DOI:10.13530/j.cnki.jlis.174004

基于语义的国家科技信息发现服务体系研究*

曾建勋　丁遒劲

摘　要　商业性资源发现系统特别是国外资源发现系统逐步得到市场认可,国内许多图书馆已经引进,但实践应用中存在资源组织揭示层次较浅、服务功能单一等问题。本文基于国家科技信息资源保障需要及国内外资源发现服务现状,提出构建我国的国家科技信息发现服务体系。从文献数据采集、数据统一管理、语义知识组织、系统服务功能、方法工具支撑五个基本方面进行详细论述,重点强调本体、叙词表等知识组织工具在文献资源语义知识组织中的应用,形成语义关联的数据资源体系,促进系统功能从单一资源级别发现服务向引文分析、科研实体识别等知识服务拓展。相关标准规范建立、知识产权管理等也是构建国家科技信息发现服务体系的重要因素。图 4。参考文献 17。

关键词　国家科技信息保障　发现服务　资源整合　知识发现　语义关联

分类号　G250.71

Research on the System of National Scientific Information Resources Discovery Service Based on Semantic Association

ZENG Jianxun & DING Qiujing

ABSTRACT

Commercial resources discovery systems, especially Primo and EBSCO Discovery Service which are provided by foreign companies, are being well recognized by the market. Many university libraries in China have introduced discovery systems for unified searching, but the systems still need to be upgraded in terms of resources organization and service functions. Particularly, semantic association should be applied to the architecture of the system of national scientific information resources discovery service.

Based on the demands of national information guarantee, this paper analyzes the present situation and the challenges of National Science and Technology Library (NSTL for short), and the application of commercial resources discovery systems in Peking University Library, Tsinghua University Library etc. is also introduced and investigated. Currently, most of the university libraries just use the discovery systems for unified retrieval. The system of national scientific information resources discovery service should be architected for the integration of information service, information resources and library systems.

The framework of the national scientific information resources discovery service is articulated from 5 aspects of data acquisition, data management, knowledge organization, system services and method support,

* 本文系国家社会科学基金重点项目"基于知识组织的图书馆资源发现服务体系研究"(编号:17ATQ002)的研究成果之一。(This article is an outcome of the key project "Research on the System of Library Resource Discovery Service Based on Knowledge Organization" (No.17ATQ002) supported by the National Social Science Foundation of China.)

通信作者:丁遒劲,Email:dingqj2011@ istic.ac.cn,ORCID:0000-0002-9890-1418(Correspondence should be addressed to DING Qiujing,Email:dingqj2011@ istic.ac.cn,ORCID:0000-0002-9890-1418)

which is focused on the process and methods of metadata management and semantic association. Finally, the application prospect of the system of national scientific information resources discovery service is described. The system can integrate the service system and business system as a whole to support information resources acquisition, data organization, user management and information service. At the same time, the system is not only for national institutions but also for the cooperative management of all university libraries and academic libraries in China.

This paper studies the construction framework and process of the system of national scientific information resources discovery service at the theoretical level, but the effective operation of the service system depends on the establishment of a unified system of standards and proper handling of intellectual property issues. Meanwhile, the service system should be updated and maintained regularly with the increase of digital resources, the failure of resource link address, the change of bibliographic information, the change of resource format and so on. A full-process resource maintenance system, from copyright negotiation and authorization data acquisition, data integration and fault tracing, should be established to construct a full-text sharing mechanism that is in line with copyright regulations and benefit balance to ensure the sustainability and stability of discovery services. 4 figs. 17 refs.

KEY WORDS

National science and technology information resources guarantee. Discovery service. Resources convergence. Knowledge discovery. Semantic association.

0 引言

伴随数字出版业特别是开放出版、自媒体出版等新兴出版模式的快速发展，科技文献资源形态呈现多元化趋势，不同机构、不同数据库系统、不同网站分散存储资源，以不同结构各自描述资源数据特征，形成一个个信息孤岛。与此同时，图书馆除采购原有印本文献外，越来越多地订购来自不同数据库商、出版商的数据资源，对采购的不同类型、不同载体的数据无法形成“资源统一管理”，难以达到用户一站式检索发现和获取的目的。为此，在摒弃原有针对异构数据和异构系统的“联邦检索”之后，新型的文献资源发现系统得以快速发展和应用。

然而，从现有发现系统的服务层次来看，其发现功能仍停留在资源“载体”层面，基于“资源类型—母体—元数据”进行数据组织，而未能真正深入到文献“内容”层面，实现关联化组织、语义化检索和智能化服务。此外，我国图书馆使用的外文资源发现服务主要依托于国外系统，国内数据库商的发现服务范围仅以中文资源为主，而随着维普、万方等将中文资源元数据加入到国外发现系统中，国外发现服务商有可能对国内发现服务市场形成垄断，而国外商业性资源发现系统主体数据尚未在中国本土进行存储，存在数据安全隐患，对我国科技信息服务的持续性、稳定性构成威胁。因此，亟待面向国家信息保障打造中国本土的公益性科技信息发现服务体系，基于知识组织工具建立语义关联的科技信息发现服务，推动科技文献资源在更大范围内的发现和共享，提升国家科技文献资源保障与服务效益。

