中国农业科学院机构知识库的实践探索*

赵瑞雪　杜若鹏
(中国农业科学院农业信息研究所　北京 100081)

摘要:【目的】建立中国农业科学院机构知识库(CAAS-IR), 促进中国农业科学院(CAAS)全院知识资产的数字化保存、集中揭示和传播利用。【应用背景】随着国内外 IR 建设和开放获取运动的迅速发展, 以及中国农业科学院院所科研信息化的驱动, CAAS-IR 将成为中国农业科学院重要的知识基础设施。【方法】以 DSpace 开源软件作为基础平台, 利用 Java 语言和 Solr 搜索引擎进行本地化优化改造。【结果】搭建中国农业科学院院所两级 IR 平台, 在 DSpace-core 基础上, 扩展分面检索、关联检索以及科研统计分析等功能。【结论】CAAS-IR 的建设实践, 提升了科研人员和科技管理部门对 IR 的认知水平。IR 的建设是技术与内容、管理与服务联合协作的产物, 有效的激励机制和增值服务有助于 IR 的实施。

关键词: 机构知识库　开放获取　知识资产　农业科学　CAAS　DSpace
分类号: G250

1　引　言

机构知识库(Institutional Repository, IR), 又称机构仓储、机构典藏等, 是对机构内成员智力成果(包括: 期刊论文、会议论文、学位论文、研究报告、演示报告、专利、专著、成果、科研数据等)进行收集、管理、长期保存、传播并提供开放利用的知识资产管理与服务系统。IR提供了展示机构科研水平和相关活动的有形手段, 在保存机构原生学术资源、提高机构学术影响力、宣示机构地位和社会价值等方面具有重要作用。

IR在世界范围内发展迅速, 截至目前, 全球机构知识库统计网站开放获取知识库名录(The Directory of Open Access Repositories, OpenDOAR)[1]收录的IR已达2 730个。中国台湾和中国香港地区各高校/科研机构的IR建设已规模化, 内地建设始于2004年, 中国科学院以及厦门大学、清华大学、北京大学等高校都较早开展了建设实践, 并且显示出蓬勃发展的趋势[2-5]。在此背景下, 中国农业科学院国家农业图书馆于2010年启动了中国农业科学院机构知识库(CAAS-IR)建设项目, 在保存、揭示、传播、共享中国农业科学院(中国农科院)全院学术成果, 促进学术交流, 进而为中国农科院的数字科研和管理提供知识服务支撑。

2　CASS-IR 建设目标与定位

中国农科院全院有 31 个研究所(中心)、1 个研究生院, 近 7 000 科技人员, 每年毕业的博硕士研究生人数在 890 人左右, 专业学位人数在 400 左右。年产出科研论文及学位论文近 7 000 篇, 其中 SCI 等国际论文 900 多篇, 科技专著近 200 部, 申请专利 390 多项, 还有大量研究报告、科研数据等灰色资源, 并呈逐年增长的态势, 是农业领域重要的科研创新和科研产出群体。但是, 长期以来, 全院的科研产出缺乏统一的保存和管理平台, 知识资产处于散落或小规模、缺少安全机制的低效保存利用状态, 面临知识资产长期沉睡和持续重建的双重风险, 不利于知识利用和知识创新。

随着科研环境和学术交流方式的变化, 中国农科院的科研信息化环境建设迫在眉睫, CAAS-IR 作为全院知识基础设施的重要内容, 已经成为全院科研信息

通讯作者: 赵瑞雪, ORCID: 0000-0002-1406-8562, E-mail: zhaoruixue@caas.cn。
*本文系国家“十二五”科技支撑计划基金项目“基于 STKOS 的知识服务应用示范”(项目编号:2011BAH10B06)的研究成果之一。

化环境的重要组成部分。通过 CAAS-IR 的建设实施，达成如下目标:

(1) 促进全院自主创造知识的收集、长期保存、统一管理以及共享传播，形成知识中心，提升科学研究和科研管理的信息化水平。

(2) CAAS-IR 可以拓宽全院各所学术成果的发布和交流渠道，提高学术成果被发现和引用的几率，增强学术成果的可见性，扩大科研人员及各所的学术影响力、地位和公共价值。

