多源异构科技资源元数据构建研究及应用实践

毛维娜 于怡鑫 苗润莲 童爱香
（北京市科学技术研究院，北京 100044）

摘要：为推动区域科技资源数据高效开放与合理共享，促进区域协同创新发展，从京津冀科技资源共建共享视角，以科技机构为纽带，将科技人才、科技成果、科研项目、基础设施等科技资源相关联，构建区域科技资源基础元数据标准，利用OAI−PMH协议对元数据进行收割管理，并构建平台，实现多源异构数据的汇集，对跨库数据进行关联检索及可视化展示，为区域科技资源开放共享提供经验借鉴。

关键词：科技资源；元数据；OAI−PMH协议；数据汇交；京津冀
DOI：10.3772/j.issn.1674−1544.2024.01.007 **CSTR：**15994.14.issn.1674.1544.2024.01.007
中图分类号：F124.3 **文献标识码：**A

Research and Application Practice of Metadata Construction of Multi-source Heterogeneous Scientific and Technical Resources

MAO Weina, YU Yixin, MIAO Runlian, TONG Aixiang
(Beijing Academy of Science and Technology, Beijing 100044)

Abstract: In order to promote the efficient openness and reasonable sharing of regional scientific and technical resources and data, and promote the collaborative innovation and development of regions. From the perspective of co-construction and sharing of scientific and technical resources in the Beijing-Tianjin-Hebei region, scientific and technical institutions are used as the link. scientific and technical resources such as scientific and technical talents, scientific and technical achievements, scientific research projects, infrastructure and other scientific and technical resources are linked. Regional scientific and technical resources basic metadata standards are constructed, and the OAI-PMH protocol is used to harvest and manage metadata. It realizes the collection of multi-source heterogeneous data, and carries out associated retrieval and visual display of cross-database data through platform construction, which provides experience for the open sharing of regional scientific and technical resources.

Keywords: scientific and technical resources, metadata, OAI-PMH, data collection, Beijing-Tianjin-Hebei

0 引言

党的二十大报告提出，要“促进区域协同发展”，“深入实施区域协同发展战略、区域重大战略”，“高质量发展区域经济布局和国土空间体系”。科技资源是区域创新和可持续发展的战略

作者简介：毛维娜（1986—），女，北京市科学技术研究院助理研究员，主要研究方向为区域创新发展；**于怡鑫**（1981—），女，北京市科学技术研究院助理研究员，主要研究方向为城市治理、产业经济（通信作者）；**苗润莲**（1969—），女，北京市科学技术研究院主任、研究员，主要研究方向为区域协同发展；**童爱香**（1986—）女，北京市科学技术研究院助理研究员，主要研究方向为区域经济产业协同。

基金项目：北京市财政项目“基于大数据的京津冀协同创新发展指数研究及分析”（23CB070）。
收稿时间：2023 年 4 月 23 日。

资源，是科技活动有效开展的基础，是推动整个经济和社会发展的要素集合，具有原始创新基础和创新驱动引擎的重要作用[1]。2022年年底颁布的《中共中央 国务院关于构建数据基础制度更好发挥数据要素作用的意见》（以下简称“数据二十条”）明确提出，数据作为新型生产要素，是数字化、网络化、智能化的基础，已快速融入生产、分配、流通、消费和社会服务管理等各环节，深刻改变着生产方式、生活方式和社会治理方式，要加强数据汇聚共享和开放开发，强化统筹授权使用和管理，推进互联互通，打破“数据孤岛”[2]。科技资源开放共享可以有效消除区域信息孤岛，帮助政府人员了解区域科技资源现状，为政府决策提供信息支撑；帮助科研人员缩短科研时间节省经费，提高科研效率；帮助企业快速了解同行业同领域技术进展、最新动态、前言热点，挖掘出解决问题的关键技术，最终形成智慧产品[3−6]。作为科技资源共建共享基础的科技资源元数据，是实现科技资源有效发现、快速检索、精准获取以及交换和整合的重要手段，是推动数据开放共享的前提条件和基本保障。

1　研究现状

国外描述元数据主要集中于网络资源、文献资料、藏品以及视频等各类数字资源对象。其中，影响最广的都柏林核心（Dublin Core，DC）元数据，是一种能很好地解决网络资源的发现、控制和管理问题的一种元数据[7]。由于DC具有较强的兼容性，很多科学数据的存储将DC作为基础，如英国eBank UK、DISC−UK DataShare以及荷兰的CCLRC等[8]。还有MARC DTD系列、MODS、Z39.50 profiles、ONIX、ETD−MS等描述文献资料的元数据[9−13]。这些数据是记录在载体上的信息资料，主要包括图书、报刊、特种资料、档案等。另外，还有CDWA、VRAcore等描述藏品的元数据，以及描述新视频资料的MusicBrainz MetaData Initiative元数据[14]。这些元数据具有一个相同的特点，就是有共同的核心元素，并在核心元素的基础上进行扩展、重复使用已经成熟的元数据标准中的一些元素，或者直接使用其他元数据标准，这也是互操作共享资源的趋势。

