农业科技信息资源一站式发现服务研究*

赵瑞雪，张洁，寇远涛，鲜国建
（中国农业科学院农业信息研究所，北京 100081）

摘要：本文调研国内外一站式资源发现服务系统建设与服务现状，结合农业科技信息服务需求，提出农业科技信息资源一站式发现服务平台的总体建设框架，并从多源异构资源汇聚、知识组织及核心服务实现三方面详细论述，最后结合平台应用服务实践和新技术发展动态，提出下一步发展设想。

关键词：资源发现服务；农业科技信息资源；资源聚合；知识组织
中图分类号：G252
DOI：10.3772/j.issn.1673-2286.2017.11.001

1 引言

互联网技术的快速发展使信息的获取、传播和规模增长速度达到空前水平，科技信息资源呈指数级增长，日渐开放的网络环境也促使科技信息资源类型多样化，如科学统计数据、开放获取资源、社交网络数据、移动互联数据等。然而，由于科技信息资源分布在不同的数据库中，且数据结构、组织方式、管理方式等各不相同，信息资源处于高度分散和混乱无序的状态。这种海量且分散的资源为用户快速、精准、全面地获取知识带来过载压力。

图书情报机构作为专业的资源及知识服务机构，需要适应时代特征，一方面为满足用户日渐多元化、个性化的信息需求而不断扩大资源建设范围和规模；另一方面，要增强大规模信息资源整合重组、深度揭示和语义化组织的建设力度，提供更深层次的统一发现和获取服务。

为此，国内外专业服务机构历经基于OPAC的印本资源发现，各类数据库的导航及联邦检索，到资源发现服务解决方案探索的过程[1]，并不断拓展资源整合的广度和深度，努力为用户提供支持资源统一检索和深度揭示的一站式资源发现服务。

本文结合国家农业图书馆资源建设与服务实践，详细探讨农业科技信息资源一站式发现服务的解决方案和初步实践情况。

2 国内外资源发现系统建设与应用现状

资源发现系统的出现为资源供应方和使用方提供全新的交互方式和互动渠道，资源供应方由数据库出版商向资源发现服务的解决方案供应商转变。

自2009年美国Proquest旗下Serials Solution公司推出Summon系统，比利时Exlibris公司（现同为Proquest旗下公司[2]）于2010年推出Primo发现系统，随后EBSCO公司发布EBSCO Discovery Service发现系统[3]；国内数字图书馆内容商和解决方案提供商也相继推出资源发现系统，如超星公司率先推出基于数十亿级海量元数据的互联网资源发现系统，维普资讯推出智立方知识发现系统，中国知网推出学术资源发现平台。

此外，可视化检索（如面向网络的Grokker搜索引擎和面向图书馆的AquaBrowser系统）和语义检索（如基于本体的GoPubmed、基于知识图谱的Google search、基于深度学习算法的IBM Watson）等智能搜索引擎的出现和普及应用，使得用户从资源发现中不断衍生出知

* 本研究得到公益性科研院所基本科研业务费项目（编号：JBYW-AII-2016-05）和中国农业科学院科技创新工程项目（编号：CAAS-ASTIP-2017-AII）资助。

识发现和智慧发现的需求，图书情报机构和资源发现服务供应商也在新机遇中不断寻找发展机会。窦天芳等结合清华大学在引入Primo发现系统构建水木搜索的应用实践中，提出数据驱动的知识服务新思路[4]；曾建勋等从国家科技信息资源保障需要出发，提出基于语义的国家科技信息发现服务体系[5]；黄永文等总结中国科学院用户的主要信息资源发现需求，提出集成化、可视化知识检索服务平台的体系框架[6]；EBSCO公司在其发现系统中加入Grokker插件，实现可视化检索功能[6]。下一代资源发现系统也正向可视化、关联化、智能化方向发展。