(3) 完善全院科研知识环境，提高知识生产和知识管理能力。

(4) 为中国农科院以及各所开展学术统计、研究评价提供数据支撑。

3　CAAS-IR 的建设实施

CAAS-IR 建设包括需求调研、建设方案确立、平台研发、内容建设 4 个阶段。

3.1　需求调研

为准确把握全院对机构知识库的建设需求，CAAS-IR 建设团队首先针对院内科研人员进行了问卷调查，问卷涉及研究所科研产出保存和共享现状以及科研人员对机构知识库的认知情况和存缴意愿等相关问题。调研结果表明: 50%以上的被调查者对机构知识库不太了解，但参与的意愿较强烈；有 80.51%的被调查者表示愿意将自己的知识资产自存储到 CAAS-IR；学科带头人、课题组长、科研管理人员对 IR 需求强烈[6]。此外，建设团队还通过咨询和走访等方式对部分科研管理人员进行了调研，结果表明: 管理者更关注科研统计信息，高产机构、学科和作者信息，以及科研产出趋势等。管理者更希望 IR 成为辅助科研管理的工具。

3.2　建设方案确立

中国农科院的各研究所都是独立的法人机构，各自相对自主地开展科研和管理活动，因此，CAAS-IR 建设采取了“集中揭示、分布部署”的院所两级模式。为每个研究所建立独立的所级 IR，承担本所知识资产的收集、管理和服务。院级 IR 通过收割或导入研究所知识元数据，实现全院知识资产的集中揭示和展示利用，如图 1 所示。同时，为了减少 IR 建设给各研究所带来的额外负担，CAAS-IR 建设初期的主要工作由国家农业图书馆承担，建设期结束后，鼓励并带动研究所承担各所数据的采集和上载工作。

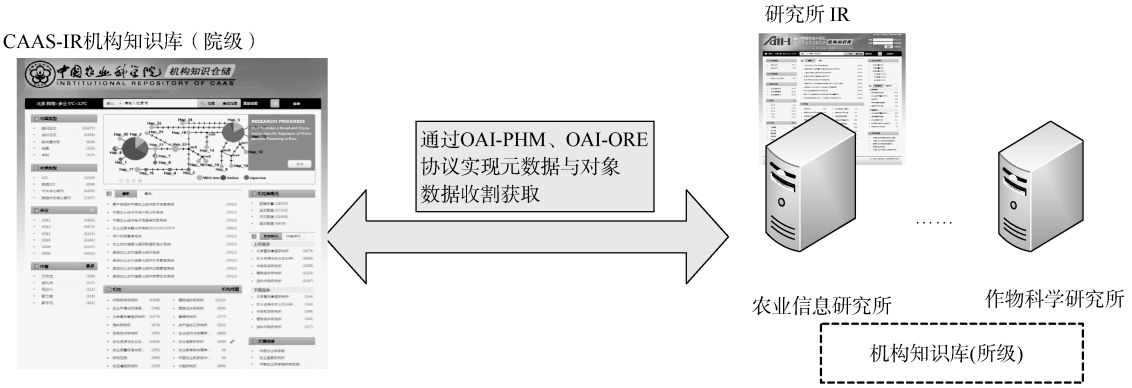

图 1　CAAS-IR 院所两级模式

3.3　平台研发

(1) 体系结构

CAAS-IR 选用 DSpace 开源软件[7]作为基础系统平台，但是 DSpace 本身具有一定的耦合性，不利于进行整体改造。鉴于 DSpace 数据模型设计合理，以及 DSpace-Core API 在数据操作方面具有的高效、稳定、独立等优点，CAAS-IR 采用了 DSpace 的数据模型，继承并扩展了 DSpace 原有表结构，通过调用 DSpace-Core API 实现对数据的操作，在此基础上采用 Java 语言和 Solr 进行二次开发和优化改造。

CAAS-IR 体系架构沿用了 DSpace 的三层结构，自下而上分别是: 数据存储层、业务逻辑层与应用表现层，如图 2 所示。其中，数据存储层位于最底层，主要由数据库管理模块与文件管理模块两部分组成。数据库管理模块用于实现数据的增删改及查询等数据库操作，文件管理模块主要用于存储机构知识库存缴的