国内针对科技领域数据的标准化研究已经形成了3套比较完整的标准规范体系，即全国科技平台标准化技术委员会科技资源标准体系、国家科技基础条件平台科学数据标准规范体系以及中国科学院科学数据标准规范体系[15]，为推进数据标准工作奠定了基础。按照“资源共享，平台先行”的原则，国家在科技平台建设工作方面先后制定了1 000余项技术标准和规范，覆盖了实验基地、科学数据、科技文献、仪器设备等领域，为科技资源规范化管理和共享发挥了重要作用。随着科学数据的快速增长，科学数据的开放共享面临新的挑战，如数据来源广、格式复杂多样、字段命名存在差异性。同时，科技资源具有较强的专业性、结构复杂[16−18]，通过简单的数据融合方案很难支撑区域科技创新资源的融合。其主要表现为以下3个方面：一是随着信息技术的发展，科技资源的数据量呈指数级增长，数据冗余性大，需要专业的技术进行筛选合并；二是区域科技资源的数据来源广，数据表示形式多样化，字段属性定义差异较大，异构现象明显，无法直接入库检索应用；三是由于数据分布于不同的区域，数据格式定义存在差异，数据汇交存在困难。

基于上述分析，鉴于区域多源异构数据融合难的问题，急切需要通过核心要素将各类科技资源相关联，实现不同区域、不同数据源之间的数据兼容。本文基于现有研究，以科技机构为纽带，将科技人才、科技成果、科研项目、基础设施等科技资源相关联，实现多源异构数据之间的关联，并依据这个关联关系构建京津冀区域科技资源元数据，利用OAI−PMH实现区域科技资源的融合汇聚，保证元数据的可扩展性，为区域科技资源开放共享奠定基础。

2　区域科技资源元数据标准体系

科技资源元数据使用的目的在于识别、评价和追踪资源，简洁高效地管理大量网络化数据，而实现信息资源一体化组织和有效管理，是实现

科技资源有效发现、共享、交换和整合的重要手段[19]。

2.1 科技资源元数据作用

科技资源元数据通常被用来描述科技资源的数据要素、范围、管理方式、数据来源、数据的提供方式等相关的信息。其作用可以归纳为以下4个方面。

（1）描述数据特征。科技资源元数据可以准确详细地描述资源数据的内容、属性、定义规则、值域等信息，可以较为完整地反映科技资源对象的特征。

（2）数据集成的基础。从数据源采集获取的数据需要按照特定的规律存入数据仓库中，数据源与数据仓库中的数据对应关系及变化规则离不开元数据知识库的支撑，如果在数据库建设过程中，忽视了元数据结构的管理，数据融合可能会比较困难。

（3）辅助用户理解。元数据是实现业务功能与数据集之间的映射，可以把数据以用户需求的方式进行可视化展示，帮助用户理解和使用数据，是实现科技资源的安全访问、准确获取、高效利用的主要途径。

（4）支持需求变化。随着信息技术的发展，会不断产生新的数据，并根据业务数据需求，需要不断更新变化数据字段。元数据管理系统可以把整个系统的业务流、数据流和信息流有效地整合在一起，使系统功能属性不再依赖于特定的开发人员，进而提高系统的可扩展性。

2.2 科技资源元数据标准体系构建

作为构建科技资源信息数据集和科技平台元数据的最基本的信息单元，科技资源元数据必须通过制定概念准确、格式统一的元数据标准对科技资源信息进行科学、准确、规范的描述，确保具有相同概念的元数据在语义上的统一、规范和无歧义。为了实现区域科技资源数据的开放共享，还需要统一规范元数据的存储形式、汇交形式等。建立科技资源元数据标准体系，既可以保证科技资源信息的规范性、准确性和实用性，又可以最大限度地消除科技资源数据描述混乱的现象，有助于快速提高科技资源整合、开放共享的效率和质量[20]。

在已有专家研究经验的基础上，结合科技资源数据特点，从元数据定义基础、元数据内容、元数据检索以及元数据汇交管理4个层面构建科技资源元数据标准体系，如图1所示。其中，科技资源元数据定义基础包括元数据定义的依据、元数据注册管理、元数据编辑管理3个方面；元数据内容包括科技资源元数据、用户管理元数据以及系统管理元数据3个方面；元数据检索包括元数据检索方法管理以及元数据检索协议管理两