综上所述，目前国内外商业化的资源发现服务产品在功能上均基于自身元数据仓储为用户提供资源整合、统一检索和检索结果分面揭示等功能。在资源整合上，国外平台的资源以外文资源为主，中文元数据覆盖率低，国内平台的资源以自有资源为主，中文资源丰富，外文元数据资源匮乏，同时商业化产品对开放获取资源的整合相对缺失；在统一检索上，则在检索结果相关性排序和排序方式备选方案等方面尚不理想；在结果揭示上，资源发现系统在结果揭示深度和知识关联展示等层面尚有提升空间。在大数据和人工智能等新技术不断发展的背景下，资源发现系统服务提供商和图书情报机构也不断纳入新技术，积极解决上述问题，寻找发展空间和转型机会，从资源发现向下一代知识发现、智慧发现过渡。

3 农业科技信息资源一站式发现服务平台建设方案

3.1 服务需求

信息技术的发展、大数据时代的到来引发了科研范式从计算科学向数据密集型科学的演化，科学研究第四范式在此环境下产生。数字时代下，无论是研究个体还是科学共同体都在渴求能够获得更深入、更便捷的数据服务。图书情报机构在科研用户的敦促下，需要主动推进知识服务升级，帮助用户扩展资源发现，揭示信息资源蕴含的知识，构筑知识间关联[7]。科研第四范式背景下，用户对信息资源发现服务的主要需求和功能表现可以概括表述为以下3点。

（1）广泛多源的资源需求。信息时代下，科学信息资源的多样化拓展了用户选取信息的时空，也更大程度地满足用户的信息需求，科研用户在传统科技型文献资源的基础上期望可以获得科学数据等数值型资源及百科、科研项目等事实型数据资源来辅助科研过程的顺利开展。

（2）全面准确的搜索结果。检索结果的全面和准确需要依托于检索意图的准确理解、检索结果的个性化展示等关键要素。检索意图即用户的检索目的，用户希望发现系统能理解检索词代表的实际检索意图及真正指代对象。如输入“袁隆平”，查询该专家的科研成果而非元数据或全文中包含“袁隆平”这个词的资源，输入“中国农业科学院”查询该机构已发表的科研成果及该机构的相关介绍。检索结果的个性化展示指用户希望检索结果的排列和呈现方式可以按照倾向的方式来排列。如在检索结果查看时，用户需要最相关、最权威、最新的信息资源靠前展示，同时当检索命中结果较多时，用户需要借助多维分面导航和浏览服务来帮助其筛选出想详细查看的资源。

（3）深度直观的知识关联。检索结果可以引领和指导科研用户需求，用户的信息需求会根据检索命中结果不断深化，知识关联揭示可以帮助用户查看文章作者的其他成果、该篇文章的被引文献、引用文献及主题相关文献，帮助用户了解与检索词高度相关的热搜词，图形可视化的展示方式是非线性关联关系直观展示的较好选择。

综上，基于上述需求，国家农业图书馆启动面向农业专业领域的科技信息资源一站式发现服务平台，期望可以为用户提供智能化、初步语义化、可视化的农业科技信息资源发现与获取服务。

3.2 平台建设框架

基于面向服务架构的分层设计思想，提出松耦合、具有可扩展性、易重用性、易维护性的农业科技信息资源一站式发现服务平台的总体框架（见图1）。该框架主要包括资源汇聚层、知识揭示层、应用支撑层、服务层和用户层五层结构。

（1）资源汇聚层。该层主要实现资源的收集、聚合及有序组织，完成农业科技信息资源元数据仓储的构建。资源汇聚层包括两方面工作：一是完成对多源异构元数据资源（包括文献类资源、数据类资源、开放获取资源等）的汇交；二是对已汇交资源进行整理、清洗、规范化与统一管理，形成同构的标准化元数据集，这些