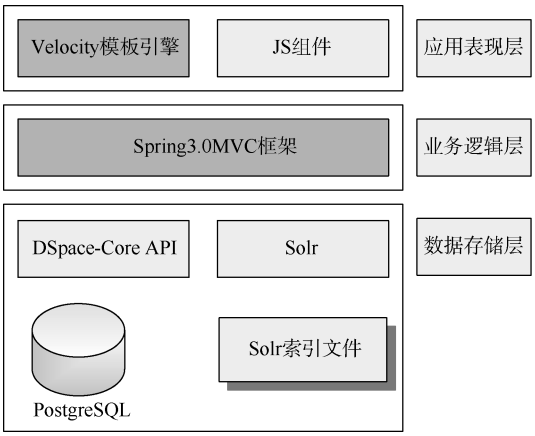

图 2　CAAS-IR 系统架构图

附件文件；业务逻辑层是系统的核心层，主要包括事务管理、用户与权限管理、搜索引擎以及工作流管理等核心模块；应用表现层位于系统整体架构的最上层，负责系统与用户的交互。CAAS-IR 采用 Spring3.0-MVC 框架构建业务逻辑层，实现业务逻辑模块之间灵活组合、可配置。由于 DSpace 原有用户界面的显示效果与风格局限，CAAS-IR 使用 Velocity 模板引擎实现 Java 代码与 Web 页面分离，为建立个性化用户界面提供支持。

(2) 系统功能

CAAS-IR 总体功能可分为服务功能和管理功能两部分。其中，服务功能包括基础服务、统计分析和深度服务，管理功能包括资源管理、数据提交和系统管理。系统功能框架如图 3 所示：

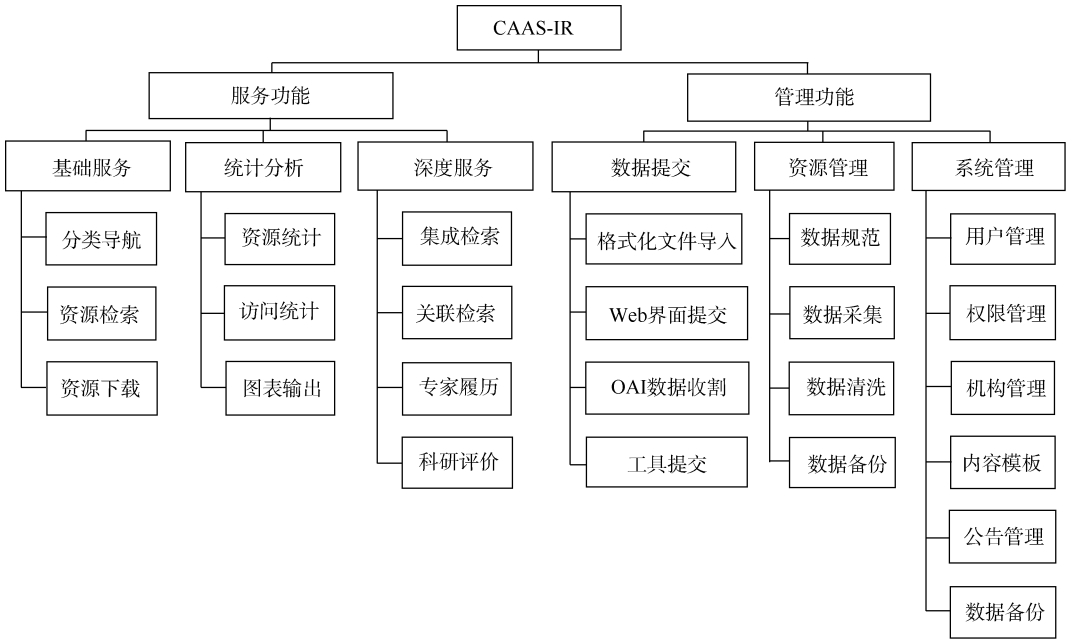

图 3　CAAS-IR 功能结构

(3) CAAS-IR 关键技术实现

①数据批量导入

为了提高数据提交效率，CAAS-IR 提供基于 Windows 环境的批量导入工具。该工具通过 getData 方法将用户提交的元数据 Excel 文件逐条分解成单行记录，同时将遍历检索用户指定的附件目录下的全部文件，根据文件名称进行匹配，若找到相符附件文件，将形成一条数据记录对象，通过 sentFiles 方法向 CAAS-IR 传送带有批次号的数据包，供接收端存储和进行批次管理。