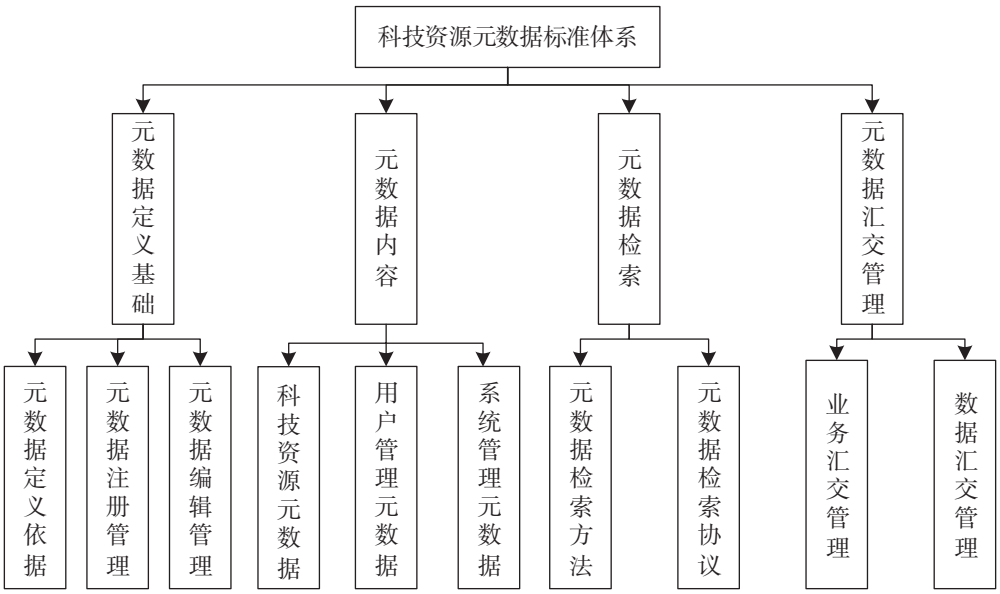

图1 科技资源元数据标准体系

个方面；元数据汇交管理包括业务汇交管理、数据汇交管理两个部分。通过元数据标准体系设计及管理，形成整个平台数据资源的准确视图，完善对数据的解释、定义，形成区域范围内一致的数据定义，并可以对数据来源、变迁等情况进行跟踪分析。

2.3　科技资源数据元素及修饰词定义和著录规则

在上述科技资源元数据标准体系的支撑下，依托国家标准《科技平台资源核心元数据》，结合京津冀区域科技资源特点，对平台科技元素修饰词进行定义，并对著录规则进行说明，为实现基于元数据的跨平台查询检索、元数据汇交提供支撑。

平台科技资源元数据包含标识符、中文名称、英文名称、定义、类型、著录内容及说明、数据类型、可选性、最大出现次数、值域、著录范例11个元素，如表1所示。其中，标识符由标识代号、注册机构标识符、类型代码、内部标识符4部分组成，即STRI（标识代号）+“：”+注册机构标识符（5位）+类型代码（2位）+内部标识符（不定长）[21]；数据类型可以设置为文本类型、数值类型、字符类型；可选性既可以是必选项，也可以是可选项。

3　区域科技资源数据整合设计

广义的科技资源数据包含科技物力资源、科技人力资源、科技财力资源、科技信息资源4个重要组成部分，但这样的科技资源分类还无法直接应用到具体的科技资源实践中。针对现代信息化数据处理的要求，需要将科技资源按照可以存储在计算机中的具体格式进行类别划分，制定更为详细的科技资源分类方法。本文通过实地调研京津冀三地科技管理部门，根据用户对科技资源数据开放共享、高效利用的需求，从数据存储管理及应用视角出发，综合考虑数据可检索性、可获得性以及可分析性，将科技资源数据分为科技机构、科技人才、科技成果、科技项目、仪器设备、统计数据及其他信息。

表1　以科技机构为例对元数据记录进行说明

修饰词定义	著录规则
标识符	由4个部分组成：STRI（标识代号）+“：”+注册机构标识符（5位）+类型代码（2位）+内部标识符（不定长）
中文名称	机构中文名
英文名称	机构英文名称
定义	机构在注册机构登记的中文名称
类型	元素修饰词
著录内容及说明	此处描述机构的正式中文名称
数据类型	文本、数值、字符
可选性	必选，可选
最大出现次数	自然数
值域	字段长度限制
著录范例	示例

3.1　区域科技资源元数据模型设计

基于上述元数据标准体系设计，为了更好地实现科技资源综合检索，需要将不用类型的科技资源通过一个关键要素进行关联，实现多元信息的跨库关联检索。科技机构作为整个科技资源的载体，是科技资源的核心资源，是科技人才工作的平台，拥有基础设施，为用户提供产品或服务。科研人员完成的科技成果隶属于科研机构，科研人才负责的项目需要科技机构承载。所以，以科技机构为纽带，可以实现科技人才、科研项目、科技成果、统计数据等科技资源信息综合检索以及空间关联，如图2所示。

基于上述研究，构建元数据模型结构，用来描述科技机构与各类科技资源之间的关联关系，如图3所示。其中，科技机构与地理信息、基础信息、联系信息、拓展信息、多媒体信息和标签是一一对应的，与科技人才、科研项目、科技成果存在一对多关系，即科研机构可以与多个科技人才、科研项目、科技成果进行对应。科技机构除了具备地理位置、基础信息、联系信息、多媒体以及标签通用属性外，还包括一些机构特有的属性，如高校有学校级别、学校类型、国家重点培育学科等特殊属性，企业有统一社会信用代码、注册时间、登记状态、登记机关、经营规模等特殊属性，这些属性无法全部融合到一起，需要设立单独的扩展信息。

