



文献 CSTR:
32001.14.11-6035.nasdc.2025.0140.zh



文献 DOI:
10.11922/11-6035.nasdc.2025.0140.zh

数据 DOI:
10.57760/sciencedb.j00001.01587

文献分类: 图书馆、情报与文献学

收稿日期: 2025-07-25

修回日期: 2026-03-24

录用日期: 2026-03-27

发表日期: 2026-06-15

水稻育种抗旱、耐盐与低镉积累科学实验事件数据集

林巧^{1#}, 鲜国建^{1,2#}, 胡智杰¹, 吴保¹, 丁南陵¹, 张学福^{1,2*}, 孙坦^{2,3*}

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
3. 中国农业科学院, 北京 100081

共同第一作者

摘要: 随着数据驱动技术在生物育种领域的广泛应用, 构建结构化、标准化的科学实验数据资源对于揭示科研活动规律、支撑智能育种决策具有重要意义。为提升水稻在非生物胁迫环境下的育种效率, 本研究构建了水稻育种抗非生物胁迫领域科学实验事件数据集。研究基于 Web of ScienceTM和 PubMed Central 数据库采集 1990–2023 年间相关文献, 结合 LightRAG 检索增强模型与改进的 BERT 语言模型, 抽取了品种, 基因, 蛋白质, 组织、细胞和器官, 环境因素, 表型, 遗传育种技术, 分子育种技术, 计算机技术 9 类实体及包含, 转录或翻译, 表达调节, 相互作用, 组织、细胞和器官发育调节, 影响, 技术合作, 技术支撑 8 类关系, 并构建具有时间与学科属性的五元组事件结构, 覆盖 10,409 条实体和 13,034 条实体关系。数据集以“主语-关系-宾语”结构呈现事件过程, 划分为属性类、功能类、因果类与技术类 4 类事件类型。在数据构建过程中, 采用标准化清洗、语义标注、专家审核与多模型对比评估等质量控制手段, 确保数据的准确性和一致性。相较于传统科研数据资源, 本数据集突出科学实验中的动态关系与技术路径表达, 能够为水稻功能基因挖掘、性状改良路径重构、知识图谱构建与智能问答等系统提供关键支撑, 具有较高的专业价值与推广意义。

关键词: 水稻育种; 非生物胁迫; 科学实验事件; LightRAG 模型; BERT 模型

数据库(集)基本信息简介

数据库(集)名称	水稻育种抗旱、耐盐与低镉积累科学实验事件数据集
数据通信作者	张学福 (zhangxuefu@caas.cn); 孙坦 (suntan@caas.cn)
数据作者	林巧、鲜国建、胡智杰、吴保、丁南陵、张学福、孙坦
地理区域	全球
数据时间范围	1990–2023年
数据量	982 KB
数据格式	*.csv
数据服务系统网址	https://doi.org/10.57760/sciencedb.j00001.01587
基金项目	国家社会科学基金 (23BTQ054、22BTQ079)
数据库(集)组成	数据集共包括2个数据文件, 其中: (1) 实体 (entity) 表中包含3个字段; (2) 关系 (relation) 表中包含11个字段。

* 论文通信作者

张学福: zhangxuefu@caas.cn

孙坦: suntan@caas.cn

引言

种业是国家战略性、基础性核心产业，是现代农业发展的“芯片”。水稻是全球最重要的粮食作物之一，其产量和品质直接关系到粮食安全与社会稳定^[1]。非生物胁迫，如干旱、盐碱、极端温度、重金属污染、强光辐射和营养缺乏等，严重影响了水稻的生长、发育和产量。因此，提高水稻对非生物胁迫的抗性，已成为当前育种领域的研究重点与热点问题之一。随着高通量测序、表型组学及人工智能等技术的发展，数据驱动的生物育种正在加速演进。《“数据要素×”三年行动计划（2024—2026年）》^[2]中明确提出，要聚焦生物育种等研发领域，深入挖掘各类科学数据和科技文献，通过细粒度知识抽取和多来源知识融合，推进跨学科、跨领域协同创新。在此背景下，构建与重用具有结构化知识的科研数据资产，成为提升科研效率与育种智能化水平的关键手段。

在科学研究的信息抽取任务中，事件被定义为文本中具有特定语义的动作或变化过程，通常由触发词（标识事件的动词或动词短语）、与事件相关的实体以及其他属性构成。在科学研究中，一篇篇科学文献作为成果性产物被看作是不同科学事件，是对科学活动过程的陈述^[3]。然而，已有科学事件多侧重于对单一语义单元或局部知识片段的刻画，通常缺乏对实验过程要素及其关联关系的系统建模，难以反映科研实践中的多要素协同与动态演化特征。

本研究在科学事件的基础上进一步定义了科学实验事件（Scientific experimental events, SE）的概念，指的是由科学研究中影响创新的各类数据或技术实体及其关联关系构成的结构化集合，用于描述科学实验过程中涉及的关键创新活动及其相互作用。与一般科学事件相比，科学实验事件更强调以“实验过程”为核心，将多类型要素及其关系进行结构化组织。因此，一个完整的科学实验事件包含主实体、主实体类型、关系类型、客体实体、客体类型的五元组结构，并附带时间与学科两个属性特征，具备层次性、方向性与动态性等特征，能够有效描绘科学实验活动的逻辑进展与创新路径。