CAAS-IR 通过 servicesDirectImport 方法接收数据，根据所获元数据信息与数据模板进行比对，检查数据是否完备，同时检查提交者是否满足权限，查询该条目是否与本机构作者相匹配。如不匹配，将该数据记录存放入提交失败记录表。同时，该接收方法具有查重功能，可以通过 service.cfg 配置文件配置查重策略，如查重对比字段、重复数据处理：拒绝提交、覆盖提交、忽略提交等。

②基于 Solr 的检索机制

资源检索是 IR 的核心功能之一，但 DSpace3.0 系统的浏览检索功能相对简单，没有分面聚类检索功能。因此，CAAS-IR 采用 Solr 重新构建资源检索模块，实现对资源的多种分面检索，主要实现机制如下：

配置 Solr 的 schema.xml，设置所要检索字段。将系统元数据字段与 Solr 的检索字段进行映射，以 key/value 形式保存在 dspace.cfg 文件中。通过 SolrIndex 的 indexContent 方法

建立资源的 Solr 索引文件。主要代码如下:

```
public static void indexContent(Context context, DSpaceObjectdso,
  SolrServer server, booleanautoCommit) throws SQLException,
SolrServerException, IOException {
if (server == null)
  return;
  SolrInputDocument doc = new SolrInputDocument();
  doc = buildDoc(context, dso, doc);
if (doc != null) {
  server.add(doc);
if (autoCommit)
  server.commit();
        }
    }
```

通过getFacetBrowserByQuery静态方法接受业务逻辑层输入的Solr查询语句，获得查询结果数量，同时对检索结果集内容进行解析，将检索条件表达式与结果集一同封装到FacetBrowser对象中。该对象将通过业务逻辑层的Controller传递给前端的SearchResult.vm页面。

通过 setConditions 方法将本次检索条件与根据用户 session 的 key 值生成的唯一号一同保存在系统缓存中。用户在检索结果页面再次点击分面导航链接时，检索条件会与上次检索条件进行叠加，从而实现递进式分面检索。

③基于叙词表的扩展检索

基于词库匹配式的中文切分词技术，除了切词算法本身的优劣外，很大程度也取决于词库本身的含词量多少以及所含词的质量。中国农业科学叙词表(简称 CAT)涵盖农业专业领域叙词和非叙词 6 万多条，将其加入到 CAAS-IR 中文分词器的切词词典中，极大提升了系统对于农业领域文献切词的准确率。

同时，CAAS-IR 调用 CAT 的 WebService 接口，通过 getCATConceptsByLabel 方法获取用户检索词的正式叙词，以及该词的上下位词、相关词与等同词。系统通过 webServiceForCAT 方法解析返回结果，封装扩展检索词链接，供用户关联扩展检索使用。

④农业科研统计

为了提高 IR 对机构科研管理的支撑，提高机构、部门和科研人员参与 IR 建设的积极性，系统增加了科研统计分析功能。由于系统沿用了 DSpace 数据模型，复杂统计依靠 SQL 实现很麻烦，所以科研统计功能主要靠 Solr 查询调用完成。主要通过 doStatisticImpl 方法，对传入人员、部门等参数进行 Solr 查询语句组装、执行查询，将结果集与检索超链接封装成 JSON 格式，供前端页面进行解析展示。

该功能可以对机构、部门、个人的科研产出进行分项统计，对统计结果以排行榜、表格、柱状图、饼图与折线图等多种方式展示，并可输出成可编辑文档，便于后续利用。同时，为了便于按照 SCI、EI、中文核心期刊等期刊指标进行统计，制定了期刊指标库，用于自动识别提交数据是否被 SCI、EI 等收录。

3.4　内容建设

(1) 内容类型

CAAS-IR 目前支持的资源类型包括: 期刊论文、会议论文、学位论文、专著、译著、编著、专利、成果、软件著作权、视频资料、研究报告、演讲稿以及科研数据等 14 种，存缴群体在建设初期主要是机构内职工或研究生。