1 构建国家科技信息发现服务体系的必要性

1.1 我国科技信息资源保障的现状与问题

自 2000 年 6 月成立国家科技图书文献中心（National Science and Technology Library，NSTL）

以来，依托国家科技信息保障体系采购、加工、整理了大量中外文文献资源，每年订购外文科技期刊达 17 000 多种，其中全国独家订购外文科技期刊 6 000 余种；非刊文献近 9 000 种，其中全国独家订购 3 000 种；面向全国开通的电子现刊达到 650 种，回溯科技期刊达 3 018 种，联合采购电子资源达 600 余种，自主加工文摘数据达到 2.8 亿条，向第三方引进 2 万余种期刊的电子资源元数据，外文期刊引文 1.8 亿条。通过 40 个服务站辐射全国进行公益文献服务，每年电子资源下载量达到 6 000 万篇左右，印本文献原文传递量约 120 万篇。目前，NSTL 国际代查代借的文献量在持续下降。经统计，2012 年，SCI 收录约 2.3 万篇在 1978—2008 年间发表的高影响力论文，其中仅有 300 篇论文需通过国际代查代借服务从国外获取，说明我国外文文献信息资源保障水平已经达到相当程度，国外科技信息资源保障状况较为乐观。

然而，我国科技文献信息资源保障仍然面临挑战[1]，具体表现在以下方面。①仅拥有自身收藏纸本或电子资源的 2—3 亿条元数据，无论是资源类型还是资源体量均难以满足资源全面发现与大数据挖掘分析的需求。可获知的检索服务、可挖掘的发现服务需要大样本文本数据，凸显出现有自给自足的元数据资源建设方式的不足。②现刊和回溯论文元数据、论文和引文元数据、全国开通电子资源或开放获取资源元数据相对分离，尚未进行有效整合与集成。各类型资源数据存储于不同平台，遵循不同的资源描述标准，亟待建立统一的元数据标准及其相互映射关系。③期刊论文、学位论文、科技报告等各类型资源间关联度有限，交叉揭示程度不够。在科技信息资源类型界限越来越模糊的情况下，信息资源检索没有打破资源类型间的界限，无法拓展检索功能，实现关联检索。④人才、机构、项目、技术知识点识别与关联度不够，不便于文献计量评价，科技信息资源加工粒度有限，以题名、作者、关键词等母体信息为主，论文中相关科研实体信息揭示不完全，科研实体关系尚未揭示，仅能满足基本检索需求，计量评价难以展开。⑤全文资源只包含自身馆藏，尚未充分集成和利用其他单位的馆藏资源，无法形成更广泛的知识联盟空间。总之，在国家科技信息资源保障水平已经达到相当程度的情况下，需要改变目前单凭自身购买、自给自足的加工方式，应借助多种渠道，依靠多个机构，整合集成多来源、多类型、多载体、多渠道的元数据资源，构建国家级科技信息发现系统，推动科技信息资源全面深入的共建共享。

1.2　国内外资源发现服务现状和趋势

国内外发现系统有两类，一类是商业性资源发现系统，其主流系统有 ProQuest 的 Summon 和 Primo、EBSCO 开发的 EDS，国内有超星发现、万方学术搜索等，其分别由系统集成商、数据集成商和资源供应商演化而成，并分别具有各自的优势；另一类是公益性发现系统，如 OCLC 建设 WorldCat Local（WCL）、澳大利亚国家图书馆推出的 Trove，CALIS 的 e 读以及中国国家图书馆的文津系统等，这些一般由国家级公益机构创建。商业性发现系统的元数据量级在 10 亿条以上，其元数据资源主要通过主动与资源出版商、数据库商谈判签约获取，部分资源来源于资源服务方（如图书馆等）的主动授权与上传。如，Summon 与超过 9 000 家出版社建立合作关系，元数据全部来自直接签约，且不依靠第三方元数据来源（如 CrossRef）。商业性发现系统全面整合异构资源，将纸质馆藏资源、本馆订购商业数据库、开放获取资源和机构知识库整合成统一的元数据索引，其整合的资源相对较全，如国家工程技术图书馆订阅的 4 000 多种国外科技期刊，在 Summon 中占到 96%，在 EDS 中占到 93%。而公益性发现系统的元数据量级则在几亿条间，如 Trove 系统主要通过机构协作方式，从1 000多家图书馆及其他文化机构获取元数据[2]。总体上，由于元数据海量特性及大数据云计算能力，发现系统的构建机构不多，绝大多数图书馆直接租用发现

服务用于资源集成检索。目前全球已有超过1 300家图书馆及相关信息机构引进各类资源发现系统。

我国图书馆对发现系统的构建实践,主要从对商业性发现系统的选型分析开始。窦天芳[3]、聂华[4]、袁玉英[5]、苏建华[6]、彭佳[7]、许丽媛[8]、包凌[9]等都结合自身图书馆的应用需要,从不同角度对主流资源发现系统的页面构造和布局、特色资源和功能等进行多层次、多角度分析,阐述发现服务的推出背景,提出保证元数据质量、改善用户体验、完善数据分析以及嵌入个性化服务功能等优化策略。国内主要大学图书馆都在引进基础上形成自身应用的发现系统,如北京大学的“未名学术搜索”、清华大学的“水木搜索”、上海交通大学的“思源探索”等;个别图书馆根据自身资源规模与服务特点对发现系统进行二次开发,如中国科学院文献情报中心通过购买Summon来检索电子资源,建立本地索引来检索本地自建资源,并基于Summon API和本地索引构建统一的资源发现服务[10],清华大学图书馆基于发现系统海量元数据开发形成学科趋势分析系统,推出知识化服务。