信息抽取（Information Extraction, IE）是研究自动提取自然语言文本中有效信息的技术，包括实体抽取和关系抽取等。近年来，人工智能的迅猛发展推动了情报技术与方法的深刻变革，大量智能化信息抽取模型的引入显著提高了情报工作的效率和准确性^[4]。实体抽取也称作命名实体识别（Named Entity Recognition, NER），其目标在于从自由文本中识别出有意义的实体边界和类别，以实现快速获取重要信息^[5]。近年来，知识抽取方法不断演进，已形成基于规则^[6-7]、基于概率图^[8-9]、基于神经网络^[10-11]及基于预训练模型^[12-13]4 大类方法体系。其中，基于规则的方法适用于结构化文本但泛化能力有限；基于 CRF 等的概率图方法可处理上下文信息但计算成本高；基于神经网络如 LSTM、CNN、RNN 的方法在特征学习和语义表达方面表现出色，但对数据依赖性强。随着深度学习的发展，预训练模型（如 BERT、ALBERT、RoBERTa 等）成为当前主流方法，具备强大的上下文理解与知识迁移能力，尤其适合处理长文本中的实体和关系抽取任务。

然而，现有农业科研数据资源与信息抽取数据集仍存在一定局限。一方面，当前主流农业数据库（如 Gramene 和 Oryzabase 等）多以基因、性状或品种等实体信息为核心，侧重于静态知识的组织与存储，缺乏对科学实验过程及其内在逻辑关系的表达，难以支撑对科研活动动态演化规律的深入分析^[14-15]。另一方面，尽管信息抽取领域已构建了如 GENIA 语料库和 BioNLP Shared Task 数据集等科学事件数据资源，但这些数据集主要面向生物医学或通用科学领域，缺乏针对农业育种场景的语义建模，且通常未引入时间属性及技术要素关系，难以反映现代育种过程中多技术融合与多要素

协同的特征^[16-17]。此外，现有数据集多服务于模型训练与评测任务，对科学实验过程建模与创新路径表达的支持仍较为有限。因此，亟需构建面向农业育种领域的科学实验事件数据集，以“事件”为核心单元，对科研过程中涉及的实体、关系及其动态交互进行结构化表达，从而揭示科学实验中的关键创新要素配置模式与技术演进路径。

本研究面向水稻育种抗非生物胁迫领域，提出并构建了一个覆盖耐盐、耐旱、镉积累等关键性状的科学实验事件数据集。基于 Web of Science™ 核心合集和 PubMed Central (PMC) 数据库获取研究型科技论文全文数据，利用嵌入大语言模型的 LightRAG 检索增强框架与改进的 BERT 深度学习模型，结合国家水稻数据中心抽取得到品种，基因，蛋白质，组织、细胞和器官，环境因素，表型，遗传育种技术，分子育种技术，计算机技术共 9 类实体，以及包含，转录或翻译，表达调节，组织、细胞和器官发育调节，相互作用，影响，技术合作，技术支撑共 8 类关系，最终构建包含 10,409 条实体及 13,034 条实体关联关系的科学实验事件数据集，涵盖属性类、功能类、因果类、技术类 4 大类事件类型。本数据资源不仅具有较高的重用价值，可服务于水稻功能基因挖掘、性状改良路径分析与政策决策支持，更为实现从分子育种向智能设计育种转型提供了坚实的数据基础。

1 数据采集和处理方法

1.1 数据采集方法

在数据采集阶段，本数据集选取 Web of Science™ 核心合集 Science Citation Index Expanded (SCIE, 科学引文索引) 数据库，以及 NCBI 数据库旗下的 PMC 数据库作为检索来源。基于水稻育种的相关研究^[18]，在领域专家指导下制定布尔检索表达式，在 SCIE 中检索主题涵盖水稻抗旱、抗盐、低镉积累等非生物胁迫研究的文献。检索式如下：TS=(((drought* or arid* or salt* or salin* or alkali* or cadmium) near/5 (resist* or toleran* or prevent* or anti) or (low* near/5 cadmium) or cadmium-free or anti-drought* or anti-arid* or anti-salt* or anti-alkali* or (cadmium near/5 accumulat*)) and (rice or paddy or "oryza sativa"))。截至 2023 年 9 月 26 日，在 SCI 数据库中共检索到相关论文 9,060 篇。

数据采集流程如下：（1）对 9,060 篇水稻抗非生物胁迫论文的文献类型进行筛选，得到研究型 (Article) 论文 8,293 篇；（2）根据 DOI 号匹配，在 PMC 数据库中获取水稻抗非生物胁迫研究型论文，共匹配得到 1,903 篇全文数据的 pmc 格式压缩包；（3）对压缩包中 NXML 格式全文进行预处理（分段、去除标签等），得到 json 格式的全文；（4）对 1,903 篇 json 格式的全文进行重要信息批量提取，提取表征科学实验结果的 Abstract 部分与表征科学实验过程的 Materials and methods、Methods、Experimental Section 部分，无实验过程的文章则删除，共提取 1,639 篇文章的实验信息并存储到 TXT 文件中；（5）对提取出的数据再次进行人工筛选和校正，最终筛选出 861 篇水稻育种抗非生物胁迫领域全文的摘要和材料与方法部分文本，作为水稻育种抗非生物胁迫领域科学实验事件抽取语料库。