(2) 建设方式

CAAS-IR 支持资源集中建设和自助建设两种模式。在 CAAS-IR 建设初期，由于平台数据缺乏，需要进行全院历史数据回溯，因此由图书馆进行数据的集中采集、清洗和入库。对于新增资源，目前在试点研究所采取了自助存缴模式，按照研究所的组织架构，为每个二级部门指派 1 名专人负责本部门知识内容的收集、整理和提交，笔者称之为“知识资产管理员”。

(3) 激励机制

为了鼓励科研人员积极参与资源建设与存缴，在试点研究所建立了科技管理部门，依据研究所 IR 进行年终科技统计，并以此作为科技奖励发放的关联机制，从而从机制上保证了内容存缴的及时和准确。

4　建设进展

CAAS-IR 建设取得如下进展:

(1) 完成了院所两级 IR 平台的研发与应用示范，部署了 1 个院级 IR 平台和 3 个所级 IR 平台，如图 4 所示。

(2) 通过人工和数据处理工具相结合的方式，完成了全院科研产出数据(重点是期刊论文和会议论文)的采集和清洗，并按照研究所、研究室、研究人员三个层级进行元数据组织和导入，元数据总量超过 4 万篇，其中有 3 个研究所进行了全文数据的采集和导入。

(3) 起草了《中国农业科学院机构知识库元数据规范》、《中国农科院机构知识库内容存缴暂行管理办法》、《中国农科院机构知识库内容传播政策》、《中国农科院机构知识库运行管理办法(暂行)》、《中国农科院机构知识库技术支持与服务规范》等规范文档。

(4) 在开放共享方面，目前 CAAS-IR 院级 IR 用户浏览总量为 64 696 次，其中 2014 年为 54 070 次。3 个所级 IR 用户浏览量分别为 4.7 万次、5.4 万次和 9.9 万次。

图 4　院所两级 IR 首页

5　总结思考

(1) 协同机制问题

虽然 IR 的建设对于机构的管理者和知识的生产者都具有重要价值，但无论是机构成员、还是科研管理部门、甚至图书馆员，对 IR 的建设热情普遍不高，尤其是作为 IR 建设的发起者图书馆，其对机构知识数据的管理者和服务者角色依然存在难度。图书馆如何和所属科研机构共同推进 IR 进程值得深思。目前，CAAS-IR 正面临着图书馆如何与院有关科技管理部门以及各研究所协同推进的难题。需要进一步明确 IR 建设的受益主体是全院各所，而不仅仅是图书馆单方面的事情。

(2) 内容建设问题

IR 理论上可以包括任何内容类型，如期刊论文、研究报告、科学数据等。但是，目前国内 IR 还是以存缴发表过的期刊论文居多，CAAS-IR 亦是如此。对于研究报告，尤其是以实验数据、调查数据、统计数据为主要特征的科学数据的存缴阻力还很大，不仅需要支撑技术的升级，更需要大环境的推动。随着国内各类资源共建共享平台的发展，IR 的内容定位、版权问题和政策支持问题等都值得进一步研究。

(3) 服务定位问题

IR 的基础功能是知识资产的存缴和检索。但是，在 CAAS-IR 的推进过程中却发现其与现行的科技成果管理系统、科研档案管理系统、国家有关资源共建共享平台以及某些科技计划的科技汇交系统等有着千丝万缕的关联，那么如何合理规划具有交集的不同系统或平台之间的关系，保持 IR 的特色同时又不给资源汇交者和管理者造成负担，需要建设者深入思考。其次，现在的 IR 建设似乎正在从最初的“典藏型 IR”向“服务型 IR”发展，建设者希望通过不断挖掘 IR 的增值服务，使其融入科研和管理过程以增加用户“粘度”、提高 IR 被接受的程度，如提供个人学术履历、研究热点分析、知识图谱、合著者网络、科研管理与评价等。这种状况也使 CAAS-IR 建设者陷入迷茫，如果不以发展的眼光来看待，刚刚起步的 CAAS-IR 会不会被扼杀在摇篮里呢？或许，现在的 IR 已经从概念、内容、功能、服务等各个方面已经发生了重大变化，正在从传统走向新形势下的知识服务。

参考文献：

[1] OpenDOAR [EB/OL]. [2014-10-08]. http://www.opendoar.org.