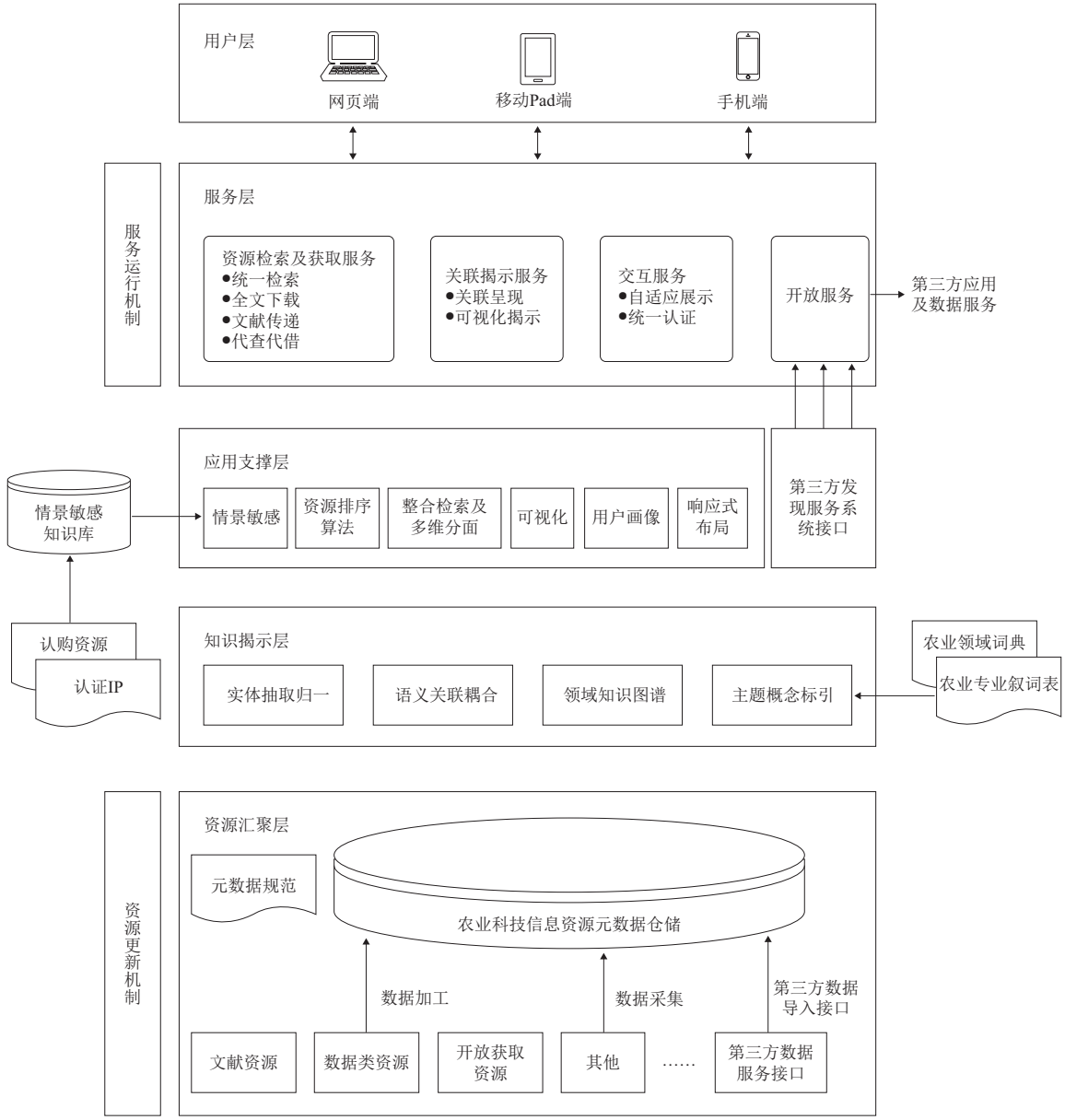

图 1 平台总体框架

数据集即元数据仓储。

（2）知识揭示层。对已聚合的科技信息资源进行再加工和数据挖掘，通过主题概念标引、实体抽取和归一、语义关联耦合及领域知识图谱的构建完成对元数据资源的知识抽取及构建富含语义的知识关联网络。