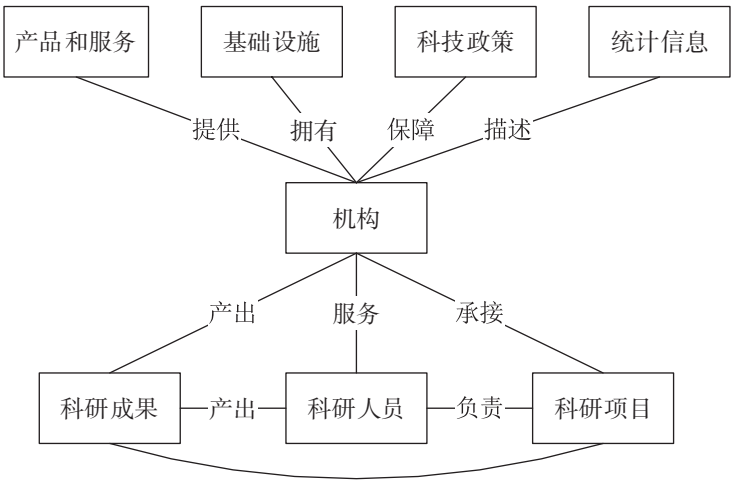

图 2　科技资源之间关联关系

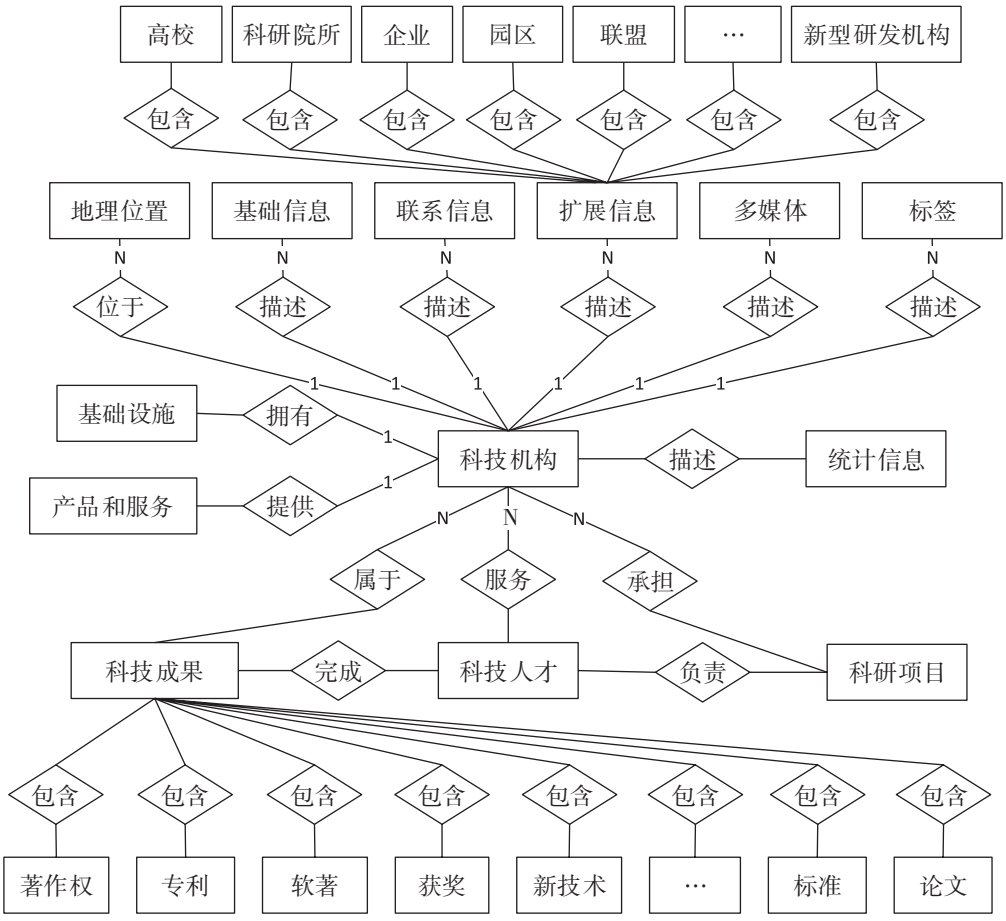

图 3　科技资源元数据关联模型

注：“1”表示 一一对应关系，“N”表示一对多关系。

3.2　区域科技资源元数据设计

整个科技资源元数据结构，采用树状分层结构：第一层为基础元素层；第二层为每一个基础元素大类下分的若干个专属的限定元素，每一个限定元素为描述科技机构元数据的最基本数据单元。

