工翻译，对其医学类期刊论文、学位论文、会议论文和外文期刊论文进行

标引，从而实现 MeSH 主题词检索功能（http://old.med.wanfangdata.com.cn/Mesh/Mesh.aspx）。

在国内，英文词表的代表是科技知识组织体系(STKOS，http://stkos.las.ac.cn/)。该平台是建设以领域本体为目标的超级词表，包括来源术语、来源词表、科技术语、STKOS 规范概念、范畴类和范畴表等 6 大类数据模型。目前平台共收录 61.5 万学术概念，232.1 万个术语，并对外提供了 API、本体可视化和关联数据等多种功能。其中与农业生物多样性相关的术语分布在生物学、植物学、动物学、农学、林业科学、畜牧科学、水产等领域，共涉及概念超过 20 万多条概念。中文主题词表的代表是《中国分类主题词表》。该词表 Web2.1 版（http://cct.nlc.cn/）共收录分类法类目 5.3 万个，主题词 11 万条。农业生物多样性的术语分散在土壤学、微生物学、水产、植物学、园林学、动物学等相关学科中。

农业科学叙词表是国内一部大型、综合性农业叙词表，共收录了包括农业、林业、生物等领域在内的 6 万多个叙词和非叙词[22]。

3.3 元数据标准规范

在农业生物多样性领域，一类元数据面向数据交互和整合，如农业元数据元素集（AgMES）旨在涵盖农业领域有关不同类型信息资源的描述、资源发现、互操作性和数据交换的语义标准问题。还有一类是面向专业领域数据描述的，包括多作物护照描述符（MCPD）和 Darwin Core 的 Germplasm 扩展等。

（1）Multi-Crop Passport Descriptors (MCPD，多作物护照描述符)标准是由国际生物多样性中心（Bioversity International）和 FAO（粮农组织）联合开展的研究工作成果，旨在方便种质资源的信息交换，同时能够兼容 FAO WIEWS（全球信息和预警系统）、PGR（植物遗传资源）和 GENESYS 全球门户的数据结构。该标准馆藏信息、采集信息、类群信息、捐赠信息、育种信息等 28 类元素。

（2）Darwin Core 及其扩展：Darwin Core 是一套用于描述生物有机体分布及其相关采集信息的规范（http://rs.tdwg.org/dwc/），并能够根据物种形态特征、古生物化石记录、参考文献、多媒体资源和种质资源等内容进行扩展。为了更好地为农业生物多样性数据整合提供支撑，GBIF 和国际生物多样性中心在 2016 年联合成立了“面向农业生物多样性领域的数据适用性任务组”来帮助原地、迁地和田间的植物数据进行交互，并将 MCPD 标准整合到 Darwin Core 的 Germplasm（种质资源）扩展中，使其能够整合到全球的生物多样性观测数据中来。

（3）国内元数据规范：国家作物种质资源数据中心针对国家作物种质资源数据中心、土壤质量数据中心、农业环境数据中心、植物保护数据中心、畜禽养殖数据中心、动物疫病数据中心、农用微生物数据中心、渔业科学数据中心、天敌等昆虫资源数据中心、农产品质量安全数据中心的监测制定了相应的农作物种质资源标准和规范，包括蔬菜、粮食作物、牧草绿肥、果树、经济作物、主要热带作物等几大类（http://www.cgrchina.cn/?page_id=12112）。更多的农林数据标准则分散在中国林业标准数据中心（http://www.lknet.ac.cn/lybz.htm）和中国农业标准网（http://www.chinanyrule.com/）两个网站中。

3.4 本体

目前为止，农业生物多样性的本体已经非常多，主要包括以下几种类型：

（1）用于描述非生物环境描述：包括 Environment Ontology（ENVO，环境本体)和 EMP Ontology (EMPO，地球微生物项目）等。

（2）用于农业数据交互：如 FAO 基于 AGROVOC 主题词表的开发建立的农业本体服务（AOS）和渔业本体（Fishery Ontology）、作物 - 有害生物本体和抗菌剂本体等等[23]。