（3）应用支撑层。应用支撑层是平台应用的核心技术支撑，该层基于应用服务需求，通过关键技术实现对底层元数据仓储及知识库的调用规则设计和业务实现。基于资源汇聚组织整合的元数据仓储及知识揭示阶段生成的实体库、知识图谱等知识服务工具，采用整合检索及多维分面、资源排序算法、情景敏感等技术为用户提供全面便捷的资源检索及获取服务；利用可视化技术发现资源间的非线性关联知识的可视化揭示；基于响应式布局技术实现多终端接入页面布局的自适应展示。

（4）服务层。服务层主要接收农业及相关专业领域科研用户的资源请求并返回相应数据，为客户端提供平台一站式发现服务应用程序的访问，平台主要为科研用户提供资源检索和获取服务、关联揭示服务、交互服务及第三方开放服务。为支撑各项服务内容，保证系统的交互友好性，提升用户使用体验，平台提供自适应终端页面布局的显示服务，也基于图书馆科研通行证

为平台用户开放统一认证服务。

（5）用户层。为提升用户使用体验，平台为终端用户提供跨时空限制的多终端访问接入方式，开放网页端和移动端协同访问渠道。移动端访问渠道目前通过平台微信公众号、安卓和IOS版本的APP客户端等开放服务。

3.3 多源异构资源汇聚

实现资源发现服务的核心在于数据资源的全面性、权威性、规范性和时效性，因此构建农业科技信息资源元数据仓储成为平台建设的关键，图2展示了多源异构资源汇聚的总体流程。

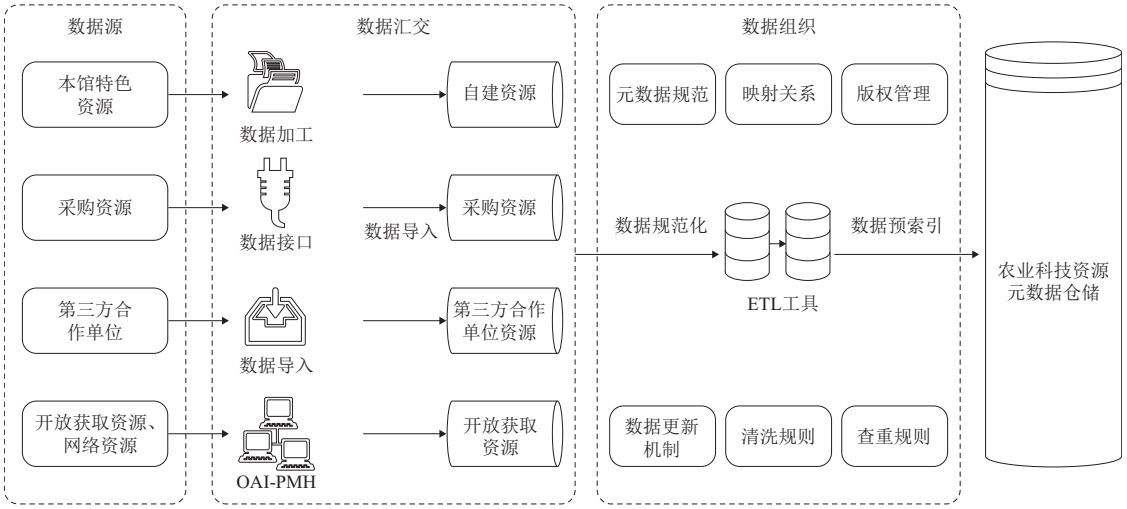

图 2 资源汇聚流程

平台资源主要来自国家农业图书馆自建资源、国家农业图书馆采购资源、第三方合作单位资源及开放获取资源。国家农业图书馆自建资源包括专业领域报告、农业百科等特色资源，主要通过收集、自组织和加工等流程来完成构建；国家农业图书馆采购资源主要包括文献、科学成果等资源，通过对方开放接口以数据导入方式直接获取；第三方合作单位资源来自于与第三方合作单位共建共享的资源，该类资源也通过开放接口方式直接导入到仓储中。目前平台正在加快推进开放期刊、学位论文、图书、课件、机构仓储等开放获取资源及网络开放获取资源（如统计数据类资源）的建设工作，该类资源主要通过定期数据采集收割的方式使用TRS SMAS采集器等数据采集收割工具，遵循OAI-PMH等相关元数据收割协议来获取。