1.2 数据处理方法

数据处理阶段包括语料预处理、标注语料构建、事件抽取建模与结构化输出等关键步骤，形成“主语-关系-宾语”结构的科学实验事件数据，进而构建支持科研活动知识发现与创新要素配置分析的数据基础。整体流程如下：

- （1）文本清洗与格式转换

基于从 PMC 数据库获取的 NXML 格式全文数据，研究首先对 Abstract、Materials and Methods、Methods 与 Experimental Section 章节进行批量提取，并保存为纯文本格式（图 1）。随后开展标准化清洗操作，包括去除 HTML 标记、符号统一（如引号、连字符等）、分句处理、去除空格与特殊字符等。此过程确保了语料输入质量，为后续事件抽取任务提供了干净、可控的输入文本基础（图 2）。

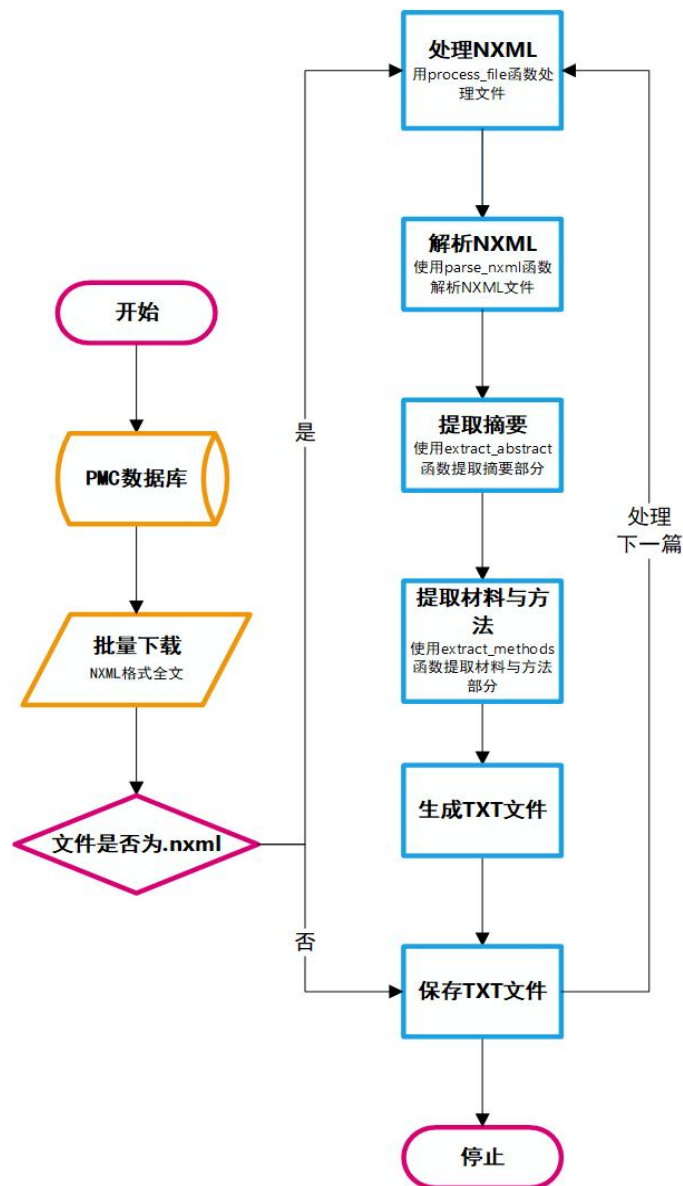


图 1 NXML 格式全文关键章节提取流程

Figure 1 Key section extraction workflow from full-text NXML format

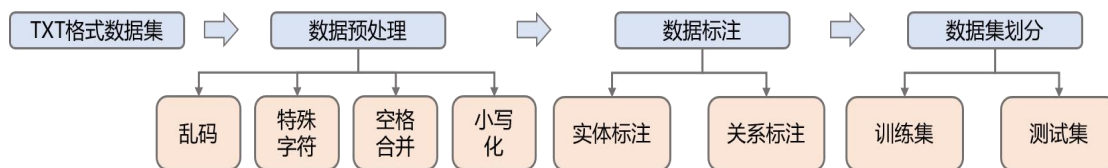


图 2 数据集规范化预处理流程

Figure 2 Dataset normalization and preprocessing workflow

(2) 实体与关系定义

为规范化语义标注与事件建模，结合水稻育种科研领域的专业语义特征，构建了科学实验事件语义结构体系，定义了 9 类实体（表 1），以及 8 类实体关系（表 2），并根据主实体类型和关系类型进一步将科学实验事件划分为属性类、功能类、因果类、技术类 4 大类（表 3）。