未来的资源发现系统需要进行服务聚合、资源整合和系统融合,将资源、服务建立在统一的系统平台上加以集成。首先,需要把印本资源、电子资源、开放资源和自身特色资源整合在一起;其次,需要把图书馆相关系统包括用户日志系统、资源管理系统、联合目录系统、数据仓储系统和前台服务系统有机地衔接起来,把文献传递系统、统一认证系统和机构知识库连接起来,构成一个完整有效的自动链接机制;第三,需要进行服务整合,即从文献层面进行多维度全文指引,既可从出版社网站直接获取,也可从图书馆原文传递,不触碰知识产权底线;第四,需要进行元数据层面的知识关联揭示,提供进行大数据分析所需要的专业数据服务,提供科研实体监测、用户行为分析、科学前沿预测等知识服务,推进数据耦合服务或者计量评价服务等。

1.3 国家科技信息发现服务体系构建的必要性

资源发现服务是一种“大数据”规模的元数据搜索服务,通过支持图书馆自动化系统,实现数据融合与语义检索服务、跨机构元数据的开放与复用和分布异构系统的关联访问[11]。资源发现系统是通过从出版商、内容商、大学图书馆、公开网站等提取各类元数据资源,利用抽取、映射、规范、融合等手段,将印刷型馆藏、数字型(开放)资源、本地自建、远程订购资源的元数据、文摘索引和全文内容等多来源、多格式和多元素的资源进行全面聚合和深度组织,构建具有统一索引功能的资源描述与揭示体系。资源发现系统具备“全媒体”资源管理能力、完整的业务流程管理能力和“全网域”的资源发现能力[12],它在全面揭示和发现元数据资源的基础上,还能多元定位资源,获得出版商、内容商版权许可,与资源订购合同相吻合,不仅整合图书馆信息检索、读者认证、预约续借等传统服务方式,还提供评论标签、全文获取、学术推荐、引文显示、馆藏揭示等新的服务功能,最终让用户在一个界面中实现全部资源的检索、显示和排序、获取,实现多种服务方式的整合,多种获取途径的选择,让更多的用户通过资源出版平台使用发现服务。

资源发现系统不仅是一个学术搜索引擎,更是一个知识发现平台,带来的是一种集成开放的思维,将改变科技信息原有的资源采访方法、数据组织方式和信息服务模式。随着资源发现服务向智能化、语义化、关联化、可视化方向发展,科技文献联合目录系统将被元数据集成整合系统所替代,科技信息资源组织将被知识组织所替代,科技信息服务将被资源发现服务所替代。届时,科技信息资源建设重心势必将转移到自身特色资源建设与应用上来,科技信息资源组织将转移到数据细粒度揭示与加工上来,科技信息服务重心将转移到数据驱动的精准服务和智能分析上来。基于此,建设国家科技信息发现服务体系有利于实施国家科技信息元数据发展战略,推进国家科技信息资源公

益服务的全面持久保障，有利于推动我国图书情报机构参与发现系统构建，建立适应发现服务的数字业务管理新流程，完善既有采访、编目、加工和服务模式。探索元数据的语义知识组织有利于深入到知识单元揭示数据间的知识关联，将语义知识组织思想和工具嵌入发现服务构建中，形成海量数据资源的挖掘利用和知识服务，丰富发现服务功能，拓展知识服务新业态。因此，无论是出于国家科技信息战略保障的需要，还是出于满足用户多元化的信息需求，我国都应基于语义化知识组织，自主构建国家科技信息发现服务体系，深化国家科技信息资源保障的内涵和能力。

2　基于语义的国家科技信息发现服务体系框架

国家科技信息保障体系中，除馆藏印本资源外，还包含商业电子资源、开放获取资源，甚至其他相关机构的特色资源，国家科技信息发现服务体系需要多渠道采集与整合各类型的科技文献信息资源，并从资源发现逐步拓展到知识发现层面。发现服务体系的建设涉及文献数据采集与规范、数据组织与集成、业务系统搭建与服务体系构建等一系列工作流程，同时需要相应的语义组织工具、标准规范来支撑资源组织及其语义关联（见图 1）。

