doi:10.3772/j.issn.2095-915x.2020.06.008

面向科技大数据的元数据仓储建设实践探索

张勇[1, 2] 苏学[2] 谢振峰[2]

1. 中国科学技术信息研究所 北京 100038；
2. 北京万方数据股份有限公司 北京 100038

开放科学（资源服务）标识码（OSID）

摘要：文章从科技领域的元数据建设与应用角度出发，首先在总结分析面向科技大数据的元数据建设研究成果和实践案例的基础上，提出了目前元数据建设存在的标准不统一、规范难度大、缺乏关联等问题。然后详细介绍了科技大数据领域元数据仓储建设的目标、具体流程，并且构建了覆盖 10 亿条科研产出、科研管理的元数据仓储数据库。最后以中文科技期刊论文元数据为例，介绍了元数据仓储建设的具体实现方法，提出一种改进的记录链接方法，并从数据质量和效率两方面验证元数据仓储建设的成果。

关键词：科技大数据；元数据仓储；仓储建设

中图分类号：G35

Research on the Construction of Metadata Warehouse for Big data of Science and Technology

ZHANG Yong[1,2] SU Xue[2] XIE Zhenfeng[2]

1. Institute of Scientific and Technical Information of China, Beijing 100038, China;
2. Wanfang Data Company Ltd., Beijing 100038, China

Abstract: From the perspective of metadata construction and application of scientific and technological big data, this paper summaries and analyzed the research results and practice cases and puts forward the existing

作者简介：张勇（1964-），本科，研究方向为知识组织、数字化加工、元数据建设；苏学（1986-），硕士，研究方向为知识组织、元数据建设；谢振峰（1979-），本科，研究方向为数据挖掘、智能检索、知识组织。

problems of metadata construction, such as the inconsistency of metadata standards, the difficulty of metadata specification, and the lack of correlation between metadata. Taking the metadata of Chinese Sci-tech Journals as an example, this paper introduces in detail the goal, specific process and specific method of metadata storage construction in the field of sci-tech big data, and finally constructs a metadata storage database covering 1 billion scientific research outputs and scientific research management, with a view to providing standard process and data construction standard for the metadata construction of sci-tech big data, and searching documents of knowledge service platform and knowledge discovery to provide data support.

Keywords: Technical data; metadata repository; warehousing construction

引言

科技创新是我国全面推进国家治理体系和治理能力现代化的主要推动力，具有支撑引领作用。加强科技大数据等数据资源库和平台建设，是推动我国科技创新，建设国家大数据中心的重要组成部分。2020 年 4 月，国家发改委首次明确新基建的范围，即：信息基础设施、融合基础设施、创新基础设施，建设内容涉及 5G 基站、人工智能、大数据中心等 7 大领域。科技大数据作为一种独特的大数据类型和新一代基础设施之一，将成为我国数字经济新的经济增长点，在建设成为世界科技创新强国中发挥重要作用。

现在全世界的科技文献正以每分钟 2000 印张的速度增长，科技论文每年发表 1000 万篇以上，科技期刊已超过 8 万种，并以每天 3 种以上的速度增长，图书每年出版 10 万种以上，平均不到 5 分钟就出版一本科技图书 [1,2]。美国互联网数据中心指出，互联网上的数据每年将增长 50%，每两年便将翻一番，目前世界上 90% 以上的数据是最近几年才产生的 [3]。随着科技文献的激增，人们查找文献、获取知识的方式也在不断改变，从图书馆系统的目录索引，到基于互联网的搜索引擎，再到学术搜索、知识发现系统。海量的文献资源与精准获取知识之间的矛盾日益明显，因此，建设面向科技大数据的元数据仓储尤为必要。

1 相关研究

1.1 科技大数据

科技大数据是元数据仓储的内容，关于科技大数据有的学者认为是一种特殊类型的大数据，是指长期积累形成的与科技创新全过程相关的各类非数值型科技信息，它涵盖了客观描述科技创新决策和具体的科技创新活动全过程的各类科技信息，最常见的一种科技大数据即科技文献数据 [4]；也有的学者认为科技大数据不同于传统论文数据，也不同于一般意义上的网络及行业大数据，数据内容包括科技成果数据、科技活动数据以及互联网自媒体科技资讯数据 [5]。

doi:10.3772/j.issn.2095-915x.2020.06.008

1.2 元数据概念内涵

元数据是关于数据的数据，按照功能的不同，元数据可分为描述元数据、结构元数据和管理元数据等。其中，描述元数据主要提供检索和定位的信息，用于发现和标识特定的数据；结构元数据主要记录数据的构成及其相互关系；管理元数据提供数据资源管理所需的信息[6]。为了能够精准的查找文献获取知识，我们必须将多来源异构的元数据映射形成统一的元数据描述框架。

1.3 元数据描述标准

目前已有多种元数据标准存在，国际上比较有影响的主要有 MARC、CDWA、Dublin Core（都柏林核心）、EAD、FADC、GILS、TEI、VRA 等[7]，国内常用的有国家图书馆中文元数据方案、中国科学院科学数据库核心元数据标准、北京大学图书馆中文元数据标准框架、NSTL 统一文献元数据标准 3.0 等标准[8]。

在科技文献方面，MARC 和 DC 被广泛使用，MARC 主要应用在图书馆的书目系统，DC 主要应用在资源发现系统，但是 DC 的使用范围教 MARC 使用更广泛。DC 元数据是一种简单易用的信息资源描述格式，是目前世界上使用最广泛的元数据格式，具有最强的适用性和最大的弹性，便于网络资源的发现和检索。DC 元数据是包含 15 个基本元素和 44 个限定词的元素集[9]，依据其描述的内容类型和范围可分为三组：①对资源内容的描述：题名、主题、说明、来源、语种、关联和覆盖范围；②对知识产权的描述：创建者、出版者、其它责任者和权限；③对外部属性的描述：日期、类型、形式和标识符。