如图 4 所示，以科技机构为例，第一层为基础元素层，分为基本信息、地理位置、联系信

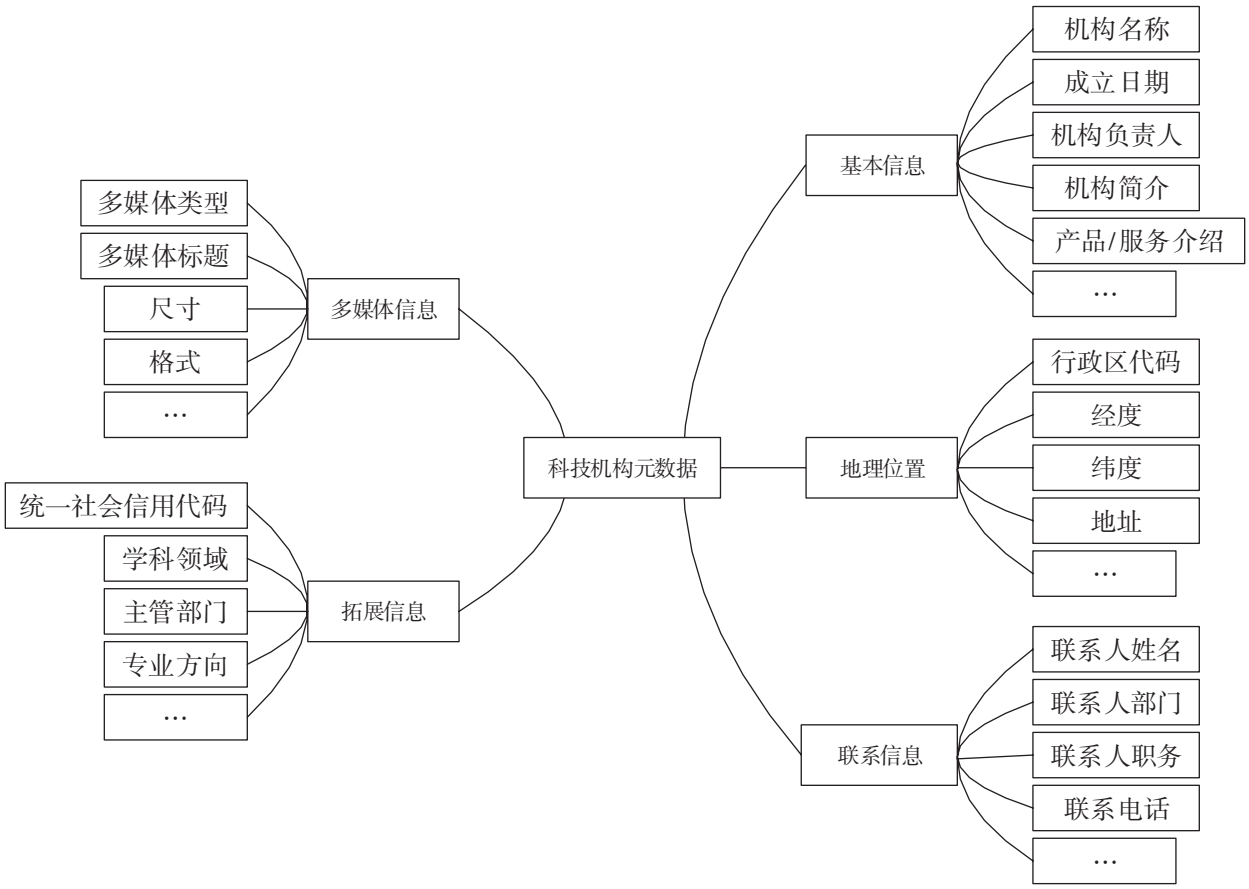

图 4　科技机构元数据

息、多媒体信息、拓展信息 5 个大类；第二层为每一个基础元素对应的限定元素。其中，基本信息包含机构名称、成立日期、机构负责人、机构简介、产品/服务介绍等；地理位置包含行政区代码、经度、纬度、地址等；联系信息包含联系人姓名、联系人部门、联系人职务、联系电话等；多媒体信息包含多媒体类型、多媒体标题、尺寸、格式等；拓展信息包含统一社会信用代码、学科领域、主管部门、专业方向等。为了更好地实现区域科技资源数据可视化展示，方便用户查询及分析，在定义科技资源元数据时除考虑一般信息具有的时效性、可处理性、共享性等特性外，还需要考虑地理因素，在科研机构地理位置要素上增加了两个特殊的元数据，即经度和纬度，用以描述科技机构的地理位置。

4　元数据汇交管理

科技资源元数据可以优化数据集成方式、提升数据安全性，保障数据应用的有效性、高效性，但是随着信息化技术的发展，科技资源数据规模不断扩大，结构多变。同时，在国家开放科学大环境下，越来越多的科技机构或者平台将开放数据。随着数据数量的增长，元数据记录的复杂性和多样性也随之增多。区域科技资源的开放共享，离不开不同区域、不同机构、不同平台的科技资源信息的支撑，而不同来源的科技资源在元数据定义方面可能存在差异，因此如何保障元数据记录的可扩展性是一个重大问题。

OAI−PMH协议是能规范实现网络环境下元数据收割功能的互操作协议标准。这个协议包含数据提供者（DP）和服务提供者（SP）。DP将元数据储存在本地数据仓库中，供SP收割，而SP会定时从DP的数据仓库中收割元数据，进而实现数据增值服务[22]。目前，这个协议多数应用于数字图书馆以及高校间文献资源的共享，如国家科学数字图书馆、北京大学中文古籍数字图书馆

等，均采用了OAI协议实现元数据的互操作[23]。

借鉴我国《国家科技计划科技报告管理办法》[24]中关于科技资源管理的相关规定，构建面向不同区域、不同机构的科技资源元数据汇交管理系统，而系统中的科技资源元数据采用DC格式，如图5所示。