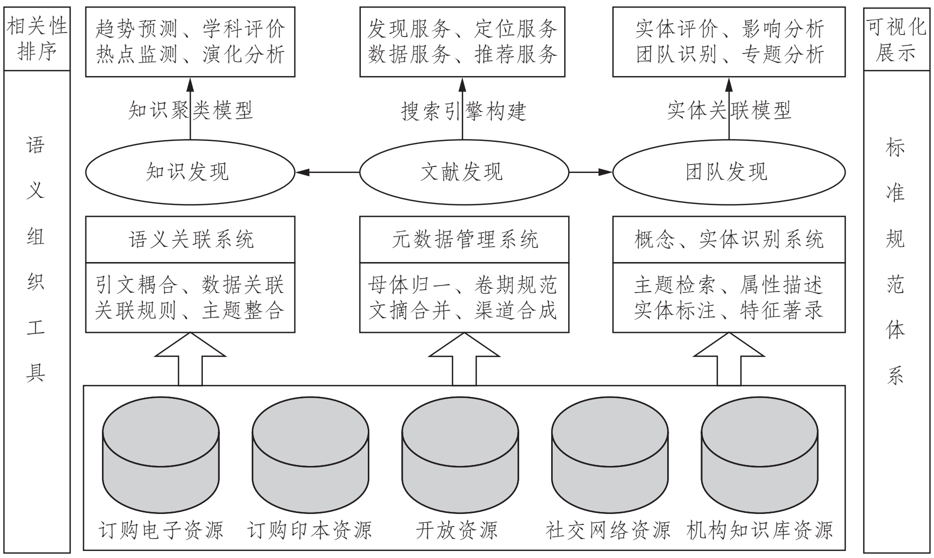

图 1　基于语义的国家科技信息发现服务体系框架

2.1　文献数据采集层

目前科技文献资源类型和来源主要分为商业出版社资源、图书馆馆藏印本资源、开放获取资源、机构知识库资源/特色资源、社交网络资源等。在文献信息采集层面，需要建立科技信息资源复合型采集渠道，在版权协议框架下，通过印本加工、网络采集、谈判购买、免费获取、呈缴、赠送等多种方式采集元数据和全文。此外，对各图书馆馆藏数据信息（如 MARC 数据、数据库开通信息）、机构知识库资源/特色资源元数据也需通过签署相应授权许可协议加入到底层元数据仓储中。

2.2　数据统一管理层

统一的资源检索及其全文指引是国家科技信息发现服务体系的基础功能，而其实现的前

提是对多渠道、多类型、多载体的元数据进行集成规范与统一管理,形成同构的标准化数据。元数据集成管理系统需要基于统一的元数据标准框架,对不同来源元数据进行映射,通过规范、校验、清洗、查重等一系列工序,分别从母体、卷期、篇级、来源渠道等维度对资源实施查重与归一,并根据资源粒度大小形成相应的资源层级结构,提高基础数据质量。

2.3 语义知识组织层

对多来源异构元数据经过清洗规范形成统一的元数据格式以后,需要经过语义化加工组织形成相应的知识资源,满足不同应用场景下的服务需求。在语义模型、规范文档、叙词表等知识组织工具支撑下,开展科技信息资源语义化加工,诸如概念标注、科研实体辨识、语义知识关联与耦合等,通过知识抽取与语义关联形成知识库。概念标注与实体辨识系统主要是为解决同一概念、实体在不同语境下表达不同、人名重复、机构名变更等问题而构建的标注系统,确保文献资源的知识属性能够被准确表达;语义知识关联与耦合系统则是借助数据挖掘、语义计算等技术,实现大批量自动分析,计算资源本身及其引文在主题、作者等方面的语义关系,构建知识网络。

2.4 服务功能实现层

国家科技信息发现服务体系建设需要突破现有发现服务的基本框架,基于语义化组织与资源计算,深化和拓展服务功能,从基于关键词匹配的一般性资源集成检索向语义检索、知识导航、文献计量分析、科研网络构建等方面拓展,尽可能以可视化方式呈现给用户,将海量知识的非结构化数据转换成视觉形式,以图形、图像形式展示知识之间错综复杂的关系和潜在联系,揭示知识的深层内涵[13],促进隐性知识向显性知识的转换。

2.5 方法工具支撑层

无论是数据最初的清洗规范、查重归一,还是后期的语义化知识组织,标准规范体系和语义组织工具都是支撑后台子系统运行和底层数据资源处理的关键。因此,需要制订包括统一元数据框架、资源遴选规则、元数据映射规则、资源查重规则、资源校验规则等标准规范,需要应用包括资源语义模型、机构规范文档、叙词表、资源推荐模型等语义组织工具。此外,概念映射、语义标注、数据挖掘、数据清洗等一系列自动化工具的开发也是国家科技信息发现服务体系构建的重要组成部分。

3 多来源元数据的统一集成管理

发现系统元数据来源各异,各来源原始数据遵循的标准格式不同,元数据薄厚程度不一,甚至存在数据内容错误等情况,需要建立元数据集成管理系统进行元数据收割/导入、转换、校验、集成,对通过各渠道采集来的元数据进行分门别类的管理,按照情报工程化原理建立数据之间的联系,实现数据输入、质量控制、数据输出、关联分析等流程化处理[14],具体如图 2 所示。

3.1 统一的标准规范制订

元数据集成管理的最终目的在于将描述同一论文资源的元数据在保留来源渠道的前提下进行查重合并,保证检索结果的准确性,结果唯一且能够提供来源渠道信息。制订统一的元数据标准规范是进行查重比较的首要前提。目前,发现服务的元数据主要由出版社(如 Wiley、Elsevier、Taylor & Francis 等)和集成商(如 Thomson Reuters 等)协议提供,以及从开放获取平台自动采集,但是不同来源的元数据格式呈现各自为阵的态势,它们在字段命名、字段结构、字段内容表达方面的规则千差万别,在元数据字段厚度方面也参差不齐,因此需要在遵循期刊文章标签集(Journal Article Tag Suite,JATS)等国际标准的前提下,建立统一的元数据框架,并与各来源元数据标准建立映射关系,确保元数

据在格式上统一。在字段内容层面，同一出版社名称、期刊名称、作者机构名称等在不同来源中有不同的表达方式，因此还需要建立各种类型的名称规范文档，将规范名称与非规范名称有序集中并建立对应关系，用唯一 ID 号进行标识，确保字段查重以及后期资源检索结果的一致性。