（3）用于普适性生物特征描述：包括 Gene Ontology（基因本体）、Plant Phenology Ontology (植物物候本体)、Plant Ontology（植物本体）和 Crop Research Ontology（作物研究本体）等。

（4）用于特定物种或品种描述：包括 Rice Ontology（水稻本体）、Wheat Ontology（小麦本体）和 Banana Ontology（香蕉本体）等。

越来越多的词表和本体被用来表述和标注农业

数据。但是，这些本体都是以不同格式、不同大小和不同结构零散分布的。因此，有必要将其以相同的格式存放在同一个平台方便使用。这其中最典型的项目是 AgroPortal（http://agroportal.lirmm.fr），它的目标是重复利用生物医学的领域本体，包括植物、农业、粮食和生物多样性科学，为各类本体提供统一调用的仓储平台[24]。通过对 AgroPortal 平台上的 35 个项目实例（190 种不同的本体，最多的是 VEST-AgroPortal Map of Standards，使用了 52 种本体）项目，对后台使用支撑的本体进行统计，可以得到使用次数超过 10 次或以上的本体，分别是：Gene Ontology（基因本体，18 次）、Plant Ontology（植物本体，12 次）、Sequence Types and Features Ontology（序列类型和特征本体，11 次）和 Plant Trait Ontology（植物性状本体，11 次）。

3.5 数据分析和工作流

在农业生物多样性的科研工作中，积累了大量的数据挖掘和分析的算法和模型。Angelo Signore 利用 Open Data Kit 和 Google Fusion Table 来根据不同蔬菜作物的基因流失的情况进行数据收集，并将其保存到 Google App Engine 平台中，然后将这样的信息转换到 Google Fusion Table 中以便进行后期制图[25]。而随着 KNIME（https://www.knime.com/）、RapidMiner（https://rapidminer.com/）和 Taverna（https://taverna.apache.org/）等工作流平台的发展，使数据挖掘和分析能够以流程化和整体化的方式进行。A.T.M Shakil Ahamed 等利用 RapidMiner 数据挖掘软件来实现 K-Means 聚类算法分析，并对孟加拉特定地区的主要农作物产量进行预测[26]。

上述基于工作流的数据分析平台都还是以桌面版为主（KNIME 和 RapidMiner 虽然有服务器版，但是商用付费的），往往有一定的局限性且缺乏共享和沟通。因此，在欧盟第七框架计划的资助下，Biovel 平台（http://www.biovel.eu/）开始上线服务，能够对 Taverna 软件的项目文件进行上传、下载和重用。myexperiment（http://ww.myexperiment.org）平台则比 Biovel 平台的内容更丰富，总计有 2905 个项目文件可在线共享。它不仅能够分享 Taverna 的工作流项目，还能分享 RapidMiner、Galaxy、KNIME 和 Kepler 等多种工作流项目。

4 农业生物多样性大数据平台框架设计

从上述的当前农业生物多样性大数据相关的发展情况来看，数据资源平台和信息化基础设施资源已经有了一定的发展，在实际科研工作中发挥了一定的支撑作用。但是基础数据的深度和广度还有待增强，在监测指标、智库平台和工作流平台等高阶系统的研发方面还有很大空缺和发展空间，不同平台之间的数据交互、整合和定制服务还亟待加强。我国除了存在上述问题之外，还存在资源零散分布、缺乏顶层设计、元数据标准规范应用不足和数据的快速实时响应比较困难等问题。基于此，本文特提出我国的农业生物多样性大数据平台建设的体系框架，总体上包括四个部分：

（1）基础层：主要是数据的来源方，包括各类研究机构（也包括标本馆、种质资源库、种植基地和野外观测台站等）、政府部门、国内外项目课题(以及在合作中建立与国际组织机构的合作关系）以及公民科学平台和个人（专业人员和爱好者）。这些来源方除了贡献数据之外，还要对贡献的数据进行明确的产权声明（如 CC0、Open Government License、Public Domain Dedication and License 等），避免后期使用的时候产权不明带来的使用纠纷。在这过程中，不同层次主体单位数据政策的执行和落实也至关重要，如国务院发布的《科学数据管理办法》、中国科学院发布的《中国科学院科学数据管理与开放共享办法(试行)》和国家农业科学数据中心的《农业科学数据汇交管理办法》等。

