NSTL元数据一体化管理研究*

丁道劲 王星 李芳菊
（中国科学技术信息研究所，北京 100038）

摘要：在数字出版潮流以及用户需求变革的双重驱动下，国家科技图书文献中心（National Science and Technology Library，NSTL）无论是源头的资源建设模式还是终端用户的服务需求都发生较大变化，亟待面向新的业务发展目标构建一体化的元数据管理模式。在NSTL总体业务流程重组再造中，NSTL元数据管理的主要目标在于对资源品种、卷期以及文摘元数据进行规范集成，并形成名称规范库，同时支持服务过程中的资源调度计算。基于上述目标，本文对NSTL元数据一体化管理核心流程以及多源异构元数据质量控制策略进行分析，并以NSTL元数据管理系统为例，论述元数据一体化管理的具体实现过程。

关键词：元数据管理；数据集成；图书馆自动化系统；NSTL
中图分类号：G254 DOI：10.3772/j.issn.1673-2286.2021.07.003
引文格式：丁道劲，王星，李芳菊. NSTL元数据一体化管理研究[J]. 数字图书馆论坛，2021 (7)：18-26.

国家科技图书文献中心（National Science and Technology Library，NSTL）经过二十多年的发展，基于统一的资源采集规划、分布式数据加工、集中的网络系统以及协同服务原则，已经建立起相对稳定的业务流程，形成以资源建设、数据加工、网络系统、文献服务为主要单元的组织结构，协调各成员单位共同开展国家科技文献保障工作。但是，面对当前不断发展变化的科技信息环境，特别是文献资源数字化以及资源开放性不断增强，NSTL亟待对原有基于印本文献的采集加工和服务的业务布局和流程进行优化，从文献服务向知识服务方向转变，以适应数字业务环境变化和用户对知识服务的需求[1]。文献元数据是关于文献资源在品种（如期刊品种、会议）、实体（如期刊卷期、会议论文集）以及论文层级的描述性信息，在以印本为主的资源建设时期，文献元数据主要是指通过编目形成的书目数据。随着文献资源出版方式多样化，元数据的采集获取方式也呈现出多源化趋势，因此在NSTL业务流程再造过程中，需要通过元数据一体化管理，实现对多来源、多载体、多类型文献的统一管理，为海量文献资源的深化利用奠定数据基础。

1 NSTL资源建设模式变革对元数据管理的要求

在数字出版潮流以及用户需求变革的双重驱动下，NSTL无论是源头的资源建设模式还是终端用户的服务需求都发生了较大变化，依托传统编目、实物登到对元数据进行管理的弊端逐渐显现，因此亟待面向新的业务发展目标构建一体化的元数据管理模式。

1.1 采集对象拓展驱动元数据对象多元化

为践行国家科技文献保障使命，NSTL在建设之初即已形成以外文印本资源为主体的资源保障模式，资源类型涉及期刊、会议录、科技报告、科技丛书等多种类型。近年来，每年外文印本期刊订购量仍然维持在1.7万种。但是，随着数字出版趋势越发明显，完全依托印本资源引进已难以满足科研人员的信息需求，因此NSTL逐渐加大了电子资源引进力度，通过全国陆续开通电子资源、开放资源建设等方式，拓宽文献资源采集获取渠道，由此形成立体化资源保障体系。

* 本研究得到国家社会科学基金青年项目“国家元数据库协同构建机制研究”（编号：19CTQ009）资助。

在以印本为主体的资源建设时期，编目业务在资源管理揭示方面发挥了不可替代的作用，它是NSTL书目元数据的主要甚至是唯一来源。但是随着科技文献资源来源渠道多样化，书目元数据不再局限于通过编目产生，视频、课件等非传统科技信息资源也难以完全用MARC进行描述，因此原本线性化的书目元数据管理方式亟待转变。