2 元数据仓储建设案例研究

元数据仓储的目的是集成多源异构数据，为文献搜索、知识发现平台提供统一的数据支撑，本文通过对国内外主要学术搜索平台的调研，发现目前的搜索平台主要分为以 Google scholar、百度学术为代表的互联网搜索平台和以 Summon、Primo、Pubmed Central、文津搜索、国家科技图书馆为代表的资源发现系统，本节分别从元数据获取和元数据集成算法等方面阐述元数据仓储建设的研究成果。

2.1 元数据获取

在元数据获取方面，Google scholar、百度学术主要通过资源合作、资源收割以及爬虫抓取等方式获取元数据。Google scholar 的元数据主要来源免费的网络学术资源、学术出版商或学术性的商业数据库、高校或其他科研机构的图书馆[10]。百度学术元数据的主要来源：一是与世界知名内容提供商进行一对一合作，授权获取到最为全面、稳定、优质的题录数据；二是对于部分开放资源，采用如 OAI-PMH 协议等的元数据收割技术进行数据收集；三是对于长尾站点，利用爬虫进行数据的收录、解析、加工处理[11]。而以 Summon、Primo、Pubmed Central、CiNii Research、文津搜索、国家科技图书馆为代表的资源发现系统主要获取方式是资源合作、资源收割、合作共享以及自建等。艾利贝斯的 Summon 平台与 1000 多家原始出版商签约，通过 OPAC 记录每日更新上传至

FTP，OAI-PMH 更新收割等方式获取元数据，涵盖 17 万种学术期刊，收录元数据总量超过 24 亿。Primo 平台也是通过与资源提供商签订协议获取元数据，对不同来源格式规范化处理之后以统一的 Primo 数据格式 PNX 存储，支持对馆藏资源的查重与 FRBR 处理，索引记录超过 5 亿。美国国立卫生研究院的 PubMed Central（PMC）通过元数据获取协议 OAI-PMH、FTP 等方式获取元数据，用日志归档和交换标签套件与出版商进行通用格式的日志内容交换。日本 CiNii Research 收集与日本研究项目相关的元数据并从聚合元数据中提取研究实体及其关系，构建大规模的学术知识 [12]。国家图书馆的文津搜索通过自己建设、外部购买等多种方式获取元数据，清理解析后将元数据字段都映射到“文津搜索”定义的统一 XML 格式上，形成覆盖 2 亿的元数据仓储库。国家科技图书馆文献中心通过自主加工、谈判引进等模式建设了 2.5 亿条的文献元数据 [13]。

2.2 元数据集成算法

在元数据集成方面，互联网学术搜索以网页去重为主，其中 Google 利用 simhash 算法处理海量文本去重 [14]，其主要思想是降维，将高维的特征向量映射成低维的特征向量，再通过两个向量的 Hamming Distance 来确定文章是否重复或者高度近似 [15]。而资源发现系统更关注各类资源的元数据集成，文津搜索系统根据每个类型资源的特点，分别制定了不同的数据合并规则 [16]，详见表 1。赵捷等在国家科技图书馆元数据集成管理系统中也根据文献资源类型制定了不通的数据查重规则 [17]，详见表 2。董微以正题名、作者名称、作者单位、语种、总页数、起始页码、终止页码等字段进行期刊论文的元数据集成实验 [18]。

表 1 文津搜索数据合并规则

数据类型	合并规则
自建资源	系统号+001 号
期刊论文	题名+作者+年份+期刊名称+期号
会议论文	题名+作者+年份+会议名称
学位论文	题名+作者+年份
外文图书	有ISBN 使用题名+作者+年份+ISBN； 无ISBN 使用题名+作者+年份+出版社
中文图书	有ISBN 使用题名+作者+年份+ISBN； 无ISBN 使用题名+作者+年份+出版社

表 2 查重规则库中的主要规则

文献类型	层 级	对应规则
期刊	母体	由ISSN、刊名、馆藏号等构成
	卷期	由上一层级规则与母体（卷期）中的年、卷、期等共同构成
	篇级	由上一层级规则与篇级唯一标识、题名、作者、单位等共同构成
会议	母体	由会议名称、主办单位、届次等构成
	卷期	由上一层级规则与母体（文集）题名、馆藏号等共同构成
	篇级	由上一层级规则与篇级唯一标识、题名、作者、单位等共同构成
文集汇编	母体	由母体（文集）题名、馆藏号等构成
	篇级	由上一层级规则与篇级标识、题名、作者、单位等共同构成
图书专著	母体	由题名、作者、出版者、版次等构成
研究报告	篇级	由题名、作者、单位等构成
学位论文	篇级	由题名、作者、学位、授予单位等构成

doi:10.3772/j.issn.2095-915x.2020.06.008

2.3 存在的问题

在元数据仓储建设的调研和实践中，我们发现元数据仓储建设存在元数据标准不统一、元数据规范难度大、缺乏多维度数据的关联分析等问题。

（1）元数据标准不统一

目前元数据的数据标准各自为阵，例如 Web of Science 数据库、Scopus 数据库、PMC、NSTL、IOP、wiley、scirp 采用的都是自定义元数据标准，OUP、CUP、emerald、De Gruyter、Hindawi、taylor、wsn 等 7 家出版社采用了 JATS 数据标准 [19]。而且同一对象在不同标准中的元数据项表述不一 [20]，例如篇名在 Wiley 中元数据字段被表述为〈article-title〉篇名，而在 Thomson Reuters 中定义为〈title type=“item”〉篇名。