表 1 水稻育种领域实体定义

Table 1 Entity definitions in rice breeding

实体类型	符号	释义	示例
品种	VAR	描述经过人工选择和培育，具有特定优良性状的一类植物或作物类型	IR8; shanyou63
基因	GEN	描述编码 mRNA 的 DNA 序列，基因名称通常采用斜体和大写字母表示	FUS3; APETALA2
蛋白质	PRO	描述 RNA 的产物，蛋白质名称通常采用普通数字和大写字母表示	CLV3; LEC1
组织、细胞和器官	TIS	描述一组共同执行特定功能的细胞，通过多个组织的功能分组形成器官	root; stem
环境因素	EF	描述影响生物体的生物或者非生物因素（如温度、光照）	salt stress
表型	PHE	描述生物体表现出的可观察或可测量的性状	salt tolerance
遗传育种技术	TT	描述通过遗传学方法改善作物品种的技术	hybridize
分子育种技术	MT	描述利用分子生物学方法进行育种改良的技术	gene editing
计算机技术	CT	描述通过计算机方法辅助育种过程的技术	deep learning

表 2 水稻育种领域实体关联关系

Table 2 Entity relationships in rice breeding

关系类型	主实体	客实体	释义
包含	品种	基因/蛋白质	描述品种中包含基因或者蛋白质
转录或翻译	基因	蛋白质	描述将 DNA 编码形成 RNA，或者将 RNA 编码形成蛋白质
表达调节	基因/蛋白质	组织、细胞和器官	描述基因或蛋白质在组织、细胞和器官中的表达
相互作用	基因/蛋白质	环境因素	描述基因、蛋白质与环境之间的相互作用关系
组织、细胞和器官发育调节	环境因素	组织、细胞和器官	描述不同环境因素对组织发育活动的影响
影响	基因/蛋白质/品种/环境因素	表型	描述基因、蛋白质、品种或环境因素对表型的影响
技术合作	遗传育种技术/分子育种技术/计算机技术	计算机技术/遗传育种技术/分子育种技术	描述遗传育种技术、分子育种技术和计算机技术之间的合作关系，按照在科学实验流程中的技术使用先后顺序确定主实体和客实体
技术支撑	遗传育种技术/分子育种技术/计算机技术	表型/基因/品种/蛋白质/组织、细胞和器官	描述遗传育种技术、分子育种技术或计算机技术对品种、基因、蛋白质、组织、细胞和器官、表型的技术支撑

表 3 水稻育种领域科学实验事件类型

Table 3 Types of scientific experimental events in rice breeding

科学实验事件类型		主实体	关联关系	客实体
属性类	材料包含事件	品种	包含	基因/蛋白质
功能类	材料翻译事件	基因	转录或翻译	蛋白质
	材料表达事件	基因/蛋白质	表达调节	组织、细胞和器官
	环境调节事件	环境因素	组织、细胞和器官发育调节	组织、细胞和器官
	相互作用事件	基因/蛋白质	相互作用	环境因素
因果类	材料影响事件	基因/蛋白质/品种	影响	表型
	环境影响事件	环境因素	影响	表型
技术类	湿实验合作事件	遗传育种技术/分子育种技术	技术合作	遗传育种技术/分子育种技术
	干实验合作事件	计算机技术	技术合作	计算机技术
	干湿合作事件	计算机技术	技术合作	遗传育种技术/分子育种技术
		遗传育种技术/分子育种技术	技术合作	计算机技术
	湿实验支撑事件	遗传育种技术/分子育种技术	技术支撑	品种/基因/蛋白质/表型
	干实验支撑事件	计算机技术	技术支撑	品种/基因/蛋白质/表型

(3) 语料标注与数据集构建

从 861 篇语料中选取 600 篇文献组成训练集，其余 261 篇文献作为测试集，构建标注语料库。依据上述语义结构定义，使用“精灵标注”软件对测试集进行 9 类实体、8 类关系的标注，标注示例如图 3 所示，每条标注结果由专家复审校验，确保语义一致性与标签准确性。