（2）资源层：这是整个平台的资源核心，包括资源的汇聚、整合、规范化和挖掘分析。一方面是从基础层获取汇聚不同类型的资源，包括出版物、标本 / 种质资源、野外观测 / 调查以及实验室数据等。另一方面，将这些资源按农业生物多样性的四个主题（植物、动物、微生物和人类活动带来的农林牧渔行业）进行梳理分类。加强现有数据集的描述规范对接 AgMES、Darwin Core、EML 和 MCPD

等国际标准规范。

（3）组织层：主要通过各种术语表、主题词表、关联数据、本体、元数据标准规范、程序代码、科研工作流以及相关新技术和新方法来对资源层的数据进行标引和重新加工，以便根据个性化的需求对资源和数据进行加工和包装，来支撑和应对应用服务层的不同需求。加强人工智能、知识图谱、NLP（自然语言处理）、语义网等新的 IT 技术在数据组织和挖掘分析中的应用，激活和优化现有的数据资源。

（4）应用服务层：主要是面向科研工作、政府管理决策、专家智库支撑和大众科普教育的实际应用场景，对已经通过组织层梳理好的结构化和标准化数据来进行组配，形成新的产品和专题资源。加强多样化服务模式的研究和应用推广，包括 API 共享、移动端服务（APP 和微信公众号）嵌入和数据出版等。同时，加强与国内外相关数据资源需求平台的合作共享，如全国农技推广信息平台等。加强对整个农业生产、农业市场和农业管理等的支撑与融合，提升数据资源的快速响应和服务能力。

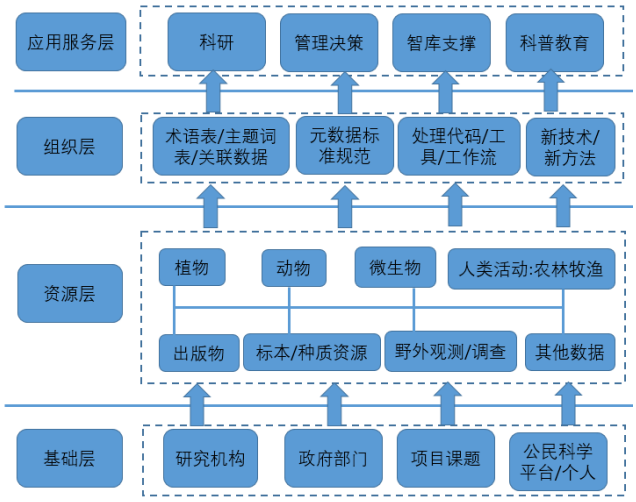

图 1　农业生物多样性大数据平台框架设计

Fig.1　The Architecture Design of Big Data Portal of Agrobiodiversity

5　总结和建议

我国农业生物多样性大数据建设虽然取得了一定的成绩，但是仍然存在数据共享程度不够、数据缺乏统一的标准规范、精度低、数据利用率低、数据流动性差、交易困难等问题。结合当前国际前沿趋势以及我国的实际国情，特提出如下展望和建议：

（1）加强国家层面的顶层设计和跨部门的合作，如科技部基础条件平台现有数据平台的打通，还包括中科院、农科院、林科院、环保部、林业局、海洋局等相关部委的开放和联合。这其中还需要配套办法和政策的推行，光靠一个《科学数据管理办法》和宏观规划难以达到目标，需要政府从上至下的推动，毕竟很多农业大数据都还是集中在政府管理部门。

（2）加强对数据的规范化处理和基于数据生命周期的平台研发，通过服务来促进建设和共享。一方面加强对历史数据的规范化梳理，参考和遵循 FAIR（可检索、可访问、可交互和可重用）原则，使其可用性增加。另一方面加强对新数据的整合，特别是各类实时传感数据的整合，加快数据处理和响应时间（特别是病虫害和入侵种等）。农业生物多样性大数据主要还是属于生产信息化的范畴，除了能够提升农业生产精准化和智能化水平之外，还需要融入大的“生态系统”来，加强同农业管理、经营和服务方面的协同，为农户、企业或信息服务商提供基础数据。