（2）元数据的规范难度大

由于原始数据的书写不规范、元数据加工过程中的数据错误、遗漏等以及元数据数量庞大等导致元数据规范难度大。期刊编辑部虽然对来源稿件的格式做了规范，但是在数据内容方面却无法做出明确的规范，作者的书写习惯往往会产生不同的元数据。此外，纸质资源的数字化加工规范的不同，也可能导致元数据的内容不一致，另外识别软件识别率、工作人员录入错误等原因导致的元数据错误。

（3）多维度数据的关联分析

大量元数据分布在不同地方，例如期刊论文数据主要来自期刊编辑部，会议论文数据主要来自举办会议单位的会议论文集，学位论文主要来自不同的高校、研究所等，专利数据主要来自各国的知识产权单位等，元数据之间缺乏关联分析，无法与他数据集实现数据共享和相互关联 [21]。

3 面向科技大数据的元数据仓储建设

3.1 建设目标

面向科技大数据的知识发现服务平台，可以帮助科研人员寻找文献、解决问题、发现知识，为科研人员提供数据支撑，推动科技创新能力的转变，提升科研效率。元数据仓储的质量和覆盖范围决定了知识发现服务平台价值大小，因此构建面向科技大数据的元数据仓储是我们的首要目标。

3.2 建设流程

元数据仓储建设的流程是元数据仓储建设的重要依据，指导整个元数据仓储建设工作。在实践探索过程中，笔者主要围绕确定元数据范围、制定元数据标准、元数据获取、元数据集成、元数据要素规范及关联等步骤确定元数据仓储建设的流程。

元数据范围主要依据元数据仓储的建设目标和科技大数据的定义确定。根据科技资源的载体不同，元数据可以大致划分为印刷型数据、电子型数据和网络型数据三类，印刷型数据主要是以纸质出版物的形式存在，电子型数据主要是以数字信号存储在计算机、硬盘等存储介质，网络型数据主要是发布在互联网上公开传播的数据，是一种特殊的电子型数据。而根据资源类型不同，元数据又划分为期刊论文、会议论文、图书、报纸、科技报告、标准、专利、学位论文、预印本资源、政策法规、科技项目、科技人才等元数据。

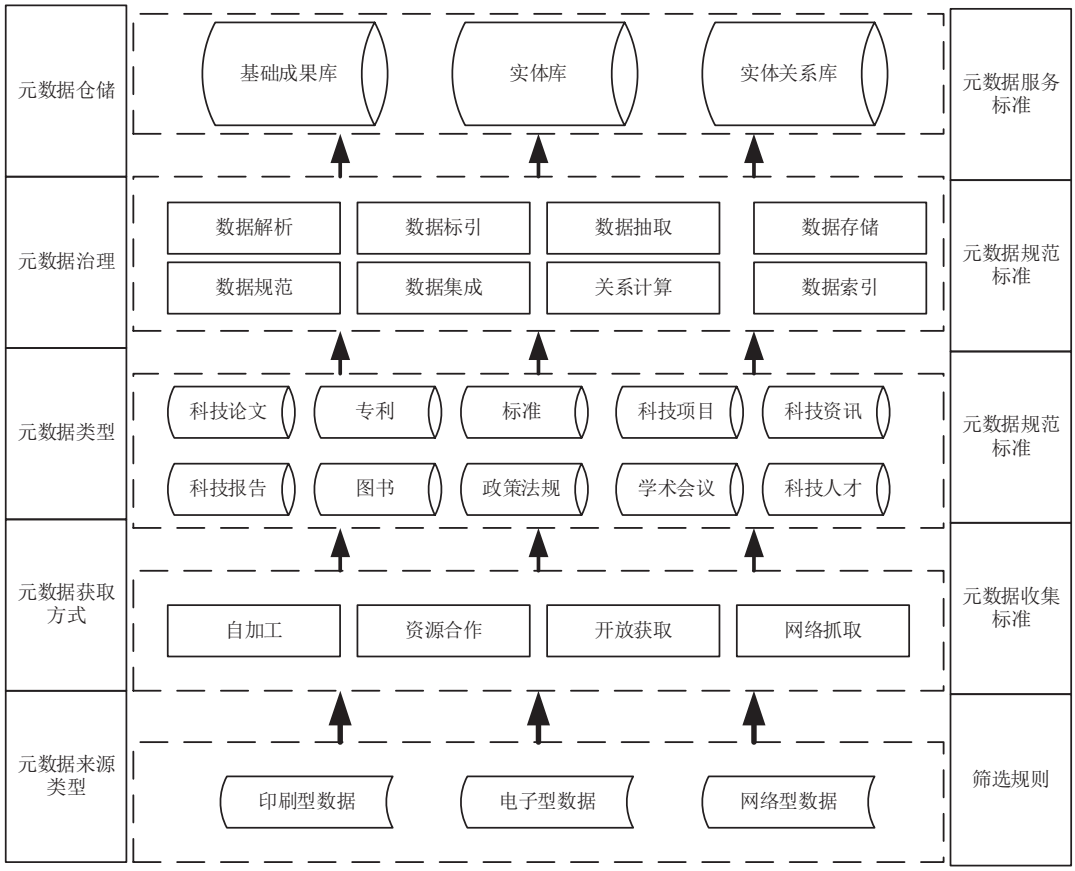

图 1　面向科技大数据的元数据仓储建设流程

元数据标准规定了各类型元数据资源的最小单元的收集规则和规范规则，元数据标准的优劣决定了元数据仓储的数据质量，也影响着知识服务平台的功能。

元数据获取是元数据仓储建设的基础，制约元数据仓储的建设规模。针对印刷型数据、电子型数据和网络型数据等不同载体的元数据，我们分别采取数字化加工、资源合作、开放获取等不同的获取方式。