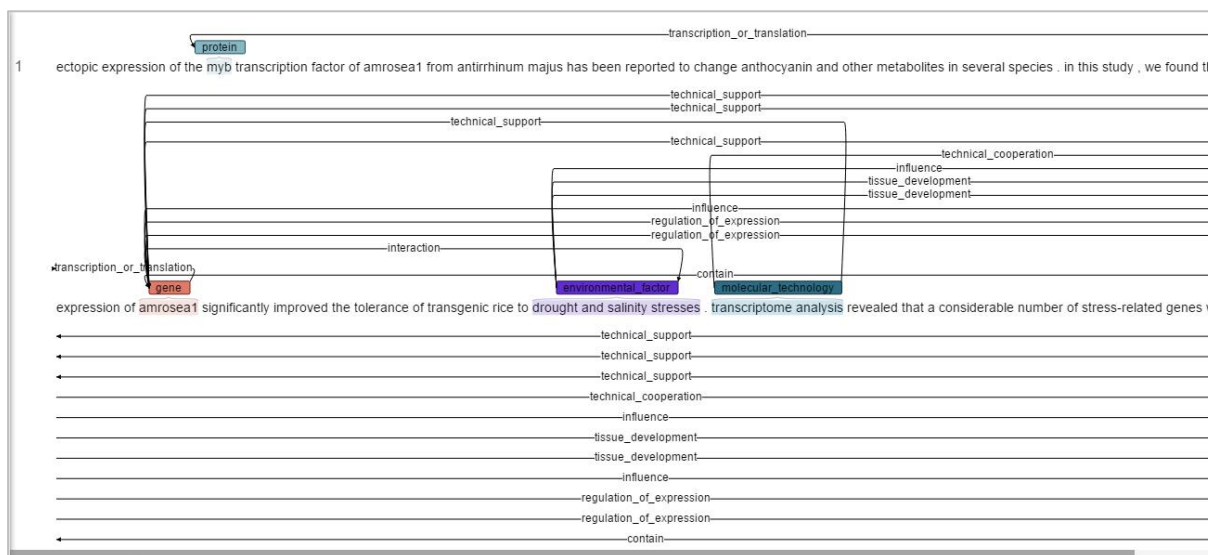


图 3 水稻育种领域标注示例

Figure 3 Annotation example in the field of rice breeding

(4) 事件抽取模型构建与训练

模型抽取阶段，在 BERT 预训练语言模型的基础上，选用 BERT-base (Chinese) 作为基础编码

模型（12 层 Transformer 结构，隐藏层维度 768，12 个注意力头）。在训练过程中，模型采用 AdamW 优化器，学习率设为 $2e-5$ ，batch size 为 16，训练轮数为 10 轮，并在单张 NVIDIA RTX 3090 GPU（24GB 显存）环境下完成模型训练与推理。在整体架构上，在 BERT 编码器基础上嵌入知识库增强的关系抽取纠错机制，构建基于知识库增强的实体与关系抽取模型，该模型采用流水线式关系抽取策略，将任务划分为实体抽取与关系分类两个阶段，以降低联合抽取任务的复杂性并提升整体抽取性能。

在实体抽取阶段，以 BERT 为基础构建表示层与编码层，并引入基于字典的“单元化”知识表示机制与联合词典与标签的“双通道”知识嵌入机制。具体而言，首先基于融合标签的前缀树结构完成外部知识库的匹配与筛选，将词典知识与标签知识映射到文本字符层；随后，在编码层中通过注意力机制分别对词组知识与标签知识进行加权嵌入，并通过知识融合网络实现多类型知识的协同建模，从而形成融合外部知识的文本表示。

在关系分类阶段，采用基于主实体的实体对关系分类方法（PL-Marker），通过枚举主-客实体对并进行关系判别，实现关系五元组的完整抽取，并进一步结合时间与学科属性构建科学实验事件五元组。该阶段利用实体抽取结果作为输入，进一步提升关系识别的准确性。

在外部知识融合方面，本研究以国家水稻数据中心本体系统“性状本体（Trait Ontology）”中 1,966 条水稻抗非生物胁迫相关基因（座）的记录作为外部知识库，同时引入轻量级检索增强框架 LightRAG，自动抽取并补充领域相关实体与关系信息，从而实现外部知识库的扩展与优化，进而实现改进 BERT+LightRAG 的水稻育种抗非生物胁迫领域实体和关系的精准抽取。

（5）事件结构标准化与输出

最终，模型抽取得到 10,409 条实体与 13,034 条实体关系，统一转换为科学实验事件五元组结构，结构如下：<主实体，主实体类型，关系类型，客体实体，客体类型>。每个科学实验事件包含时间和学科两种属性，其中，时间信息从原始文献元数据的发表年份中提取，学科分类从记载该科学实验事件的文献的学科类别中提取。

2 数据样本描述

本数据集包含水稻育种非生物胁迫领域实体（entity）和关系（relation）两张表，为方便读取和查看，将 json 格式清洗整合后以 csv 格式存储输出，其中实体（entity）表中包含 3 个字段，如表 4 所示：filename 表示实体所在元数据的文件编号；entity_name 表示实体类型，包括品种（variety），基因（gene），蛋白质（protein），组织、细胞和器官（tissue），环境因素（environmental_factor），表型（phenotype），遗传育种技术（traditional_technology），分子育种技术（molecular_technology），计算机技术（computer_technology）共 9 类；entity_value 表示具体的实体名称。

表 4 水稻育种抗非生物胁迫领域实体表（示例）

Table 4 Examples of in the field of abiotic stress resistance in rice breeding

filename	entity_name	entity_value
6993	variety	suijing 18
3847	computer_technology	clustal x
3944	protein	14-3-3 proteins

filename	entity_name	entity_value
7696	molecular_technology	crisp/cas9
3923	variety	dongjin
2769	phenotype	cd tolerance
7433	gene	apx8
7076	gene	osamt2;1
3392	traditional_technology	backcross
4042	tissue	anther
3054	environmental_factor	cd stress

关系（relation）表中包含 11 个字段，如表 5 所示：filename 表示关联关系所在元数据的文件编号；relation_name 表示关系类型，包括包含（contain），转录或翻译（transcription_or_translation），表达调节（regulation_of_expression），组织、细胞和器官发育调节（tissue_development），相互作用（interaction），影响（influence），技术合作（technical_cooperation），技术支撑（technical_support）共 8 类；entity_name_from 表示该科学实验事件的主实体类型；entity_value_from 表示该科学实验事件的主实体具体名称；entity_name_to 表示该科学实验事件的客实体类型；entity_value_to 表示该科学实验事件的客实体具体名称；科学实验事件类型表示该科学实验事件所属的具体小类，科学实验事件大类表示该科学实验事件所属的具体大类；Publication Year 表示该科学实验事件所在文献的发文年份；学科数量表示该科学实验事件所涉及的学科类别数量；被引频次表示该科学实验事件所在文献的被引频次。