1.2 揭示途径扩充驱动元数据处理流程化

同样是基于印本文献资源，出于数据质量优化和版权管理的考虑，NSTL长期通过“自主编目+加工”方式进行资源的组织揭示。但是在立体化资源模式牵引下，资源揭示方式也正在向多来源采集方向转变。从2014年起，NSTL开始陆续与科睿唯安、爱思唯尔、施普林格·自然等国外知名数据库商、集成商以及出版社合作，直接获取XML格式的论文元数据。目前，NSTL元数据合作渠道达到20余家，累计采集超过1亿条论文元数据。此外，为了突破商业资源发现服务中存在的馆藏壁垒，NSTL在2020年启动“国家外文科技期刊联合目录”建设，目前已与上海图书馆、中国科学院文献情报中心、CALIS管理中心以及中国医学科学院医学信息研究所/图书馆等达成合作。

无论是从其他机构通过免费或少量付费方式获取的论文元数据，还是通过合作共享而来的馆藏数据（包括书目数据和登到数据），都需要通过依次从品种、卷期（册）、论文层层挂接对应，形成相互关联统一的资源整体。从外部引进元数据，同时囊括了书目、卷期以及论文层级的描述信息，元数据管理成为多来源元数据采集的管理入口，与以往数据管理业务相比，需要增加对论文元数据的清洗归并等流程，为元数据后期的融合计算奠定规范的数据基础。

1.3 服务目标提升驱动元数据管理标准化

无论是立体化资源保障，还是多来源论文元数据以及馆藏元数据的合作引进，其最终目的在于构建资源发现服务，最大程度地保障科研用户对资源的发现与全文获取。基于大元数据体系，NSTL将在数据、资源以及知识层面构建多层次服务体系。但是，海量多源异构数据融合面临的首要问题在于数据格式标准化问题。因此，NSTL在2017年推出《NSTL统一文献元数据标准》，该标准在充分借鉴都柏林核心元数据倡议（Dublin Core Metadata Initiative，DCMI）、主流文献服务商的数据标准和ANSI/NISO Z39.96等基础上形成，为NSTL数据集成融合、数据分析和数据挖掘，以及不同应用服务系统间的互操作建立统一的数据描述体系[2]。

目前，NSTL各个业务系统均以《NSTL统一文献元数据标准》为基础进行数据描述、交换和互操作，以XML为编码语言，因此原有基于MARC的书目数据需要进行相应的扩展和转换，以适应NSTL整体业务流程的变革。

2 NSTL元数据一体化管理框架

面对当前不断发展变化的信息环境，NSTL从资源建设到数据加工，再到用户服务模式都做出了诸多改变，着力从文献服务向知识服务方向拓展。元数据一体化管理作为NSTL业务流程再造的重要环节，对多来源论文数据的规范管理及其深化应用都具有重要意义。

2.1 NSTL总体业务布局

NSTL业务流程再造的总体目标在于建立支持知识化服务的业务布局和流程，总体业务框架见图1。在新的业务布局中，NSTL着重加强了业务模块的整体化建设，主要体现在内部业务资源建设、数据管理以及面向用户的系统服务。

2.1.1 立体化信息资源建设

在商业出版、开放获取等科技信息资源出版传播模式的共同影响下，NSTL资源建设对象从印本资源扩展到电子资源、开放资源乃至第三方数据资源，购买不再是资源建设的唯一途径，资源的采集和合作共享成为开放环境中资源建设的重点，亟待通过强化资源发现、评估、采集（合作获取）来扩大资源获取范围。

从业务层面而言，对资源建设流程的再造重点具体包括：资源类型扩展，同时涵盖传统类型文献和新型数据资源等；从文献订购管理向采集渠道管理扩展，按照资源获取渠道，可以分为订购管理和采集共享管理；强化资源版权属性和过程文档管理，以确定各类资源的具体服务方式和服务对象。