元数据的治理和元数据的集成是元数据仓储建设的关键，影响元数据仓储建设的质量。元数据的治理主要分为数据的解析、转换、清洗等过程。利用不同获取方式获取到的元数据存储格式、结构各不相同，因此我们首先要通过数据解析，把分布在不同文件的元数据记录逐条逐字段的存储到数据库中，然后利用数据转换技术将解析出来的不同结构的元数据规范化。其中，数据转换主要包括三个步骤：一是将不同结构的元数据资源映射到按照资源类型划分的、结构相对统一的元数据标准；二是确定元数据间的关联关系；三是按照映射关系和关联关系进行元数据的转换。最后利用数据清洗技术规范元数据。数据清洗主要利用规则处理、标准映射、相似度计算等技术解决数据格式、信息冗余的去重以及错误、不完整信息的修正剔除问题。

doi:10.3772/j.issn.2095-915x.2020.06.008

元数据的集成是把多来源异构数据集成融合在一起，解决不同来源的数据去重问题，保证同一资源的多条重复元数据记录能够聚合归并为一条记录，实现元数据记录唯一化，来源多源化。

元数据要素规范和元数据关联是元数据仓储建设的核心，决定了元数据仓储建设的增值服务的价值。元数据要素规范主要是针对人、机构、主题、学科、基金等五要素内容进行合并归一，把表述不规范的同一实体合并为一个实体。元数据关联主要是各类型元数据资源之间引用、被引用的引用关系关联和元数据五要素与各类型元数据资源的关联关系。

以上几个流程环环相扣，原始的数据格式、数据质量直接影响了数据集成规范的效果，集成的仓储数据也会对资源的获取方式提出要求。因此在进行元数据仓储构建时，要从整体上进行规划，通盘考虑所有的环节。

4 实证研究

笔者以科技期刊论文为例，依据元数据的建设流程主要从元数据标准制定、元数据获取、元数据集成和元数据要素规范及关联等方面详细讲述万方数据元数据仓储建设的具体流程和实现方法。

4.1 科技期刊论文元数据标准

早在 2003 年万方数据就开始研究元数据标准，并针对文献类资源制定了相应的元数据标准，2016 年又在之前标准的基础上制定了万方文献信息规范集，规定了万方数据文献信息模型包括通用性、种册型和单篇型。其中，通用性信息为各类型的公共信息；种册型信息包括期刊信息、图书信息、会议文集信息、标准信息等；单篇型包括期刊论文信息、图书篇章信息、专利信息、学位论文信息等，详见图 2。

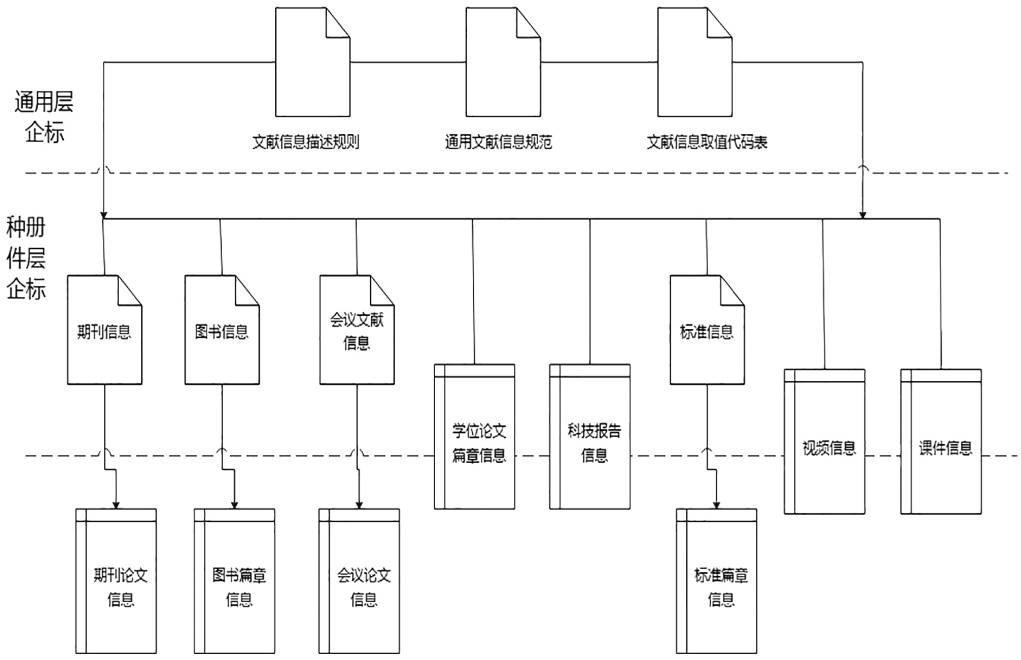

图 2 万方数据文献信息类型模型

万方数据针对期刊论文又详细制定了《万方文献信息规范集：期刊论文信息规范》，分别定义了期刊信息结构和论文信息机构[22]。期刊信息用于描述期刊整刊信息的一组数据，包括基本信息单元（标识数据组、题名数据组）、卷期信息单元（标识数据组、题名数据组、出版数据组、计数数据组）、贡献者信息单元（责任机构数据组、责任人数据组）、主题信息单元、获取信息单元、操作信息单元、评价信息单元、投稿信息单元和沿革信息单元共计 9 个信息单元构成。期刊论文信息用于描述期刊中单篇论文信息的一组数据，包括基本信息单元、卷期信息单元、单篇信息单元（标识数据组、题名数据组、出版数据组、文摘数据组、计数数据组）、引文信息单元、附加信息单元、会议信息单元、贡献者信息单元、基金信息单元、主题信息单元、获取信息单元、操作信息单元、评价信息单元和投稿信息单元组成。