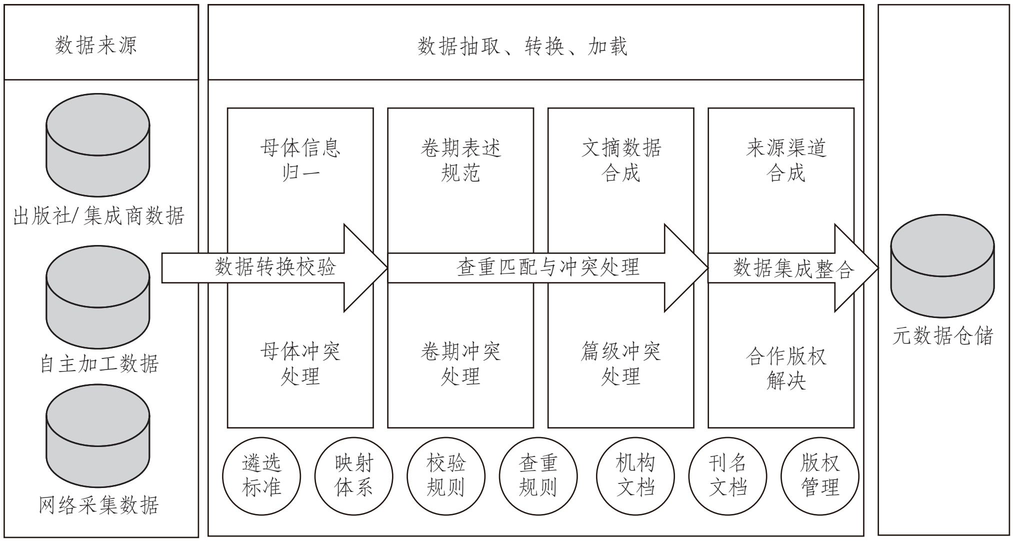

图 2　元数据集成管理系统框架

3.2　数据转换、校验、清洗、规范

标准各异的元数据在导入元数据集成管理系统之前，首先需要根据前期来源渠道的元数据标准管理信息，按照系统设定的规则与统一的元数据标准进行字段映射和格式转换，使之按照统一标准格式描述资源信息。其次，数据转换后经 Schema 形式校验，进行数据逻辑验证和媒体对象损坏验证，任何一步验证失败都要被退回，经过机器和人工修正后再次进行校验，直至完全通过。最后，校验数据在系统内嵌的各类型规范文档的作用下，通过名称识别并归一重名、别名、名称繁简、名称演变等，将不同形式的表达统一映射和指向同一对象。

3.3　数据查重与归并

不同类型的文献资源具有不同的描述粒度，其中以期刊的粒度层级最为丰富，自上而下包含母体、卷期和篇级元数据三个层面，同时多来源渠道成为期刊元数据的另一个重要标识。因此，需要从母体、卷期、篇级元数据和来源渠道四个方面对元数据进行查重、归并。其他文献资源类型的查重体系与期刊基本类似。

（1）母体层级：通过校验和修正的数据可根据不同类型资源的查重规则，通过机器算法自动筛选出疑似重复的母体资源；根据机器筛选的结果进行人工比对，对重复资源进行母体合并、编辑，对不重复而相似资源进行人工标记，用于再次疑似重复的排查参考。

（2）卷期层级：系统自动对冲突母体的卷期信息以及卷期下设篇级信息进行冲突合并，对于不同卷期表述的情况则需要人工进行核实和判断。

（3）篇级层级：对同一母体、同一卷期下篇级元数据进行查重和冲突处理，根据查重规则，当 DOI 不为空时，同一母体下相同 DOI 且年卷期、起始页码、总页数、文摘任意一组完全一致

即视为重复冲突;DOI为空时,同一母体下,元数据标题进行左右匹配且起始页码、总页数完全一致视为重复冲突。机器自动筛查后,以人工方式再次进行比较判断,确定是否为同一篇级信息并进行相应操作。

(4)来源渠道层级:来源渠道随篇级元数据一同进行删除或合并,并根据版权规定和使用约定、来源渠道可信任程度、数据质量进行优先级排序。

3.4 馆藏配置与资源调度

元数据对应印本文献、商业数据库、开放资源等资源类型,不同元数据所反映的资源对象(全文)受自身资源特性、版权约定、来源渠道影响,使用方式有所不同,直接影响到后期全文服务的提供效果。因此需要赋予其对应的馆藏信息和使用权限信息,如包含印本资源馆藏单位的OPAC信息、商业数据库开通单位的开通数据库品种、可访问期刊品种、可访问年限、开放资源的版权使用许可信息等,并根据用户身份及其使用环境,按照版权约定、来源渠道信任程度、数据质量等进行来源渠道优先级排序,准确、合理地提供全文指引和获取服务。根据用户使用环境识别用户身份,基于情景敏感进行用户可使用资源的配置,根据不同用户类型及资源方的权益要求,建立用户文献资源获取的流程、范围及策略(如服务流程、版权管理策略等),建立资源调度知识库,实现不同类型数字资源和不同服务方式之间的开放动态链接,提供从文献发现到文献传递之间的无缝链接,提高资源利用率和区域保障能力。

4 大规模元数据的知识化关联揭示

海量科技文献资源的知识抽取与集成,乃至后期知识服务的形成均有赖于有效的知识组织。基于领域本体、主题词表、术语表等知识组织工具,在规范、可控的概念语义体系与自然语言术语体系之间建立映射关系,通过实体辨识、概念标注、语义知识关联与耦合对抽象的知识单元进行规范化、关联化描述,从而实现大规模聚合信息资源,形成结构清晰的语义层面知识,实现知识发现[15](见图3)。