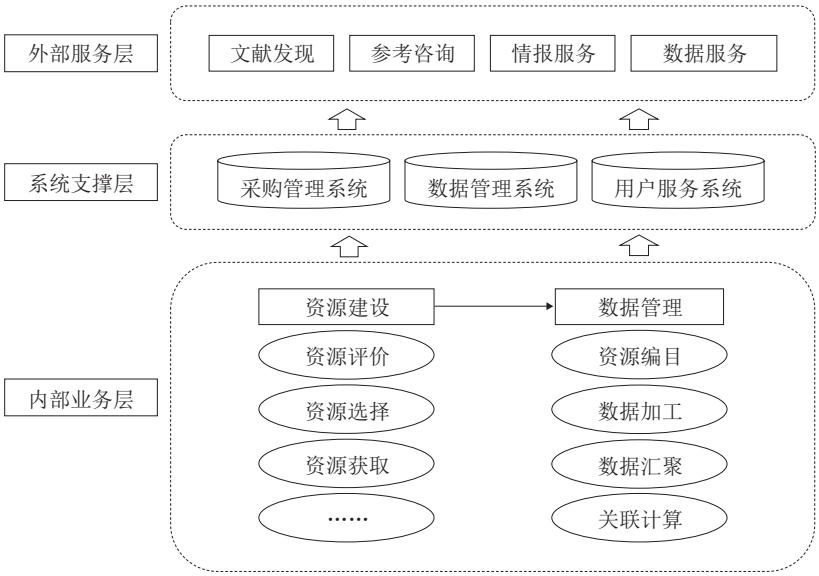

图1 NSTL总体业务框架

2.1.2 关联化文献数据管理

数据管理包括文献元数据和非文献元数据等各类元数据的管理，根据NSTL数据管理业务现状，文献数据管理划分为书目数据管理、文摘元数据加工集成、元数据增值计算和主题标引[3]。其中，书目数据管理是对各来源资源进行编目以及名称规范，涉及NSTL订购或合作获取的印本文献、开放资源和数字资源。同时，从书目数据集成库析出的调度信息将融入NSTL发现系统的资源调度知识库。此外，文摘元数据加工集成的重点在于自主加工处理包括引文数据在内的文摘元数据，并与经元数据管理系统处理的第三方元数据进行关联挂接，形成统一的文摘元数据集成库。

2.1.3 智能化系统服务增强

NSTL用户服务系统的建设将以知识发现为目标，以知识与知识、数据与数据、用户与用户、知识（数据）与用户之间的关联、计算与聚合为基础，构建NSTL知识发现系统。该系统在资源端能够对多载体、多类型、多来源资源进行统一集成揭示，通过知识组织与关联揭示实现资源增值。在服务端，系统支持用户元数据快速搜索发现与排序，并通过统一认证与分级服务，实现资源统一配置与调度，基于增值数据与关联计算结果，帮助用户发现相关的资源和服务。

2.2 元数据管理在NSTL业务流程中的功能定位

元数据管理介于资源建设和数据管理两大业务模块，它既需要对从第三方采集获取的论文元数据进行规范处理，又要从中析出品种和卷期信息，与编目数据融合形成书目数据集成库，支持后续文摘元数据融合以及资源调度计算，元数据的具体管理思路见图2。

2.2.1 实现多来源、多层级元数据规范集成

元数据管理对象同时涵盖论文元数据、馆藏信息以及书目数据。在论文元数据层面，元数据管理需要对从第三方获取的论文元数据进行格式转换与校验、人工质检与规范、品种挂接和卷期规范等，向数据加工模块流转规范化的论文元数据；在实体馆藏层面，元数据管理需要做好第三方图书馆OPAC数据与NSTL本地馆藏数据的融合处理；在品种层面，联合编目系统除了继续做好原有印本文献编目外，还需加强对开放资源和E-only资源的编目，同时扩展对于非析出文献的增强编目，形成多来源元数据获取、规范以及匹配融合机制。