4.2 科技期刊论文元数据的获取

科技期刊论文的元数据获取主要有三种方式：数字化加工数据、资源合作数据、开放获取。万方数据经过多年的积累，逐步形成了一支集资源收集、数字化加工于一体的专业队伍，建设形成了国内领先的现代化数据加工基地，拥有全套规范化加工生产线、高清晰扫描、以及拥有自主知识产权的人工智能标引系统和自动化的采集工具软件等。收集到的期刊数据先与期刊编辑部签订版权协议后，然后利用自有的先进的数字化加工技术，实现科技期刊论文的结构化存储。万方数据还通过资源合作的方式获取元数据。目前万方数据已经与 30 多家数据库厂商签订元数据合作协议，进行元数据层面的资源合作互换。例如国家哲学社会科学学术期刊、中国科技论文在线、NSTL 外文文献数据库、Wiley 期刊数据库、医学文献检索服务系统（PubMed）数据库、日本科学技术信息集成系统（J-STAGE）等中外文电子全文数据库。同时，万方数据还与 CrossRef 建立了元数据合作协议，覆盖已经注册 DOI 的所有外文论文元数据。

此外，万方数据利用开放获取的方式获取互联网的公开资源。随着互联网和开放存取运动的发展，OA 期刊和 OA 仓储越来越受到人们的青睐，成为科技大数据的一种重要来源。目前全世界已有 5225 个人和 534 个相关研究机构签署了信息自由传播会议（Budapest Open Access Initiative，简称 BOAI）计划协议[23]。由于开放存取资源一般是支持 OAI-PMH 协议，因此我们开发了元数据接口收割工具，通过 OAI 命令时间戳收割开放存取资源。

4.3 科技期刊论文元数据集成方法

4.3.1 记录链接方法

记录链接是指在两个或两个以上的文档中找出代表相同实体记录的方法[24]，本文中笔者把每条元数据记录做为一个实体，元数据的查重集成问题也就演化成从多来源数据中查找相同实体的问题。记录链接方法的实现主要有精确匹配和概率匹配两种方法，从而记录链接可以分为确定性记录链接和概率记录链接两种。确定性记录链接通过某个唯一标识符字段或者几个质量较高的字段进行记录匹配，概率链接

doi:10.3772/j.issn.2095-915x.2020.06.008

记录则通过比较记录间相同的概率确定记录匹配程度，该过程需要对两个匹配的记录中的某个识别符的值相同的概率进行参数估计。由于论文是发表在特定母体文献上的数据，母体文献不同则代表论文是不同的，同时由于论文元数据的字段通常是短文本，例如作者、年、期等，基于期刊的特性，本文主要选取了记录链接的精确匹配算法，但是在长文本匹配上采用相似度匹配代替精确匹配。

4.3.2 记录链接步骤

记录链接的方法分为数据标准化、分块、记录链接等步骤。

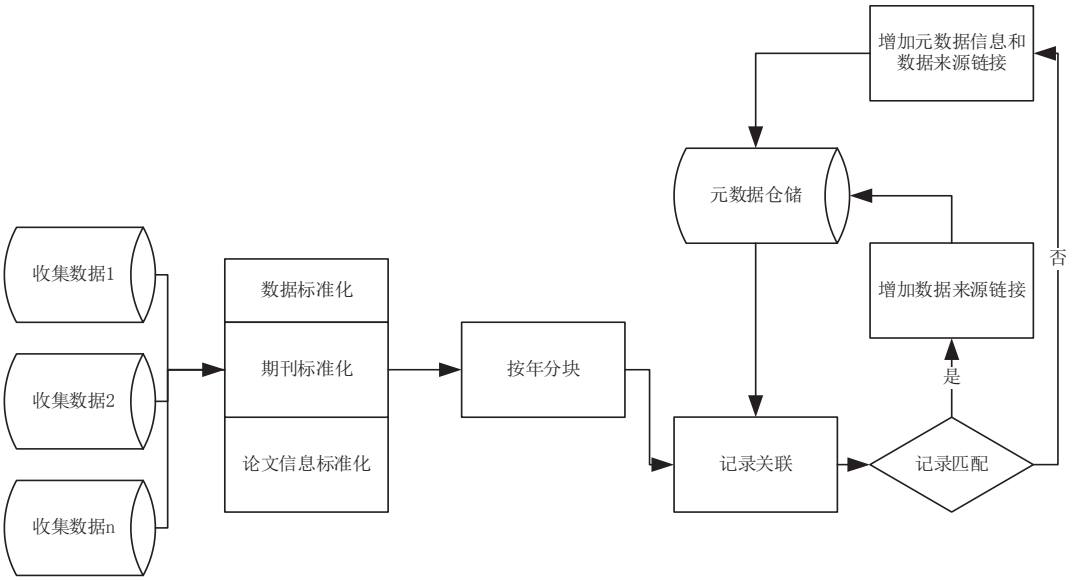

图 3 元数据集成

（1）数据标准化

记录链接对连接数据集的数据质量敏感度很高，因此，我们所研究的数据集在进行记录链接之前都需要进行严格的标准化处理，数据标准化主要是把不同来源的数据格式化，把缩写、简写、曾用名等名称进行规范标准化，将同一指代的词汇转换成同一格式或者同一词汇以及数据编码等。