DP必须在系统注册中心进行注册，注册成功后才可以提供元数据。注册中心主要有3个作用：一是汇聚DP的地址信息，根据主题对DP进行分类，方便SP发现DP；二是存储SP的地址信息，并对SP的主题进行分类，方便用户寻找SP；三是DP注册成功后，注册中心将对这个地址进

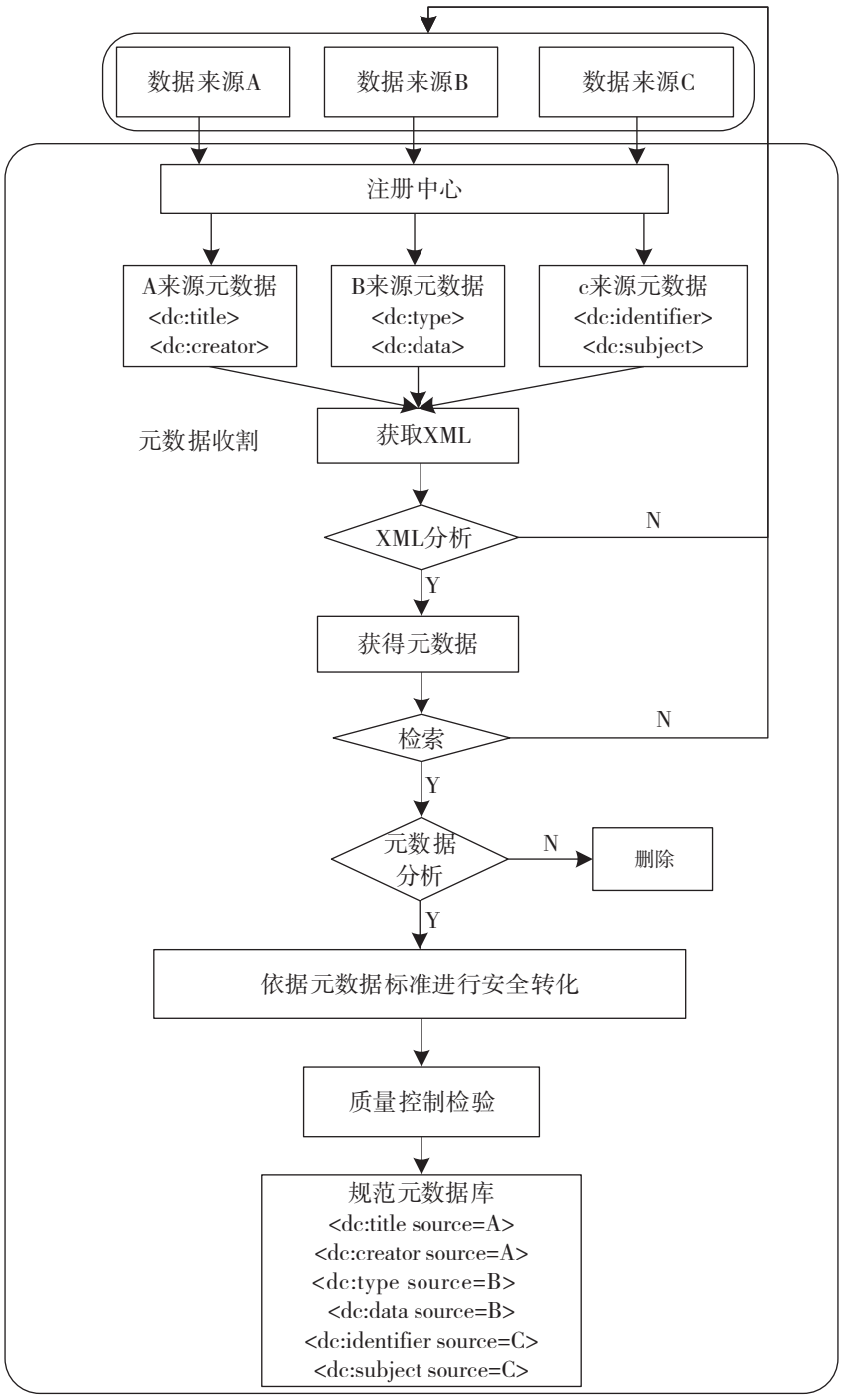

图5　元数据汇交管理

行有效性检验，进而判断DP是否支持OAI−PMH协议。

元数据汇交过程主要是元数据收割过程。SP接收到不同DP的元数据后对DP进行响应。首先，获取数据源中的元数据 XML 文件，如果获取失败则直接反馈到对应的DP，如果获取成功则提取元数据。其次，将提取的元数据在本地元数据仓储中进行检索，去除重复和过期的元数据，并对元数据进行分析。最后，将分析后的元数据依据元数据标准进行安全转化和质量检测，最终将通过质量检测的元数据补入到本地元数据库中，并标明元数据来源，新的本地数据库将作为后续数据检测的数据库。

通过OAI−PMH技术手段实现多源异构数据的汇交，确保了科技资源元数据格式的一致性，规范化后的科技资源元数据既能够确保元数据项的完整、统一，又能够追溯来源，实现兼容和互操作。