资源组织工作主要完成对已聚合资源的有序组织和关联，通过数据规范化、数据预索引等实现对多源异构科技信息资源元数据仓储的构建。

（1）数据规范化。将已获取的数据资源进行清洗、梳理、分析，基于每种资源的特点和内容意义，确定统一的元数据规范并对不同来源元数据进行映射；同时，基于数据动态生成规律，确定数据资源定期更新机制。

（2）数据预索引。遵循已确定的元数据规范和数据更新机制，定期将获取的科技信息资源元数据去重、过滤和合并后，将基于国家农业图书馆自建农业科学叙词表、名称规范库自主开发的AJAX插件整合进ETL工具，并使用该工具完成对元数据资源的预索引工作，实现对农业科技信息资源元数据仓储的索引式存储。

3.4 语义关联的知识组织

知识揭示过程是对农业科技信息资源元数据仓储进行语义化加工，抽取其中的知识资源，为用户提供知识服务，以满足当前信息环境下更多应用场景的服务需求。基于语义模型、叙词表等一系列知识组织工具，使用主题概念标引、实体抽取归一、语义关联耦合及领域知识图谱构建等知识组织方法实现实体规范库及知识库的构建。

（1）主题概念标引。在文献关键词指向不明或缺失的情况下，使用主题概念标引方法依据文献的标题、摘要等字段中的内容，用最能够表达文献主题内容的概念词语作为标引主题词来标引文献。标引流程主要包括生成主题词列表、计算主题词权重、从主题词列表中选择最终标注的主题词。主题标引输出需要记录的信息包括文献ID、关键词、权重评分等。

（2）实体抽取归一。在不同语境下，当机构、作者、期刊、会议、地点等元数据信息存在重名或多种表述方式时，对这些实体信息进行抽取归一，构建对应的机构、专家等实体规范库，以保证其准确指代，并支持数据规范化及检索的语义调用。

（3）语义关联耦合。为实现仓储资源间的语义关联与耦合，借助数据挖掘、知识抽取等技术，计算资源在作者、主题、相互引用关系等维度的关联，构建知识关联网络。

（4）领域知识图谱。基于仓储资源元数据抽取出关键实体的三元组（实体、属性、值），借助知识图谱工具构建农业领域知识图谱，供知识库和语义检索调用。

3.5 核心服务体系设计

农业科技信息资源一站式发现核心服务体系主要包括资源检索及获取服务、关联揭示服务、交互服务及开放服务。

（1）资源检索及获取服务。主要包括统一检索、全文下载、文献传递及代查代借等服务内容。针对已整合的仓储资源，平台为用户提供统一检索及多维分面导航服务的同时，也提供合法权限内的全文下载及文献传递等平台收录资源的获取服务；针对平台暂未收录资源，则通过代查代借的辅助方式提供获取服务。

（2）关联揭示服务。主要包括关联呈现及可视化揭示等服务内容。关联呈现实现与检索结果在主题词、作者、机构等方面存在关联关系的资源展示；可视化揭示主要实现对抽取出的语义关联知识构建可视化图谱。

（3）交互服务。在用户界面设计上，首先基于响应式布局，兼顾不同设备分辨率，消除浏览终端，包括PC机、智能手机及平板电脑等对网页展示效果的影响，提供自适应的终端展示服务；其次，考虑到目标用户群体的特殊性，平台面向用户开放统一认证服务。

（4）开放服务。遵循开放服务、共建共享理念，平台基于OpenURL接口技术面向第三方应用提供数据资源调用及资源检索服务。

4 关键技术实现

在平台建设过程中突破了资源整合与组织、知识抽取、语义关联、整合检索、多维分面、情景敏感等关键技术，实现对多源异构资源元数据的汇聚组织和初步知识抽取，构建多维语义索引，并在此基础上面向农业领域科研工作者提供一站式资源检索与获取服务。平台基于农业科技信息资源元数据仓储支持对资源的语义检索和分面导航，基于情景敏感的资源获取，对关联知识的结果可视化揭示，以及多终端设备对平台的协同访问。