来源	标题	作者	ISSN	刊名	年	卷	期	页码
1	2017 年武汉市武昌区学龄期儿童生长发育现况分析	何兵%陆文峰%徐丽娟%张如洪%张明霞%万俊华	1671-8852	武汉大学学报(医学版)	2020		01	141-144
2	2017 年武汉市武昌区学龄期儿童生长发育现况分析	何兵%陆文峰%徐丽娟%张如洪%张明霞%万俊华	1671-8852	武汉大学学报（医学版）	2020	41	1	141——144
1	Angiogenesis and neuronal remodeling after ischemic stroke	Masahiro Hatakeyama%Itaru Ninomiya%Masato Kanazawa	1673-5374	Neural Regeneration Research	2020		01	16-19
2	Angiogenesis and neuronal remodeling after ischemic stroke		1673-5374	中国神经再生研究（英文版）	2020		1	16——19
1	具有贵金属行为的等离子 Bi 纳米颗粒负载到还原 TiO₂微米球及其高效全光谱光催化产氧(英文)	赵行%梁婵倩%刘香%邱鹏源%崔洪芝%田健	0253-9837	Chinese Journal of Catalysis	2020		02	333-344
2	具有贵金属行为的等离子 Bi 纳米颗粒负载到还原 TiO2 微米球及其高效全光谱光催化产氧	赵行%梁婵倩%刘香%邱鹏源%崔洪芝%田健	0253-9837	催化学报	2020	41	2	353——364

图 4 元数据在不同来源的格式

在科技期刊论文元数据标准化过程中，一是对字符格式进行标准化，剔除无意义的字符；二是对期刊进行编码。利用期刊的 ISSN 号、CN 号等唯一标识符，结合期刊的名称、出版周期，对期刊进行唯一编码。

（2）分块

分块是减少记录链接过程中需要比较记录对数量的一种方案，尤其是在计算大量数据时，合理的分块可以减少记录匹配数量，大幅提高工作效率。分块时一般会选取记录的某个属性或者多个属性作为分块变量，分块变量间的记录不重叠。鉴于国内外发表的期刊论文数量已经到达数亿条，期刊论文发表的年份在论文中是单一且不重叠的，而且字符规范，因此笔者使用期刊论文的发表年份作为分块变量，在进行记录链接时只匹配关联相同年份的数据。

（3）记录链接

记录链接首先是建立不同的记录链接策略，将记录链接策略以规则的方式存入规则库，然后基于规则进行记录的匹配。笔者根据科技论文数的特性并通过调研和反复实验，制定了表 3 中不同的链接规则。

表 3　科技论文元数据链接规则

序号	链接规则
1	题名、期刊、年、期
2	题名、期刊、年、作者
3	期刊、年、期、作者、页码
4	题名、年、作者、页码

长文本的相似度计算采用余弦相似度算法。为了提高效率，向量模型本文选择字向量，现在我们假设：论文 1 中出现的字为：Z1*c*1,Z1*c*2,Z1*c*3,Z1*c*4……Z1*cn*；它们在章节中的个数为：Z1*n*1,Z1*n*2,Z1*n*3……Z1*nm*；论文 2 中出现的字为：Z2*c*1,Z2*c*2,Z2*c*3,Z2*c*4……Z2*cn*；它们在论文中的个数为：Z2*n*1,Z2*n*2,Z2*n*3……Z2*nm*；其中，Z1*c*1 和 Z2*c*1 表示两个文本中同一个字，Z1*n*1 和 Z2*n*1 是它们分别对应的个数，最后我们的相似度可以这么计算：

$$\frac{(Z1n1*Z2n1)+(Z1n2*Z2n2)+\cdots+(Z1nm*Z2nm)}{\sqrt{Z1n1^2+Z1n2^2+\cdots+Z1nm^2}*\sqrt{Z2n1^2+Z2n2^2+\cdots+Z2nm^2}}$$

我们随机抽取不同来源的期刊论文数据，利用期刊、年、期、作者、页码等短文本就行数据的查重，然后人工比对数据匹配情况，筛选出 19458 条匹配数据，发现题名相似度大于 90.1% 的数据是属于同一论文，我们把题名相似度大约 90% 做完记录关联匹配的条件之一。

短文本字段我们采用精确匹配的算法，结合长文本字段限定在特定的相似度下，新增更新数据要与元数据仓储已有的数据进行匹配，只要符合表 3 中的链接规则中的一种就认为两条记录是匹配记录，匹配的记录进行数据来源的标识关联，增加元数据仓储记录的来源标识；不匹配的记录则认为是新增数据，该条记录增加到元数据仓储中。元数据信息根据元数据的质量厚薄程度优先选择自己数字化加工的数据，同时补充其他来源的数据，重点补充数据的来源字段以便查找该数据的来源。

4.4　科技期刊论文元数据的关联

科技期刊论文的元数据关联是指科技期刊论文与自身以及其他科研实体间的关联关系。通过参考文献可以建立科技期刊论文与科技期刊论文、学位论文、会议论文、专利等多种科

doi:10.3772/j.issn.2095-915x.2020.06.008

研产出数据的关联，通过基金信息可以建立科技期刊论文与科研项目的关联，通过机构信息可以建立科技论文与科研机构的关联，通过作者信息可以建立科技论文与科研人员的关联，而通过这种二维的关联关系逐层拓展就可以建立一个科技信息的关联网络，从而发现一些隐形知识。本文将简单阐述科技论文与科研机构关联方面的一些探索。

作者单位信息是科技论文数据的一个必备字段，也是建立科研机构与科技论文元数据关联的纽带。理想情况下，科技论文数据中的作者单位信息与实际的科研机构是相同的，需要查找某机构的科技论文数据是直接通过作者单位检索某机构就可以。但是，实际上由于机构的变革和个人书写习惯等导致机构名称具有多样性，机构合并前的名称、机构的曾用名、机构的简称、不规范的机构名称等等。因此科研机构与科技论文的关联的关键就是科技论文中作者单位信息的规范。我们采用叙词表编制模型中的“用”、“代”、“属”、“分”、“族”的结构体系依据万方数据自身建设的机构库构建了机构多层级词表[25,26]。