2.2.2 构建覆盖各类资源的实体名称规范库

基于NSTL自有馆藏资源以及从第三方获取的论文元数据以及馆藏信息，书目数据集成库将基本覆盖各

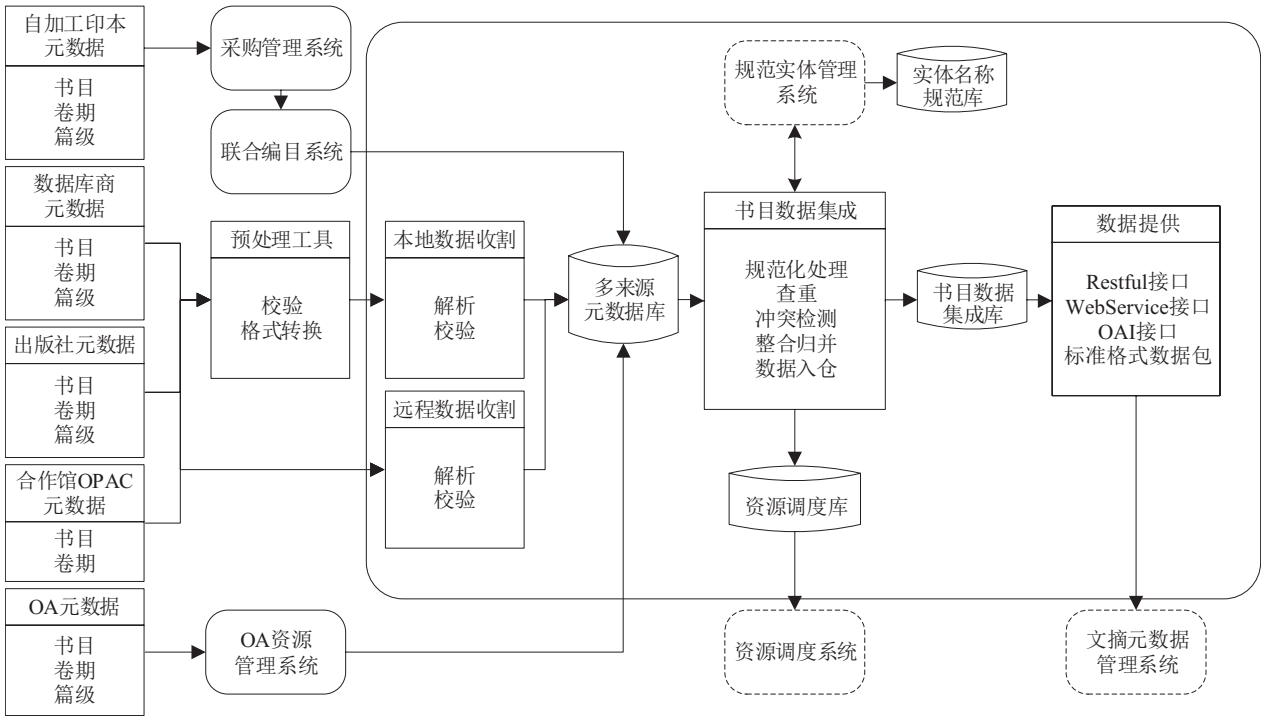

图2 NSTL元数据管理逻辑框架

类主流科技文献资源，各类资源在历史沿革、名称规范方面存在诸多交叉重复，需要构建统一的实体名称规范库。根据NSTL资源建设现状，实体名称规范库主要包括期刊名称规范库以及会议名称规范库。期刊名称规范库能够显示期刊关停并转等历史沿革关系，会议名称规范库能够显示会议举办的历史变化情况。实体名称规范库作为一项基础工具，它对支持元数据的高效融合，在服务系统中实现资源的精准导航具有重要作用。

2.2.3 基于资源多元属性支持服务调度计算

文献资源属性包含多个方面，除了题名、摘要等偏重内容的描述性信息外，不同来源渠道元数据的版权特征也是影响文献资源与用户契合度的关键因素。一般而言，图书馆的元数据主要是对其纸质馆藏的描述，因此对应的全文服务方式以纸质馆藏借阅和文献传递为主，存在服务时效滞后问题，但是用户受众面较广。除了开放获取资源可以直接访问全文外，学术出版商等来源元数据描述的一般为电子订阅资源，只有处于特定IP范围内的用户才能访问全文。元数据同时包含资源描述信息和渠道信息，因此能够有效支持资源调度知识库对资源对象、服务主体以及用户的匹配关联，从而在最大程度上为用户提供情景敏感的服务。