4.1 基于元数据检索的统一发现

基于规范元数据标准的科技信息资源仓储为平台的统一发现提供数据基础，整合检索为平台的统一发现提供了功能保障。平台检索基于开源搜索引擎Apache Solr实现，除为农业领域科研主体提供整合检索、多维分面、命中词高亮等功能外，基于已构建的多维语义索引，初步实现了对自然语言检索式的语义浅层理解和分析，提供统一发现过程中的实体命中、语义扩展及跨语言检索等功能。

基于知识组织过程所构建的实体规范库，匹配检索词中实体（如机构和作者等），命中实体相关资源并在结果呈现时优先展示；基于农业科学叙词CAT-skos词表、农业领域词典DIC等，在资源发现过程中，对检索词的关系词（用、代、属、分、参）进行扩展，扩大检索范围提高检索查全率和查准率；使用优化后的农业领域中英文高频词表，实现检索词的中英文互译，完成元数据的跨语言统一检索（见图3）。

以检索词“分子标记”为例，在检索外文电子文献资源时，英文对照词为“molecular makers”，同时基于叙词表给出相应扩展词，检索发生时，检索词组包括“分子标记”“molecular makers”“molecular mapping”“molecule marker”“molecular mark”及“molecular marker”等。

4.2 基于情景敏感的资源获取

为感知用户信息和实时使用环境，为用户匹配合理的资源和服务获取权限，引入情景敏感技术。情景敏感知识库的构建基于预先收集的用户资源订购情况和IP地址范围，用户访问资源时，平台将用户访问IP与情景敏感知识库进行校验判断其对检索结果的获取权限。对于权限校验通过者，平台给出检索结果的参考链接，此处链接解析以OpenURL及SFX Link Server技术为基础[6]；权限校验未通过的用户，则可选择通过原文传递服务来获取资源（见图4）。