表 4 机构多层级词表

机构名称——	北京大学第一医院	
代项——	D 北京大学第一临床医学院	——曾用名、简称、不规范名称
	北大医院	
属项——	S 北京大学医学部	——上级机构名称
族项——	Z 北京大学	——一级机构名称
分项——	F 北京大学第一医院癫痫诊疗中心	——下属机构名称
	北京大学儿童视力保护中心	
	……	
机构名称——	中国药物依赖性研究所	
用项——	Y 北京大学中国药物性依赖性研究所	——规范机构名称
属项——	S 北京大学医学部	——上级机构名称
族项——	Z 北京大学	——一级机构名称

由于科技论文元数据中的作者单位信息多样，在深入研究机构命名规则的集成上，根据不同的机构类型构建机构名称特征词及属性字典，根据已构建的机构特征词字典，遵循正向识别的原则，查找机构数据中的名称特征词，依次识别作者单位的一级机构、二级机构、三级机构等。

OneOrg	CX_Org_...	CX_Org_EJ	CX_Org_SJ	CX_Org_ZJ
北京大学物理学院大气与海洋科学系	北京大学	北京大学物理学院	北京大学物理学院大气与海洋科学系	北京大学物理学院大气与海洋科学系
北京大学物理学院大气与海洋科学系,北京 100871, 白城兵器试验中...	北京大学	北京大学物理学院	北京大学物理学院大气与海洋科学系	北京大学物理学院大气与海洋科学系
北京大学物理学院大气与海洋科学系,北京100871,民航贵州空管分局...	北京大学	北京大学物理学院	北京大学物理学院大气与海洋科学系	北京大学物理学院大气与海洋科学系
北京大学物理学院大气与海洋科学系气候与海-气实验室	北京大学	北京大学物理学院	北京大学物理学院大气与海洋科学系	北京大学物理学院大气与海洋科学系气候与海-气实验室
北京大学信息科学技术学院电子学系	北京大学	北京大学信息科学技术...	北京大学信息科学技术学院电子学系	北京大学信息科学技术学院电子学系
北京大学遥感与地理信息系统研究所;北京市空间信息集成与3S工程...	北京大学	北京大学遥感与地理信...	北京大学遥感与地理信息系统研究所	北京大学遥感与地理信息系统研究所
北京大学药学院化学生物学系	北京大学	北京大学药学院	北京大学药学院化学生物学系	北京大学药学院化学生物学系
北京大学药学院化学生物学系,北京	北京大学	北京大学药学院	北京大学药学院化学生物学系	北京大学药学院化学生物学系
北京大学医学部公共卫生学院妇女与儿童青少年卫生学系	北京大学	北京大学医学部公共卫...	北京大学医学部公共卫生学院妇女与儿童青少年卫生...	北京大学医学部公共卫生学院妇女与儿童青少年卫生学系
北京大学医学部公共卫生学院劳动卫生与环境卫生学系	北京大学	北京大学医学部公共卫...	北京大学医学部公共卫生学院劳动卫生与环境卫生学系	北京大学医学部公共卫生学院劳动卫生与环境卫生学系
北京大学医学部公共卫生学院社会医学与健康教育系,北京,100191	北京大学	北京大学医学部公共卫...	北京大学医学部公共卫生学院社会医学与健康教育系	北京大学医学部公共卫生学院社会医学与健康教育系
北京大学医学部基础医学院病原生物学系,北京,100191	北京大学	北京大学医学部基础医...	北京大学医学部基础医学院病原生物学系	北京大学医学部基础医学院病原生物学系
北京大学医学部基础医学院病原生物学系和感染病中心	北京大学	北京大学医学部基础医...	北京大学医学部基础医学院病原生物学系	北京大学医学部基础医学院病原生物学系和感染病中心

图 5 计算机自动识别机构层级关系

经过上述两步建立了多层级机构词表和论文中作者单位数据的层级识别，我们根据最小单位优先的原则，利用计算机自动识别的机构名称按照层级由深到浅的次序分别于多层级机构词表由深到浅的次序进行数据关联，当数据关联上之后，不再参与后续的数据关联，没有关联的数据则继续后续的数据关联。数据关联完成后，对于没有给出规范机构名称的数据还需要辅以人工的方式给出机构的规范名称。

4.5 结果检验

为验证元数据仓储建设的合理性，本文以中文期刊论文为例，从数据质量和效率两方面实证元数据仓储建设的成果。

4.5.1 数据质量

数据齐全性和重复率是验证数据质量的两个重要方面，因此笔者随机选择关键词“大气污染”、“人工智能”分别在万方学术搜索和百度学术搜索平台限定在题名里检索，检索条数如下：

表 5 发现系统检索结果

	大气污染			人工智能		
	2017	2018	2019	2017	2018	2019
万方学术搜索	878	1032	769	3992	8041	6470
百度学术搜索	409	270	122	1875	1316	816

从检索结果上看，万方学术搜索在“大气污染”、“人工智能”2017 年、2018 年、2019 年三年的数量都是高于百度学术搜索。

随机选取 2019 年度标题里含有“人工智能”的元数据 6470 条，利用相似度算法筛选出题名相似度达到 90% 以上的元数据 4071 条，经过人工核实校验，逐条比对论文的作者、发文期刊，发现重复数据 3 条，重复率 0.046%，属于可接受范围。

4.5.2 效率

数据更新速度是效率的一个重要体现，在单机环境下，期刊论文库数量达到 9000 万条记录时，采用短文本的精确匹配和长文本的相似度计算结合的记录链接算法，每次更新 20 万条数据大约需要 1 个小时左右，完成全部的数据查重和数据规范工作，之前未分块的精确匹配算法每次更新需要 20 小时，与之相比效率有了大幅提高，数据的重复率也大大降低。