[2] 聂华，韦成府，崔海媛. CALIS 机构知识库：建设与推广、反思与展望[J]. 中国图书馆学报，2013，39(4)：46-51. (Nie Hua, Wei Chengfu, Cui Haiyuan. CALIS Institutional Repository: Construction and Promotion, Reflection and Prospects [J]. Journal of Library Science in China, 2013, 39(4): 46-51.)

[3] 张冬荣，祝忠明，李麟，等. 中国科学院机构知识库建设推广与服务[J]. 图书情报工作，2013，57(1)：20-25. (Zhang

Dongrong, Zhu Zhongming, Li Lin, et al. Construction, Promotion and Service of CAS IRs [J]. Library and Information Service, 2013, 57(1): 20-25.)

[4] 朱梦皎, 武夷山. 中、日、印高校机构知识库建设现状比较分析[J]. 图书与情报, 2012(6): 69-72. (Zhu Mengjiao, Wu Yishan. Comparative Analysis of University Institutional Repository Construction in China, Japan, and India [J]. Library and Information, 2012(6):69-72.）

[5] 台湾学术机构典藏[EB/OL]. [2013-10-08]. http://tair.org.tw. (Taiwan Academic Institutional Repository [EB/OL]. [2013-10-08]. http://tair.org.tw.)

[6] 王洪蕾. 中国农业科学院机构仓储框架设计与资源建设研究[D]. 北京: 中国农业科学院, 2011. (Wang Honglei. A Study on the Frame Design and Resources Construction of Institutional Repository of Chinese Academy of Agricultural Sciences [D]. Beijing: Chinese Academy of Agricultural Sciences, 2011.)

[7] 毛广卫. 基于 DSpace 的中国农科院机构仓储系统的研究与实现[D]. 北京: 中国农业科学院, 2011. (Mao Guangwei. Research and Achieve Institutional Repository of Chinese Academy of Agricultural Sciences [D]. Beijing: Chinese Academy of Agricultural Sciences, 2011.)

[8] 李晨英, 韩明杰, 洪重阳, 等. 建立服务可扩展型机构知识库方法探索——中国农业大学机构知识库构建与服务实践[J]. 现代图书情报技术, 2014(3): 19-25. (Li Chenying, Han Mingjie, Hong Chongyang, et al. Research on Methods of Building an Expandable Institutional Repository: Constructing China Agricultural University Institutional Repository to Deliver Effective Services [J]. New Technology of Library and Information Service, 2014(3): 18-23.)

[9] 邓红. 高校机构知识库建设实践与探索——以北京工业大学图书馆为例[J]. 现代情报, 2013, 33(7): 80-83, 129. (Deng Hong. Practice and Exploration of Building Institutional Repository in University: Taking the Library of Beijing Technology University as an Example [J]. Journal of Modern Information, 2013, 33(7): 80-83, 129.)

作者贡献声明:

赵瑞雪: 提出主要研究思路, 设计方案, 论文撰写及最终版本修订;
杜若鹏: 平台技术方案实施。

收稿日期: 2013-11-04
收修改稿日期: 2014-11-06

Practice on Institutional Repository of Chinese Academy of Agricultural Sciences

Zhao Ruixue　Du Ruopeng
(Institute of Agricultural Information, Chinese Academy of Agricultural Sciences, Beijing 100081, China)

Abstract: [**Objective**] The goal of the construction of Institutional Repositories of Chinese Academy of Agricultural Sciences (CAAS-IR) is to promote the preservation and dissemination of digital assets utilization. [**Context**] With the rapid development of domestic and foreign IR construction and the open access movement, CAAS-IR will become the important knowledge infrastructure of Chinese Academy of Agricultural Sciences. [**Methods**] The CAAS-IR uses DSpace as the prototype system and is optimized by Java programming and application of Solr. [**Results**] CAAS-IR platform extends the functionality of faceted search, retrieval and statistical analysis and other functions that are based on frame of DSpace-core. [**Conclusions**] Practice on CAAS-IR promotes cognitive level of IR for the scientific research personnel and management of science and technology department of CAAS. The construction of IR involves many aspects such as technology, resources construction, management and service. The effective incentive mechanism and value-added service will help the implementation of IR.

Keywords: Institutional Repository　Open access　Knowledge assets　Agricultural science　CAAS　DSpace