3 数据质量控制和评估

为验证数据集的抽取效果，选择 BERT、BERT+Union 以及 BERT+Intersect 模型作为对比模型，从准确率、召回率和 F1 值三个指标对抽取结果进行评估。从整体效果看，本数据集构建的抽取模型在水稻育种抗非生物胁迫领域的实体和关系抽取中均取得了最佳的效果，并且相较于其他对比模型，效果提升显著（表 6）。同时，为进一步验证抽取结果的可靠性，从数据集中随机抽取 200 条科学实验事件，邀请 3 名领域专家进行人工复核。依据实体识别、关系判定及事件结构完整性等标准进行评估，并采用四级评分方式进行量化。结果表明，抽取结果平均得分为 2.73，一致性系数（Kappa）为 0.81，表明数据集具有较高的准确性与一致性。

4 数据价值

当前农业科研仍存在从海量非结构化文献到结构化、可计算知识转化困难的问题，尤其缺乏能够刻画科学实验过程及其要素关系的高质量数据资源，在一定程度上制约了科研知识的有效组织、关联与深层利用。因此，构建一个面向水稻育种抗非生物胁迫领域、以科学实验事件为核心的结构化数据集，对于填补海量文献与可计算、可推理的科研知识之间的语义鸿沟，具有迫切必要性和重要意义。本研究构建的数据集涵盖耐盐、耐旱、低镉积累等典型水稻非生物胁迫性状，通过构建事件抽取语料集、语义类型体系与事件抽取模型，引入“科学实验事件”作为核心抽取单元，采用“主语-关系-宾语”五元组结构并附加时间与学科双属性，最终形成了 10,409 条实体和 13,034 条关系的

表 5 水稻育种抗非生物胁迫领域实体关联关系表（示例）

Table 5 Examples of entity relationships in the field of abiotic stress resistance in rice breeding

filename	relation_name	entity_name_from	entity_value_from	entity_name_to	entity_value_to	科学实验事件类型	科学实验事件大类	Publication Year	学科数量	被引频次
7059	contain	variety	guangluai	protein	14-3-3 genes	材料包含事件	属性类	2022	1	14
4497	influence	gene	simyb56in	phenotype	aba accumulation	材料影响事件	因果类	2020	1	72
2644	interaction	gene	oscbcsbspb4	environmental_factor	abiotic stress	相互作用事件	因果类	2018	2	14
4100	regulation_of_expression	gene	osduf668	tissue	flag leaf	材料表达事件	功能类	2019	1	21
7104	technical_cooperation	computer_technology	origin	molecular_technology	16s rna	干湿合作事件	技术类	2022	1	29
7059	technical_support	computer_technology	clustal omega	protein	14-3-3 genes	干实验支撑事件	技术类	2022	1	14
2620	tissue_development	environmental_factor	drought stress	tissue	aleurone layers	环境调节事件	功能类	2018	1	17
3909	transcription_or_translation	gene	oscam1-1	protein	cam	材料翻译事件	功能类	2019	1	34
1886	technical_cooperation	computer_technology	clustal x	computer_technology	bwa	干实验合作事件	技术类	2016	1	69

表 6 实体和关系抽取效果

Table 6 Performance of entity and relation extraction

模型类型	实体抽取准确率	关系抽取 准确率	实体抽取召回 率	关系抽取召 回率	实体抽取 F1 值	关系抽取 F1 值
BERT	79.00%	59.45%	72.01%	49.43%	75.34%	53.99%
BERT+Union	25.75%	22.68%	81.82%	98.88%	39.18%	36.90%
BERT+Intersect	97.18%	70.00%	57.66%	31.46%	72.37%	43.41%
本模型	89.51%	79.97%	85.73%	70.67%	86.53%	76.26%

结构化数据集。相较于现有以基因、性状或品种为节点的农业数据，本数据集通过事件抽取揭示科研活动背后的创新要素配置过程，从而更贴近科研实践逻辑。本数据集有效弥补了当前农业科研领域数据结构单一、语义层级薄弱的问题，为农业科研知识图谱、科研评价、智能问答等系统提供了重要的数据支撑。

5 AIGC 使用声明

在撰写本手稿期间，作者仅使用 ChatGPT-5 进行语法检查、语言润色和语句流畅度优化。所有生成的内容均由作者进行修改和确认，作者对论文的整体内容、数据真实性及学术诚信负全部责任。