经过实证，元数据仓储覆盖的数据相对齐全，数据重复率低，基本实现了元数据仓储覆盖资源广泛且资源唯一的目标。新增数据更新时间较短，也保证了数据更新的及时性。

5 总结建议

本文以万方数据的科技资源元数据建设实践为例，介绍了面向科技大数据的元数据仓储建设的目标、具体流程、建设方法和解决措施，对相关元数据仓储建设具有参考价值。但是由于元数据资源类型的复杂性和时间的限制，整个元数据仓储建设的过程还有很多需要进一步完善的地方，尤其是在元数据集成算法、各元数据要素间的规范关联方面，还需要进一步深入研究。

在元数据仓储建设方面，笔者建议从数据源头抓起，从国家、行业层面规划，在生产数据阶段就为每一篇论文、每个作者、每个机构

doi:10.3772/j.issn.2095-915x.2020.06.008

赋予一个唯一标识，方便后续数据集成融合；在政策方面，也需要国家层面规划，形成一套由上而下的数据共建共享机制，建立公共的基础仓储数据，不同领域可以共享基础仓储，也可以在基础仓储数据上建立各自特色的仓储数据。

参考文献

[1] 浅谈影响信息时代文献信息数量增长的相关原因 [EB/OL]. (2019-11-14) [2020-05-01]. http://blog.sina.com.cn/s/blog_4e56b8ca01000akn.html.

[2] 李燕 . 科技文献对园艺科技发展的作用研究 [D]. 泰安 : 山东农业大学 , 2005.

[3] 带您了解大数据 [EB/OL]. (2019-11-14) [2020-05-01]. http://www.thebigdata.cn/YeJieDongTai/8470.html

[4] 曾文 , 车尧 , 张运良 , 等 . 服务于科技大数据情报分析的方法及工具研究 [J]. 情报科学 , 2019(4): 92-96.

[5] 钱力 , 谢靖 , 常志军 , 等 . 基于科技大数据的智能知识服务体系研究设计 [J]. 数据分析与知识发现 , 2019(1):4-14.

[6] 李善青 , 郑彦宁 , 赵辉 , 等 . 大数据背景下科学元数据的重要问题研究 [J]. 科技管理研究 , 2018(10): 184-188.

[7] 冯项云 , 肖珑 . 国外常用元数据标准比较研究 [J]. 大学图书馆学报 , 2001(4): 15-21.

[8] 陈彩红 . 国内外元数据标准宏观比较研究 [J]. 河北科技图苑 , 2011(1): 66-67.

[9] 张崇 . DC 元数据在国内的应用及思考 [J]. 现代图书情报技术 , 2004(11): 6-9.

[10] 马彩峰 . Google 学术搜索的信息组织探究 [J]. 情报杂志 , 2010(3): 173-175.

[11] 王卓 . 基于信息觅食理论的高校图书馆用户信息获取行为研究 [D]. 哈尔滨 : 黑龙江大学 , 2016.

[12] 山地一祯 , 李颖 . 日本国立信息研究所研究数据基础设施概述 [J]. 情报工程 , 2020, 6(1): 4-10.

[13] 马袁燕 . 面关系发现服务的元数据集成整合研究 [D]. 北京 : 中国科学技术信息研究所 , 2018.

[14] Google 用来处理海量文本去重的 simhash 算法原理及实现 [EB/OL]. (2016-11-27) [2020-05-01]. https://www.sohu.com/a/120031197_468732

[15] simhash 算法及原理简介 [EB/OL]. (2018-04-02) [2020-05-01]. https://blog.csdn.net/lengye7/article/details/79789206

[16] 蔡颖 , 才小川 . 资源发现系统服务能力提升初探——以文津搜索系统为例 [J]. 图书情报导刊 , 2020, 5(3): 31-36.

[17] 赵捷 , 董微 . 面向发现服务的图书馆元数据集成管理系统构建研究 [J]. 数字图书馆论坛 , 2018(7): 11-21.

[18] 董微 , 杨代庆 . 面向学术资源集成的真值发现算法 [J]. 情报工程 , 2017, 3(1): 66-71.

[19] 马袁燕 . 面向发现服务的文献元数据集成整合研究 [J]. 图书馆 , 2019(1): 76-81,87.

[20] 于倩倩 , 张建勇 , 黄永文 . 大数据环境下的文献元数据标准设计特点分析 [J]. 图书馆杂志 , 2018(11): 36.

[21] 范丽鹏 , 王曰芬 , 李璞 . 大数据与创新双驱动的知识创新服务需求与趋势研究 [J]. 情报工程 , 2019, 5(1): 22-32.

[22] 北京万方数据股份有限公 . Q/WF1.4-2016, 万方文献信息规范机 [S]. 北京：北京万方数据股份有限公司 .

[23] 开放存取 [EB/OL]. (2019-11-14) [2020-05-01]. https://baike.baidu.com/item/%E5%BC%80%E6%94%BE%E5%AD%98%E5%8F%96/1688963?fromtitle=%E5%BC%80%E6%94%BE%E8%8E%B7%E5%8F%96&fromid=733660&fr=aladdin

[24] 徐刚 . 记录链接理论及应用 [J]. 统计与决策 , 2016(7): 78-81.

[25] 杨奕虹等 . 机构多层级词表的编制及在文献计量评价与科研绩效管理中的应用 [J]. 数字图书馆论坛 , 2013(6): 57-63.

[26] 陈田田 , 吴广印 . 研究者唯一识别及其在专家档案系统中的实施 [J]. 情报工程 , 2015, 1(3): 31-37