5　功能实例应用

在遵循整体性、规范性、易扩展的总体设计思想的基础上，系统采取从数据层、管理层、应用层、功能层和价值层 5 个层级进行优化设计和标准化处理。平台采用SOA架构，建立服务仓库，将服务仓库中来自京津冀各部门的科技资源信息与用户需求反馈结合起来，消除不同应用服务或者不同地域平台之间的技术差异，使不同应用的服务器有需协调运行。平台采用SQL Server 2008 作为数据库管理系统，利用增量备份功能保证数据库的安全性。平台系统总体框架如图 6 所示。

平台系统总体框架分为数据采集层、数据管理层、应用服务层、系统功能层及价值实现层。数据采集层主要是实现对各类科技资源数据的收集、整理、分类，借助数据采集处理工具，按照统一的数据标准规范进行入库[6]。建立规范的数

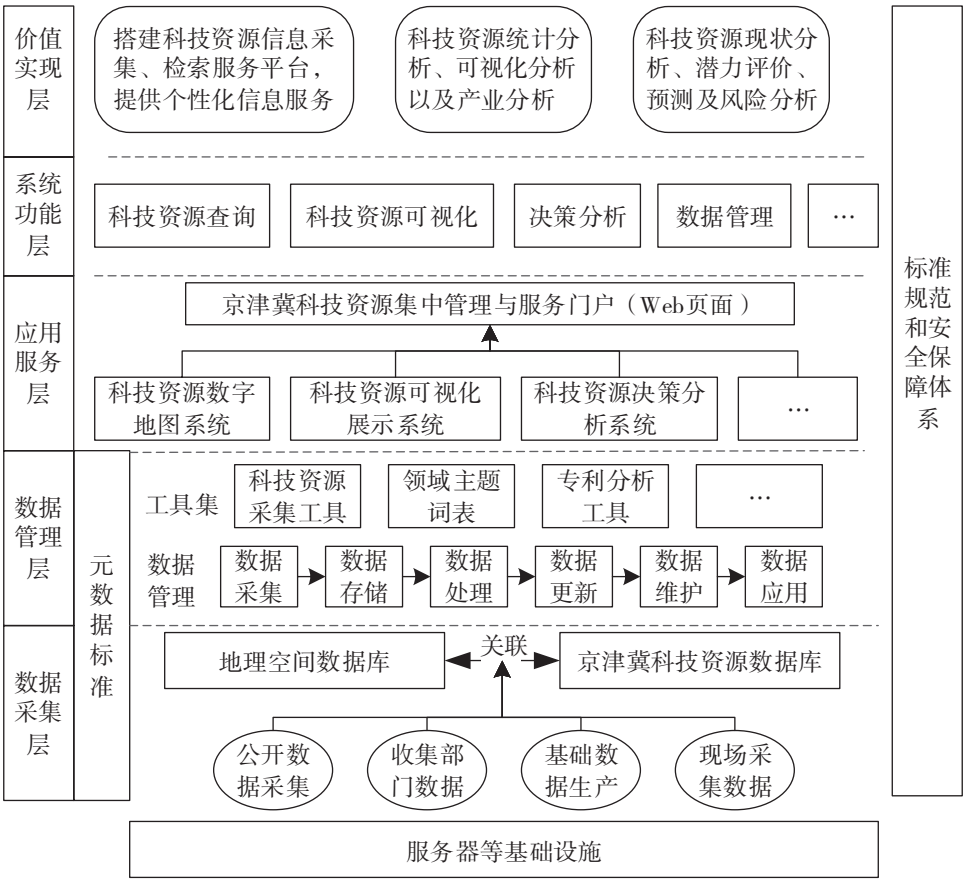

图 6　平台系统总体框架

据库结构，并且实现不同子库之间的关联，通过每个数据表的唯一值字段标识（主键），实现主表和副表的关联。数据管理层主要实现对各类数据资源的管理功能。应用服务层主要实现各类的应用服务，是整个项目的核心内容，由于项目涉及多种服务功能模块，需要有一个综合的门户网站进行功能集成和数据统一共享，使得整个项目的不同部分之间可以紧密地联系和结合，而不是孤立地堆砌功能模块。系统功能层主要实现各种功能的集成。价值实现层是基于多源数据形成的研究成果，为京津冀三地科研机构、政府部门、科技企业和公众提供个性化服务内容，实现科技资源的创新价值。

平台依据建立的科技资源元数据，汇集了京津冀三地科技资源数据，可以实现关键字检索、分类检索以及空间检索。关键字检索可以针对特定区域特定数据库进行综合检索，实现多类科技资源的综合呈现。分类检索可以通过设定的科技资源分类，对科技资源进行跨库检索。如图7所示，以高校检索为例，可以直观地显示京津冀三地的高校分布情况，并将高校相关联的科研项目、科技人才、仪器设备、基础设施等进行综合呈现，打破信息壁垒，实现跨库数据检索。空间检索可以实现任意空间、任意资源的检索及可视化。