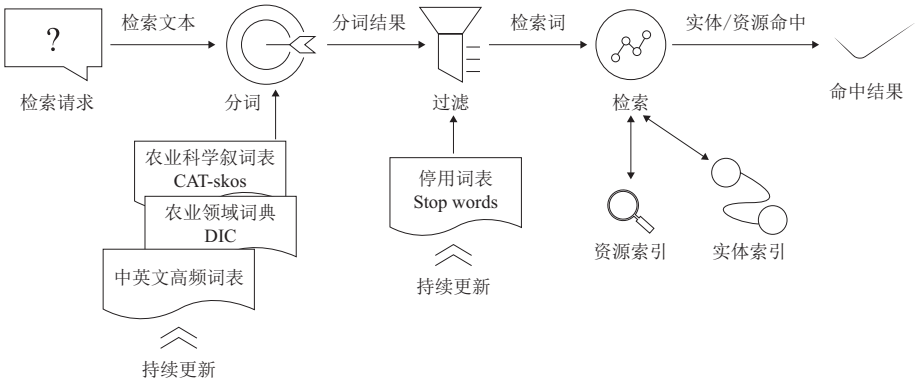

图 3 平台检索流程

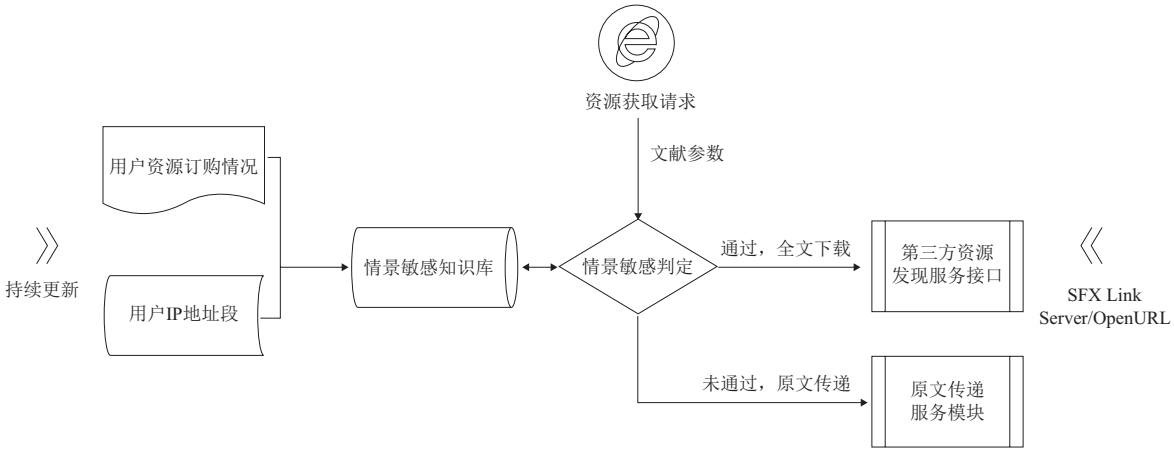

图 4 基于情景敏感的获取服务

4.3 基于关联知识的结果揭示

平台基于文献资源的二次加工和深度挖掘，在检索结果页面，将主题标引和语义关联耦合抽取出的关联知识进行可视化展示。如以“水稻”作为检索词，主题词云展示了检索结果中文献关键词聚类产生的高频词集合，知识立方则展示了检索主题下主题概念、领域专家、科研机构及科学数据间的关联网络关系，图5展示了“水稻”的知识立方图谱效果。

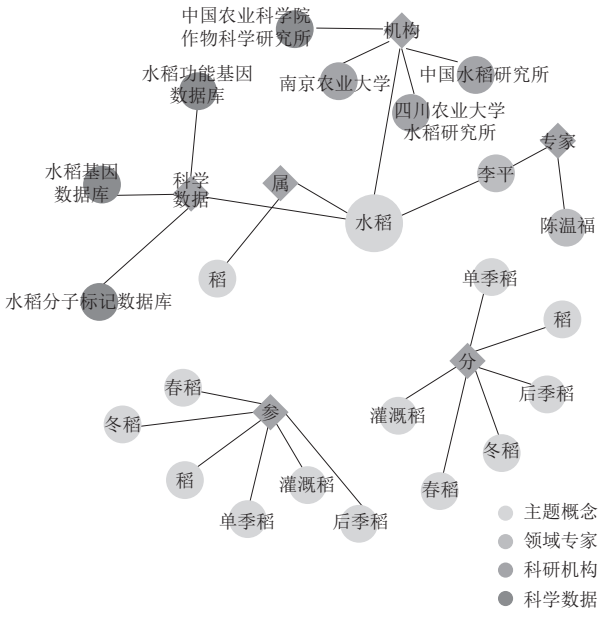

图 5 “水稻”知识立方图谱

4.4 基于共享应用的多终端协同服务

基于多终端的访问接入渠道，图书馆为用户提供跨物理时空限制的资源发现服务，除网页端外，目前已开通移动端访问（包括微信公众号、IOS版和安卓版的APP客户端），移动端页面基于HTML5实现且支持响应式布局。为保证多终端协同服务效果，平台采用应用共享方式，资源检索及获取请求全部通过接口传入后台应用，统一处理完成后，通过接口传递回网页端及移动端的前台服务页面，保证多终端的服务一致性和协同性。

5 总结与展望

目前，平台一期已经建设完成并于2016年11月16日面向全国农业科研用户服务。截至2017年9月，平台注册用户数量超过2万人，累计访问量超过30万人次，日均文献下载量超过200篇（来自百度统计对平台服务网站的实时监测数据）。

随着平台服务的不断深化及大数据等新技术的不断涌现，用户对平台服务提出更高的要求，如检索结果的个性化呈现、多类型数据资源的深度融合及数据挖掘分析等。未来服务平台将围绕以下方面进一步完善和提升完善。