作者分工职责

林巧（1991—），女，福建省南平市人，博士，副研究馆员，研究方向为文本与数据挖掘。主要承担工作：数据采集与数据论文撰写。

鲜国建（1982—），男，四川省南充市人，博士，研究员，研究方向为大数据融汇治理与知识图谱。主要承担工作：研究思路设计与数据审核。

胡智杰（1995—），男，山东省日照市人，博士研究生，研究方向为情报分析、知识组织。主要承担工作：数据标注与论文修改。

吴侯（1998—），男，内蒙古呼和浩特市人，博士研究生，研究方向为信息资源管理。主要承担工作：数据采集与数据校对。

丁南陵（1998—），女，湖北省宜昌市人，硕士，研究方向为科技评价。主要承担工作：数据标注与论文修改。

张学福（1966—），男，黑龙江省伊春市人，博士，研究员，研究方向为信息资源管理。主要承担工作：论文指导与论文修改。

孙坦（1970—），男，黑龙江省七台河市人，博士，研究馆员（二级），研究方向为数字信息描述与组织。主要承担工作：数据质量控制与论文指导。

参考文献

[1] 朱英国. 浅谈水稻事业与国家粮食安全[J]. 中国乡镇企业, 2011(7): 41–43. [ZHU Y G. Discussion on rice cause and national food security[J]. Leisure Agriculture and the Beautiful Countryside, 2011(7):

41–43.]

- [2] 国家数据局, 中央网信办, 科技部, 等. “数据要素×”三年行动计划(2024—2026年) [EB/OL]. (2024-01-04)[2025-9-13]. https://www.nda.gov.cn/sjj/zhuanli/ztsjysx/qt/0902/20240830174038137859023_pc.html. [National Data Administration, Cyberspace Administration of China, Ministry of Science and Technology, et al. “Data Elements ×” Three-Year Action Plan (2024–2026)[EB/OL]. (2024-01-04)[2025-09-13]. https://www.nda.gov.cn/sjj/zhuanli/ztsjysx/qt/0902/20240830174038137859023_pc.html.]
- [3] 白如江, 周彦廷, 王效岳, 等. 科学事件知识图谱构建研究[J]. 情报理论与实践, 2020, 43(9): 107–114, 124. DOI: 10.16353/j.cnki.1000-7490.2020.09.016. [BAI R J, ZHOU Y T, WANG X Y, et al. Research on the construction of knowledge graph of scientific events[J]. Information Studies (Theory & Application), 2020, 43(9): 107–114, 124. DOI: 10.16353/j.cnki.1000-7490.2020.09.016.]
- [4] 吴玉浩, 王亚茹, 杨敏. 数智驱动下标准情报场景化服务的价值创造机理研究[J]. 图书馆, 2024(5): 57–65. DOI: 10.3969/j.issn.1002-1558.2024.05.009. [WU Y H, WANG Y R, YANG M. Research on the value creation mechanism of standard information scenario service driven by digital intelligence[J]. Library, 2024(5): 57 – 65. DOI: 10.3969/j.issn.1002-1558.2024.05.009.]
- [5] 许华. 基于有监督学习的医疗实体抽取方法研究[D]. 武汉: 武汉科技大学, 2016. [XU H. Research on supervised learning-based method for medical entity extraction[D]. Wuhan: Wuhan University of Science and Technology, 2016.]
- [6] 夏迎春. 基于知识图谱的农业知识服务系统研究[D]. 安徽: 安徽农业大学, 2018. [XIA Y C. Research on agricultural knowledge service system based on knowledge graph[D]. Anhui: Anhui Agricultural University, 2018.]
- [7] LU G X, TIAN L, ZHENG X, et al. Integrating knowledge-based sparse representation for image detection[J]. Neurocomputing, 2021, 442: 173–183. DOI: 10.1016/j.neucom.2020.09.089.
- [8] 李悦. 农作物病虫害知识图谱构建研究[D]. 北京: 中国农业科学院, 2021. [LI Y. Research on the Construction of Knowledge Graph of Crop Diseases and Pests[D]. Beijing: Chinese Academy of Agricultural Sciences, 2021.]
- [9] 吴茜. 基于知识图谱的农业智能问答系统设计与实现[D]. 厦门: 厦门大学, 2019. [WU Q. Design and Implementation of Agriculture Intelligent Q&A System Based on Knowledge Graph[D]. Xiamen: Xiamen university, 2019.]
- [10] 鲍彤. 番茄病虫害领域实体识别和关系抽取研究[D]. 江苏: 江苏大学, 2022. [BAO T. Research On Entity Recognition and Relation Extraction in the Field of Tomato Pests and Diseases[D]. Jiangsu: Jiangsu university, 2022.]
- [11] KRISHNAMOORTHY V. Evolution of reading comprehension and question answering systems[J]. Procedia Computer Science, 2021, 185: 231–238. DOI: 10.1016/j.procs.2021.05.024.
- [12] 袁培森, 李润隆, 王翀, 等. 基于BERT的水稻表型知识图谱实体关系抽取研究[J]. 农业机械学报, 2021, 52(5): 151–158. DOI: 10.6041/j.issn.1000-1298.2021.05.016. [YUAN P S, LI R L, WANG C, et al. Entity relationship extraction from rice phenotype knowledge graph based on BERT[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(5): 151–158. DOI: 10.6041/j.issn.1000-1298.2021.05.016.]