图7　科技资源检索结果——以北京大学为例

6　结语

科技资源数据开放共享是实现科学、社会和经济价值的重要途径，促进科技资源数据的共享和重用，是区域高效协同创新发展有效保障。然而，现有研究成果较少基于区域科技资源角度探讨元数据，且关于公共数据开放共享的研究多从策略研究、政策保障等宏观层面展开，缺乏基础元数据的研究和实践操作。本文以京津冀科技资源开放共享为例，研究制定科技资源元数据标准体系，并以科技机构为载体，建立关系型数据库，实现多源要素的空间关联，形成一套完整的科技资源元数据标准，并利用元数据收割算法实现本地元数据的扩展和优化，为多源异构数据汇交提供保障。同时，通过平台搭建，实现多源异构数据的关联检索和可视化展示，为区域科技资源开放共享提供经验借鉴。后续将针对科技资源数据标准的具体应用情况、技术问题等方面开展进一步的深入研究。

参考文献

[1] 尹婵娟, 朱晓峰, 黄霞. 近10年来我国知识获取研究综述[J]. 情报理论与实践, 2011, 34(9): 123−126.

[2] 中共中央国务院关于构建数据基础制度更好发挥数据要素作用的意见[EB/OL] . [2022−12−27]. https://zycpzs.mofcom.gov.cn/html/gwy/2022/12/1672128008887.html.

[3] 孙坦, 鲜国建, 黄永文, 等. 面向外文科技文献的科技知识组织体系建设与应用[J]. 数字图书馆论坛, 2020(7): 20−29.

[4] 曹琳. 内蒙古大学图书馆资源整合研究[D]. 呼和浩特: 内蒙古大学工商管理学院, 2012.

[5] 顾复, 刘杨圣彦, 顾新建. 科技资源描述模型和建立方法研究[J]. 知识管理论坛, 2020, 5(2): 69−81.

[6] 李梅, 苗润莲, 张岸, 等. 基于GIS的京津冀科技资源数字地图服务平台构建[J] .现代情报, 2017, 37(6): 172−177.

[7] Dublin Core Metadata Initiative. Guidelines for Dublin Core application profiles[EB/OL]. [2023−03−02]. https://www.dublincore.org/specifications/dublin−core/profile−guidelines/.

[8] Huygens. Huygens institute for the history of the Netherlands (Huygens ING) data policy memorandum [EB/OL].[2023−03−02].Https://www.huygens.knaw.nl/wp−content/uploads/2013/10/20140327−data−policy−document.pdf.

[9] Library of Congress Network Development, MARC Standards Office. MARC DTDs document type definitions background and development[EB/OL]. [2023−03−03].https://www.loc.gov/marc/marcdtd/marcdtd−back.html.

[10] 赵英, 陈文飞. 基于Z39.50数据库更新扩展服务的联合目录规程[J].现代图书情报技术, 2000(6): 32−35.

[11] ONIX for Books. Maintenance and support [EB/OL]. [2023−04−27]. https://www.editeur.org/16 /Maintenance−and−support/.

[12] 魏建琳, 马莉萍. 基于XML的MARC元数据互操作的实现[J]. 西安文理学院学报(自然科学版), 2006(2): 99−102.

[13] GETANEH A, BRETT S, PENNY R. Towards a conceptual framework for user- driven semantic metadata interoperability in digital libraries: a social constructivist approach[EB/OL].[2023−03−04]. https://www.emerald.com/insight/content/doi/10.1108/03074801211199031/full /html.

[14] 朱艳华, 高瑜蔚, 胡良霖, 等. 我国科学数据标准规范实践与思考[J]. 中国科学数据, 2023(8): 1−10.

[15] 朱兴国, 武少波, 夏显鄂. 基于元数据的数据共享系统框架设计研究[J]. 科协论坛, 2010(4): 51−52.

[16] 袁烁峰, 林小露. 基于共性元数据规范的科技计划项目数据资源整合[J]. 科技管理, 2012(4): 19−21.

[17] 全国信息与文献标准化技术委员会. 科技报告元数据规范: GB/T 30535—2014 [S]. 北京: 中国标准出版社, 2014.

[18] 刘春燕, 安小米. 基于生命周期的科技信息资源共享元数据研究[J]. 情报理论与探索, 2018, 41(16): 38−43.

[19] 王志强, 杨青海. 科技资源元数据标准化研究[J]. 标准科学, 2021(5): 29−33.

[20] 徐迪威, 张颖. 中国地方科技平台标准化研究[J]. 科技管理研究, 2016, 36(16): 74−78.

[21] MALY K, ZUBAIR M, LIU X M. An OAI data service provider for the individual[J]. D−Lib magazine, 2001, 7(4): 1082−9873.

[22] 王蜀安, 汪萌, 张铭. 支持OAI−PMH的元数据互操作体系结构设计与实现[J] .计算机工程与应用, 2003, 39(20): 168−172.

[23] 科技部. 国家科技计划科技报告管理办法[EB/OL]. [2023−03−05]. http://www.gov.cn/gongbao/content/2014/content_2567187.htm.

[24] 付强, 陈晓玲, 李沫, 等. 基于深度整合的科技资源系统的设计与实现[J]. 吉林大学学报(信息科学版), 2022, 40(6): 924−929.