（3）通过多种方式加大国际合作力度。总的来说，我国的农业生物多样性数据的国际化参与不足，特别是高端农业知识的服务和高端智库尤为不足。在信息技术和专业领域两个方面，加强同国际农业研究磋商组织（CGIAR）及其 15 个国际农业研究中心的联合与合作，学习技术、专业知识和经验。加大项目管理开放力度，吸引全球人才加强对我国农业生物多样性问题的研究，特别是 SDGs 相关的前瞻性课题的布局和支持。笔者曾就研究中国 SDG 的文章做过分析，在 800 余篇英文文章中，除了中国作者之外，美国、澳利大亚、加拿大和日本等国家的作者也分别有 30 篇以上的文章提及中国的 SDG 研究工作。通过对中国案例研究，可以让更多人了解中国，良好的中国案例也能更好地推向全球，提升中国的国际影响力。

（4）加强对企业创新的支持和数据科学人才培养。农业生物多样性作为一门应用性学科，有着很

大的商业市场，相关信息技术型企业也非常多，对数据的需求也非常大。但是，由于开放性和规范化不足，导致利用率很差。因此，未来需要大力扶持和培育数据驱动型的企业进行创新活动，加大对基础数据的利用和挖掘，并在这个过程中不断培养数据科学人才，增强数据科学人才的数据搜寻、获取和分析能力。

参考文献

[1] FAO. 1999a. Agricultural Biodiversity, Multifunctional Character of Agriculture and Land Conference, Background Paper 1. Maastricht, Netherlands. 1999.

[2] Gruber, Karl. Agrobiodiversity: The living library.Nature. 2017, 544,S8,DOI:10.1038/544S8a.

[3] 黎玲萍, 毛克彪, 付秀丽 , 等. 国内外农业大数据应用研究分析[J]. 高技术通讯, 2016,26(4):414-422.
Li L P, Mao K B, Fu X L , et al. Analysis of the Research on Agricultural Big Data Applications at Home and Abroad[J]. Chinese High Technology Letters, 2016,26(4):414-422.

[4] 李明, 彭培好, 王玉宽, 等. 农业生物多样性研究进展[J]. 中国农学通报, 2014, 30(9):7-14.
Li M, Peng P H, Wang Y K, et al. Progress of Agrobiodiversity Research [J]. Chinese Agricultural Science Bulletin, 2014, 30(9):7-14.

[5] 万方浩,郭建英,王德辉.中国外来入侵生物的危害与管理对策[J].生物多样性,2002(01):119-125.
Wang F H, Guo J Y, Wang D H. Alien invasive species in China: their damages and management strategies [J]. Biodiversity Science, 2002(01):119-125.

[6] 章家恩, 饶卫民. 农业生态系统的服务功能与可持续利用对策探讨[J].生态学杂志,2004(04):99-102.
Zhang J E, Rao W M. Discussion on Agroecosystem Services and Sustainable Utilization [J]. Chinese Journal of Ecology, 2004 (04):99-102.

[7] 张金萍,张保华,刘衍君,等.中国农业生态安全及相关研究进展[J].世界科技研究与发展,2005(02):42-46.
Zhang J P, Zhang B H, Liu Y J, et al. The Progress of Chinese Agri-ecological Security and Its Correlative Research[J]. World SCI-TECH R&D,2005(02):42-46.

[8] 李琴,陈家宽.长江大保护事业呼吁重视植物遗传多样性的保护和可持续利用[J].生物多样性,2018,26(04):327-332.
Li Q, Chen J K. The primary task of watershed-scale comprehensive conservation of Yangtze River Basin: Conservation and sustainable utilization of plant genetic diversity[J]. Biodiversity Science,2018,26(04):327-332.

[9] 唐晓玲, 刘振湘. 我国家养动物多样性现状与持续利用对策[J]. 畜禽业, 2001(7):6-7.
Tang X L, Liu Z X. Diversity and Sustainable Utilization of Domestic Animals in China. Livestock and Poultry Industry [J], 2001(7):6-7.