- [13] LI Y B, WANG X H, HUI L H, et al. Chinese clinical named entity recognition in electronic medical records: development of a lattice long short-term memory model with contextualized character representations[J]. JMIR Medical Informatics, 2020, 8(9) DOI: 10.2196/19848.
- [14] TELLO-RUIZ M K, NAITHANI S, GUPTA P, et al. Gramene 2021: harnessing the power of comparative genomics and pathways for plant research[J]. Nucleic Acids Research, 2021, 49(D1): D1452–D1463. DOI: 10.1093/nar/gkaa979.
- [15] KURATA N, YAMAZAKI Y. Oryzabase: an integrated biological and genome information database for rice[J]. Plant Physiology, 2006, 140(1): 12–17. DOI: 10.1104/pp.105.063008.
- [16] KIM J, OHTA T, TSURUOKA Y, et al. Introduction to the bio-entity recognition task at JNLPBA[C]. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA), Geneva, Switzerland, 2004.
- [17] PYYSALO S, GINTER F, HEIMONEN J, et al. BioInfer: a corpus for information extraction in the biomedical domain[J]. BMC Bioinformatics, 2007: 8(1):50. DOI: 10.1186/1471-2105-8-50.
- [18] 丁文家, 胡峻铭, 王嘉力. 水稻育种主要目标性状基因挖掘研究进展[J]. 杂交水稻, 2023, 38(3): 1–19. DOI: 10.16267/j.cnki.1005-3956.20220714.283. [DING W J, HU J M, WANG J L. Research progress on gene mining of main target traits in rice breeding[J]. Hybrid Rice, 2023, 38(3): 1–19. DOI: 10.16267/j.cnki.1005-3956.20220714.283.]

论文引用格式

林巧, 鲜国建, 胡智杰, 等. 水稻育种抗旱、耐盐与低镉积累科学实验事件数据集[J/OL]. 中国科学数据, 2026, 11(2). (2026-00-00). DOI: 10.11922/11-6035.nasdc.2025.0140.zh.

数据引用格式

林巧, 鲜国建, 胡智杰, 等. 水稻育种抗旱、耐盐与低镉积累科学实验事件数据集[DS/OL]. V2. Science Data Bank, 2025. (2026-06-15). DOI: 10.57760/sciencedb.j00001.01587.

A dataset of rice breeding experimental events for drought tolerance, salinity tolerance, and low cadmium accumulation

LIN Qiao^{1#}, XIAN Guojian^{1,2#}, HU Zhijie¹, WU Yu¹, DING Nanling¹,
ZHANG Xuefu^{1,2*}, SUN Tan^{2,3*}

1. Institute of Agricultural Information, Chinese Academy of Agricultural Sciences, Beijing 100081, P.R. China

2. Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081, P.R. China

3. Chinese Academy of Agricultural Sciences, Beijing 100081, P.R. China

co-first authors

*Email: zhangxuefu@caas.cn (ZHANG Xuefu), suntan@caas.cn (SUN Tan)

Abstract: With the widespread application of data-driven technologies in biological breeding, the construction of structured and standardized scientific experimental datasets is important for revealing research patterns and supporting intelligent breeding decision-making. To improve the breeding efficiency of rice under abiotic stress conditions, this study constructed a scientific experimental event dataset in the field of rice breeding for abiotic stress tolerance. Based on the Web of Science™ and PubMed Central databases, relevant full-text literature published between 1990 and 2023 was collected. Combining the enhanced LightRAG retrieval-augmented model and an improved BERT language model, this study extracted nine types of entities (variety, gene, protein, tissue, environmental factor, phenotype, genetic breeding technology, molecular breeding technology, and computer technology) and eight types of relationships (containment, transcription or translation, expression regulation, interaction, regulation of tissue, cell and organ development, influence, technical cooperation, and technical support. Furthermore, a five-tuple event structure with temporal and disciplinary attributes was constructed, incorporating 10,409 entities and 13,034 entity relationships. The dataset presents experimental events in a subject–predicate–object structure, and is categorized into four types: attribute, function, causality, and technical. During the data construction process, quality control measures such as standardized cleaning, semantic annotation, expert review, and multi-model comparative evaluation were used to ensure the accuracy and consistency of the data. Compared with traditional scientific data resources, this dataset emphasizes the dynamic relationships and technical pathways embedded in experimental processes. It can provide key support for rice functional gene mining, trait improvement pathway reconstruction, knowledge graph construction, intelligent question answering systems and so on, possessing substantial professional value and broad application potential.

Keywords: rice breeding; abiotic stress; scientific experimental events; LightRAG model; BERT model

Dataset Profile

Title	A dataset of rice breeding experimental events for drought tolerance, salinity tolerance, and low cadmium accumulation
Data corresponding authors	ZHANG Xuefu (zhangxuefu@caas.cn); SUN Tan (suntan@caas.cn)
Data authors	LIN Qiao, XIAN Guojian, HU Zhijie, WU Yu, DING Nanling, ZHANG Xuefu, SUN Tan
Geographical scope	Global
Time range	1990-2023
Data volume	982 KB
Data format	*.csv
Data service system	https://doi.org/10.57760/sciencedb.j00001.01587
Sources of funding	National Social Science Fund of China (23BTQ054, 22BTQ079).
Dataset composition	The dataset consists of 2 data files. Specifically: (1) an entity table containing 3 fields; and (2) a relation table containing 11 fields.