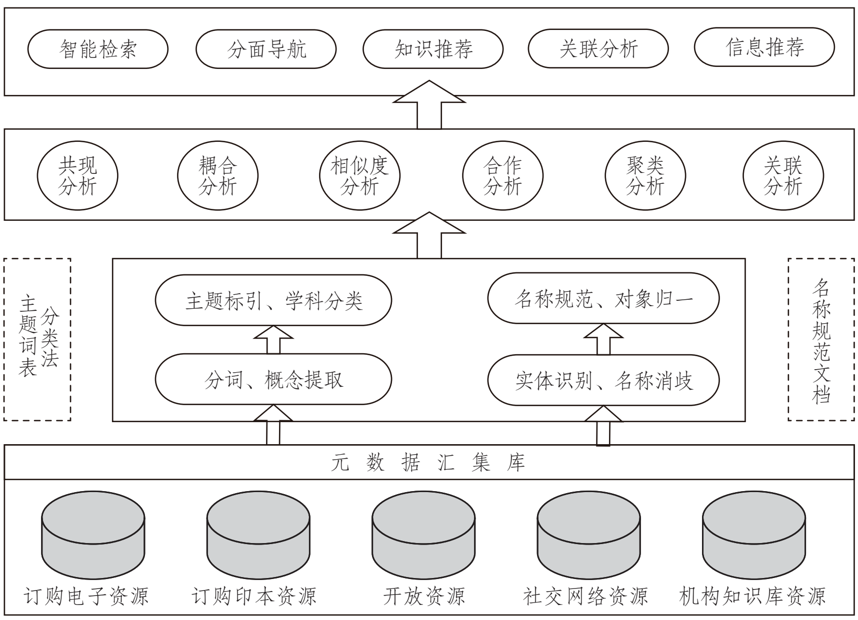

图3 语义知识组织框架

4.1 概念标注

人类理解自然语言的过程是一个语义概念的联想和关联的过程，建立基于概念的文本表征模型是实现基于语义的文本内容处理的途径[16]。一般来说，概念标注方法主要分为手工和自动两大类。手工概念识别主要依靠专家实现，存在时间和人工成本比较高的弊端，难以适应大规模科技文献资源的概念识别。因此，国家科技信息发现服务采用“机器为主，人工辅助”的方式进行概念标注，以领域本体及其表达概念为基础，自动标注科技信息中包含的相关概念及其语义形式。

4.2 实体辨识

科研实体包含作者、机构、基金、主题、论文及引文等，系统辨识科研实体的准确性有赖于建设各类科研实体的规范文档，细分为期刊论文与引文库、期刊名称库、作者专家库、机构名称库、基金项目库等，围绕着期刊论文数据，论文主题、作者、作者机构、期刊和基金等科研实体相互关联[17]。基于统一的规范文档，通过相关论文题名、作品主题范围、标识符（ISBN、ISSN）以及作者、合著者等匹配确认实体归属。

4.3 语义关联

用户期望真正意义上的“资源发现”，不仅要解决异构资源间的互操作问题，实现多媒体多类型资源的多维度、多脉络揭示，更要实现基于语义关联的文献推荐及知识发现。因此需要在抽取概念和实体等知识单元的基础上，通过知识的语义关联与引证耦合，应用关联数据、语义技术等，从语义层面进行科技信息资源的组织和聚合，实现真正意义上的知识关联和知识发现。

（1）构建文献关联关系。以文献资源中抽取的知识单元为基础，按照不同学科领域，利用不同领域本体、主题词表等知识组织工具进行数据映射与链接，建立和扩展更为丰富的知识关联，并与书目数据对应的对象相匹配，进行资源聚类和分面，扩展知识发现范围。

（2）揭示科研实体关系。通过元数据规范化处理和语义关系揭示，提取文献资源中的科研实体和实体间关系，形成“机构—人员—科研成果”关系图，将大规模文献资源中所隐含的科研合作关系、机构合作主题等知识内容充分揭示出来，从而实现资源的深层次聚合和专家团队识别。

（3）计算共现耦合关系。科技文献中存在相互引用的耦合关系、关键词共现关系、作者合作关系、主题相似关系和学科交叉关系等，利用耦合分析、关联分析、相似性算法、聚类分析和共现分析方法可以分析科技文献中作者、期刊、论文、机构和主题间的关联关系，从而为分面导航提供相应维度的分类层级，为知识推荐提供交叉关联关系。

5 国家科技信息发现服务体系展望

国家科技信息发现服务体系中既包含多渠道、多媒体、多类型海量元数据的集成与聚合，又包含用户体验的多层次、多模式服务功能的一站式融合，还包括资源采集、编目、流通、原文传递，以及用户认证和用户管理等相关系统的一体化对接。因此，不能将国家科技信息发现服务体系简单理解为一个独立的文献资源集成检索系统，它融通图书馆自动化系统的前台服务系统、后台业务系统，是图书馆文献采集、数据组织、用户管理、功能实现乃至各层次信息服务的集中体现。同时，具有国家保障职责的科技信息发现服务体系，不仅要实现国家级科技文献保障机构数字业务流程的互通互联和数据流转，也需要联合全国范围内的高校图书馆、科研图书馆进行资源共享与服务保障的协同管理（见图4）。

5.1 科技文献数字业务管理系统的整合

我国目前建有诸如国家高等教育文献保障体系（CALIS）、国家科技图书文献中心（NSTL）、中国高校人文社会科学文献中心（CASHL）等国家级文献保障机构。以NSTL为例，其本身拥有