（1）提升资源元数据厚度。考虑在当前元数据的基础上增加文献的引用情况和被引情况，并且通过文献作者标识符（如农业科学家ORCID）将作者信息及相关科研成果情况关联到平台中对应的文献元数据，提高元数据厚度，扩大检索范围，同时结合可视化技术直观展示相关领域成果及其合作者情况。

（2）向下一代语义检索过渡。当前平台已初步应用语义检索的研究成果，实现检索词的概念扩展及实体识别，未来将在资源和知识组织层面继续深度挖掘元数据价值，打造基于本体的资源组织方法，揭示资源元数据文本中隐含的深层语义，将关键词级的信息检索提升到概念级的知识检索，并利用本体丰富的层次结构和语义关系进行检索条件的语义概念补充和拓展；同时持续优化检索排序算法，引入资源打分分级机制，构建兼顾资源元数据质量、权威性与查准率、查全率的检索排序模型，提升结果展示的智能化和人性化水平。

（3）基于用户画像提供个性化服务。用户画像指通过对用户数据的挖掘提炼，尽可能全面细致地抽取出一个用户的信息全貌，帮助用户解决如何将数据转化为价值的问题[8]。平台下一步建设工作将基于用户行为日志（登录、浏览、下载、请求等行为数据），收集分析用户显性需求、挖掘用户隐性需求，构建基于画像的用户兴趣分析模型及用户关系图谱，为用户提供更多的个性化定制服务和推送。

参考文献

[1] 孙宇.2013年图书馆前沿技术论坛(IT4L):“资源发现之旅”研讨会综述[J].数字图书馆论坛,2013(7):68-70.

[2] 佚名.ProQuest与艾利贝斯合并加速全世界图书馆的创新步伐[J].大学图书馆学报,2015,33(6):127.

[3] 包凌,蒋颖.图书馆统一资源发现系统的比较研究[J].情报资料工作,2012(5):67-72.

[4] 窦天芳,姜爱蓉.资源发现系统功能分析及应用前景[J].图书情报工作,2012,56(7):38-43.

[5] 曾建勋,丁遒劲.基于语义的国家科技信息发现服务体系研究[J].中国图书馆学报,2017,43(4):51-62.

[6] 黄永文,张智雄,吴振新,等.集成化可视化的知识检索服务平台建设[J].科研信息化技术与应用,2013,4(2):34-42.

[7] 张颖.美国研究型图书馆研究数据服务的实践进展及趋势[J].图书情报工作,2017(9):33-41.

[8] 朱前东.国外资源发现系统评价策略研究[J].图书与情报,2014(4):6-10.

作者简介

赵瑞雪，女，1968年生，博士，研究员，博士生导师，研究方向：信息管理与信息系统、信息资源管理、知识组织与数字图书馆，E-mail：zhaoruixue@caas.cn。
张洁，女，1991年生，硕士，馆员，研究方向：信息管理与信息系统、语义检索，E-mail：zhangjie07@caas.cn。
寇远涛，男，1982年生，博士，副研究馆员，研究方向：信息管理与信息系统、数字图书馆理论与技术，E-mail：kouyuantao@caas.cn。
鲜国建，男，1982年生，博士，副研究馆员，研究方向：知识组织、关联数据，E-mail：xianguojian@caas.cn。

Research of Discovery Service of Agricultural Sci-Tech Information Resource

ZHAO RuiXue, ZHANG Jie, KOU YuanTao, XIAN GuoJian
(Agricultural Information Institute of CAAS, Beijing 100081, China)

Abstract: Through the literature research of worldwide resource discovery systems' construction and service status, this paper proposes the overall construction framework of Agricultural Technology Information Resources Discovery System according to the domain requirement, and then gives out detailed description in three dimensions: converge of multivariate heterogeneous resources, knowledge organization and core services. At last, combining with system service practice and new technology trends, the paper brings forward development outlook of agricultural technology information resources discovery system.

Keywords: Resource Discovery System; Agricultural Sci-Tech Information Resource; Resource Converge; Knowledge Organization

（收稿日期：2017-10-31）