[10] 姜侯,杨雅萍,孙九林. 农业大数据研究与应用[J]. 农业大数据学报, 2019, 1(1): 5-10.
Jiang H ,Yang Y P, Sun J L. Research and Application of Big Data in Agriculture[J]. Journal of Agricultural Big Data, 2019, 1(1): 5-10.

[11] 周国民. 我国农业大数据应用进展综述[J]. 农业大数据学报, 2019, 1(1): 16-23.
Zhou G M. Progress in the Application of Big Data in Agriculture in China [J]. Journal of Agricultural Big Data, 2019, 1(1): 16-23.

[12] 赵瑞雪,赵华,朱亮. 国内外农业科学大数据建设与共享进展[J]. 农业大数据学报, 2019, 1(1): 24-36.
Zhao R X, Zhao H, Zhu L. Progress in the Development and Sharing of Big Data in Agricultural Science between China and Foreign Countries [J]. Journal of Agricultural Big Data, 2019, 1(1): 24-36.

[13] 曹永生,方沩.国家农作物种质资源平台的建立和应用[J].生物多样性,2010,18(05):454-460.
Cao Y S, Fang W. Establishment and application of National Crop Germplasm Resources Infrastructure in China [J]. Biodiversity Science,2010, 18(05):454-460.

[14] 马克平,朱敏,纪力强, 等. 中国生物多样性大数据平台建设[J]. 中国科学院院刊,2018,33(8):838-845.
Ma K P, Zhu M, Ji L Q, et al. Establishing China Infrastructure for Big Biodiversity Data[J]. Bulletin of the Chinese Academy of Sciences.2018, 33(8) : 838-845.

[15] F Fischer G F, Nachtergaele S, Prieler H T V L, et al. Global Agro-ecological Zones Assessment for Agriculture (GAEZ

2008) [M]. IIASA, Laxenburg, Austria and FAO, Rome, Italy. 2008.

[16] Steffen F, Linda S, Juan C L B, et al. A Comparison of Global Agricultural Monitoring Systems and Current Gaps [J]. Agricultural Systems. 2019, 168:258-272.

[17] 邵廣昭. 臺灣魚類資料庫網路電子版(http://fishdb.sinica.edu.tw, 2019-5-12).
Shao G Z. Fish Database in Taiwan (http://fishdb.sinica.edu.tw, 2019-5-12).

[18] Bioversity International. The Agrobiodiversity Index: Methodology Report v.1.0 [R]. Bioversity International, Rome, Italy.2018.

[19] McGann, James G. 2018 Global Go To Think Tank Index Report[R]. 2019.

[20] 梁丽.中国国家级农业智库建设研究. 北京:中国农业科学院.2018.
Liang L. Research on Construction of China National Agricultural Think Tank[D]. Beijing: Chinese Academy of Agricultural Sciences. 2008.

[21] Akshat Pant. Gap Analysis of Agrobiodiversity data in GBIF and the NAL Thesaurus. Ag Data Commons.2018. http://dx.doi.org/10.15482/USDA.ADC/1466041.

[22] 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.
Xian J L, Zhao R X, Kou Y T, et al. Study and Practice on Converting and Publishing Chinese Agricul tural Thesaurus as Linked Open Data[J]. New Technology of Library and Information Service, 2013, 29(11): 8-14.

[23] 钱平,孟宪学,郑业鲁,等.中国农业本体服务的初步研究[J]. 农业网络信息.2009 ,8:5-8.
Qian P, Meng X X,Zheng Y L, et al. Preliminary Study on Agricultural Ontology Services in China [J]. Agricultural Network Information. 2009, 8:5-8.

[24] Clément Jonquet, Anne Toulet, Elizabeth Arnaud, et al. AgroPortal: A vocabulary and ontology repository for agronomy[J]. Computers and Electronics in Agriculture.2018,144: 126-143.

[25] Angelo Signore. Mapping and sharing agro-biodiversity using Open Data Kit and Google Fusion Tables[J]. Computers and Electronics in Agriculture.2016,127: 87-91.

[26] A.T.M Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, et al. Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh [C]. 2015IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. 2015,1-6.