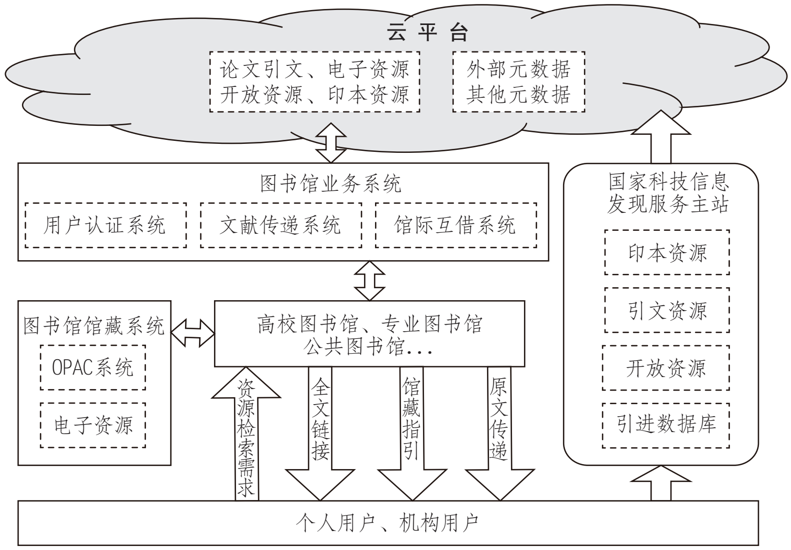

图 4　国家科技信息发现服务体系

文献综合管理系统、联合编目系统、数据加工系统、文献传递服务系统等一套完整的图书馆业务流程系统,近年来除采购大量印本资源外,不断开拓全国回溯数据库采购、开放获取资源建设等多种资源保障模式,但是无论是业务系统还是服务系统均未完全实现互联互通,不乏信息孤岛存在,在一定程度上造成业务流程重复,数据冗余。因此,亟待基于云架构,针对原本分散的数据资源、工作流程和服务系统进行统一整合和协同管理,通过数据集成、流程整合、系统互联形成统一的工程化管理模式和一体化服务入口,以提升管理效率,优化数据质量。

5.2　科技信息资源建设与服务力量的整合

国家科技信息发现服务体系中的文献资源不应仅包括国家保障体系自身拥有的全文资源,同时还要发挥全国范围内相关文献机构资源共建的优势。近年来,我国高校图书馆、公共图书馆、专业图书馆采集了大量的电子资源和印本文献,一些科研机构建立了颇具特色的机构知识库,通过构建国家科技信息发现服务体系,将这些图书馆的馆藏资源(如印本 MARC 数据、数据库订购信息、特色资源列表)信息定期纳入/更新至国家科技信息资源仓储中,将有利于全国范围内对科技信息的全方位检索发现;将这些图书馆的原文传递服务系统与国家科技信息资源调度系统对接,对用户发现检索请求进行文献传递服务回应,能够提升全国范围内全文获取的保障率。同时,以服务对等、费用补贴等方式予以结算,将资源与服务共享层级延伸至相关机构,能够整合各级图书文献机构的服务能力,调动全国范围内的资源与服务力量参与资源共建共享。

5.3　科技文献资源多层级服务功能的整合

目前国家科技信息保障服务主要局限于自有资源的检索及其全文传递,国家科技信息发现服务体系的建立,不仅强化文献发现服务,而且推进知识服务和数据服务,实现多层级服务功能的整合。在数据服务层面,通过第三方数据接口服务实时调用相关学者、机构的科技成果信息,建立机构知识库,将元数据嵌入诸如科研管理系统中,保证科研管理部门在项目论证、项目申报等各个环节直接获取相关文献信息。在信息服务层面,用户可统一检索各类型资源,搜索引擎根据词频、近义词的模糊归类、关键词

匹配位置等影响因素设置权重，对检索结果进行相关性排序，提供学科、作者、机构等分面显示功能，并根据相关学者、主题等推荐文献资源。系统通过感知用户使用环境提供全文获取途径，对无法直接获取的全文提供原文传递、代查代借等服务。在知识服务层面，利用元数据的规模优势以及数据的语义关系组织，通过科研实体抽取、引文网络耦合、规范文档建立、概念主题识别、资源解析关联等操作，形成“科研人员—科研机构—学科主题”的各类关联知识网络，相关科研实体的科研表现和科研活动脉络在可视化技术的呈现下清晰展现，实现科研团队识别、科研绩效评估、主题趋势分析等。

6　结语

国家科技信息发现服务体系建立在经语义知识化处理的海量元数据及其全文资源的获取指引基础之上，而发现服务体系的有效运行则有赖于统一的标准规范体系建立和知识产权问题的妥善处理。在文献数据采集阶段需要明确各来源渠道的使用权限，通过统一的标准规范实现多来源数据资源的规范化揭示与集成，在全文指引获取阶段则需要遵循资源使用权限，实现在符合知识产权相关规定前提下的文献资源共享服务。与此同时，伴随数字资源不断扩充、资源链接地址的失效、母体信息的变动、资源载体形式的变化、资源类型的演化，以及新型资源形态的出现等，发现服务体系应同步更新，定期维护，建立从版权谈判与授权、数据获取、数据集成和故障追踪的全流程资源维护体系，形成多来源数据采集更新机制，构筑既符合版权规定，又能够利益均衡的全文获取共享机制，确保发现服务的稳定性和可持续性。

参考文献

[1] 彭以祺，吴波尔，沈仲祺. 国家科技图书文献中心“十三五”发展规划[J]. 数字图书馆论坛，2016(11)：12-20.(Peng Yiqi，Wu Boer，Shen Zhongqi. The 13th Five-Year Plan for the development of National Science and Technology Library[J]. Digital Library Forum，2016(11)：12-20.)

[2] Cathro W，Collier S. Developing trove：the policy and technical challenges [R/OL]. [2016-03-13]. http://pandora. nla. gov. au/pan/21336/20110629 - 1337/www. nla. gov. au/openpublish/index. php/nlasp/article/download/1666/2025html.pdf.

[3] 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.(Dou Tianfang，Jian Airong. Function analysis and application prospects of resource discovery system[J]. Library and Information Service，2012，56(7)：38-43.)

[4] 聂华，朱玲. 网络级发现服务——通向深度整合与便捷获取的路径[J]. 大学图书馆学报，2011，29(6)：5-10，25.(Nie Hua，Zhu Ling. Web-scale discovery service：on the way to a deep integration and easy access[J]. Journal of Academic Libraries，2011，29(6)：5-10，25.)

[5] 袁玉英. 常用几种资源发现系统对比分析研究[J]. 图书馆工作与研究，2015(9)：38-41.(Yuan Yuying. Research on the comparison on the most commonly used resource discovery systems[J]. Library Work and Study，2015(9)：38-41.)

[6] 苏建华. 图书馆选择资源发现系统的策略分析——以资源发现系统与学术搜索引擎的比较为视角[J]. 情报科学，2015(6)：91-94，105.(Su Jianhua. Strategy analysis of library selection of resource discovery system：a comparative perspective of resource discovery system and academic search engines [J]. Information Science，2015(6)：91-94，105.)

[7] 彭佳，郑巧英. 信息资源聚合与组织研究——以发现系统为例[J]. 图书馆杂志，2016(3)：80-85.(Peng Jia，Zheng Qiaoying. Study on integration and organization of information resources based on information resources

discovery systems [J]. Library Journal,2016(3):80-85.)

[8] 许丽媛,吴振新,谢靖. 图书馆主流资源发现平台分析[J]. 图书馆学研究,2015(8):33-38.(Xu Liyuan,Wu Zhenxin,Xie Jing. Analysis on mainstream discovery service of library [J]. Researches in Library Science,2015(8):33-38.)

[9] 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.(Bao Ling,Jiang Ying. Comparative study of library uniform resource discovery systems[J]. Information and Documentation Services,2012(5):67-72.)

[10] 中国科学院文献情报中心选用 ProQuest Summon 发现服务支撑科研信息查找[J]. 图书情报工作,2015,59(2):146.(National Science Library of Chinese Academy of Sciences selected the discovery service of ProQuest summon for academic information retrieving[J]. Library and Information Service,2015,59(2):146.)

[11] Summon 网络级发现服务简介及功能说明[EB/OL]. [2017-01-13]. http://lib.csu.edu.cn/pubnew/zndxtsgnew/summon/Summoncpjs.pdf.(Introduction and function description of summon [EB/OL]. [2017-01-13]. http://lib.csu.edu.cn/pubnew/ zndxtsgnew/summon/Summoncpjs.pdf.)

[12] 殷红,刘炜. 新一代图书馆服务系统:功能评价与愿景展望[J]. 中国图书馆学报,2013,39(5):26-33.(Yin Hong,Liu Wei. On the new generation of library service system:functions,visions and prospects[J]. Journal of Library Science in China,2013,39(5):26-33.)

[13] 贺德方,曾建勋. 基于语义的馆藏资源深度聚合研究[J]. 中国图书馆学报,2012,38(4):79-87.(He Defang,Zeng Jianxun. Study on in-depth integration of library collections based on semantics[J]. Journal of Library Science in China,2012,38(4):79-87.)

[14] 杨新涯. 图书馆文献搜索研究[M]. 重庆:重庆大学出版社,2015:115.(Yang Xinya. A study of library bibliographic search[M]. Chongqing:Chongqing University Press,2015:115.)

[15] 盛东方,孙建军. 基于语义搜索引擎的学科知识服务研究——以 GoPubMed 为例[J]. 图书情报知识,2015(4):113-120.(Sheng Dongfang,Sun Jianjun. Research on subject knowledge service based on semantic search engine:a case study with GoPubMed[J]. Document,Informaiton & Knowledge,2015(4):113-120.)

[16] 涂新辉,何婷婷,李芳,等. 基于排序学习的文本概念标注方法研究[J]. 北京大学学报(自然科学版),2013,49(1):153-158.(Tu Xinhui,He Tingting,Li Fang,et al. Learning to rank concept annotation for text[J]. Acta Scientiarum Naturalium Universitatis Pekinensis,2013,49(1):153-158.)

[17] 曾建勋,王立学. 面向知识评价的规范文档建设方法[J]. 图书情报工作,2012,56(10):101-106.(Zeng Jianxun,Wang Lixue. Construction of knowledge evaluation-oriented authority files [J]. Library and Information Service,2012,56(10):101-106.)

曾建勋 中国科学技术信息研究所研究馆员,博士生导师。北京 100038。

丁遒劲 中国科学技术信息研究所馆员。北京 100038。

(收稿日期:2017-03-14)