2.3 NSTL元数据一体化管理核心流程

元数据集成管理系统实现了为下游NSTL大数据平台提供经校验转换、质检规范后的论文元数据，为NSTL资源发现系统提供多馆藏书目集成数据、多馆藏卷期集成数据、物理馆藏信息数据和数据库品种信息等，支持服务系统的资源发现功能，主要工作流程包括数据预处理、书目元数据归并集成以及实体名称规范库构建。

2.3.1 第三方元数据预处理

针对NSTL通过合作共享获取的20余家外部机构论文元数据，NSTL以《NSTL统一文献元数据标准》作为统一标准规范，来统一各类型资源的描述格式，建立对不同格式、不同类型的元数据进行统一规范控制的方法和策略，构建完备一致的多来源元数据规范模型，促进多条薄元数据整合为单条厚元数据。在同一标准规范基础上，多源异构的论文元数据将经过格式验证与转换、查重、质检等形成统一的元数据资源。

2.3.2 书目元数据归并集成

对于不同来源、不同类型的书目数据建立归一规则库及冲突发现机制，利用规则最大程度自动化合并多条重复书目数据，减少人工操作，保证多条重复书目数据记录聚合归并为一条记录，形成多来源书目数据集成数据库[4]。同时，需要完整存储和分析存在冲突的数据记录，采用机器学习与人工结合的方法解决数据冲突，并对多来源书目数据聚合时通过遴选与测评基准数据源建立归一规则，使不同来源的书目数据能够相互补充，为资源发现系统提供全面、准确的书目数据。

2.3.3 实体名称规范库构建

针对第三方元数据预处理结果，对规范书目信息进行管理并形成NSTL实体名称规范库，以“品种”为单位，汇聚同一资源名称的不同表达形式，梳理资源间的有效关联关系，逐步形成覆盖各类文献资源的NSTL实体名称规范库，在规范名称的基础上进一步形成卷期规范库，支撑NSTL资源融合与知识服务。

3 多源异构元数据质量控制策略

各来源原始数据遵循标准格式不同、元数据薄厚程度不一，甚至存在数据内容错误等情况，因此需要对各渠道的元数据进行分门别类地管理，涉及多来源元数据格式解析与验证、格式统一映射与转换、相似度计算以及数据增强等多个环节[5]。例如，期刊论文元数据质量控制策略如图3所示。

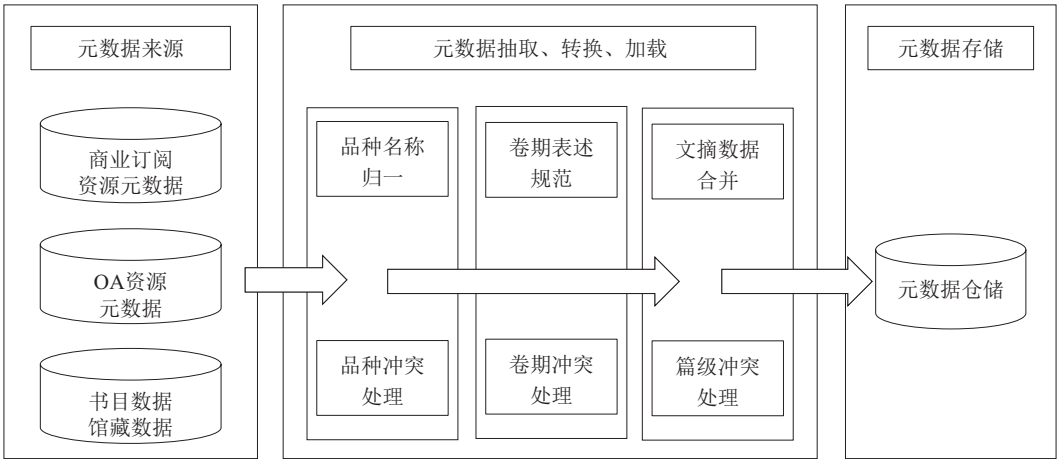

图3 期刊论文元数据质量控制[5]

3.1 基于来源元数据格式的解析与验证

网络通信过程中需要传输数据，常用的数据格式有两种，即JSON（JavaScript Object Notation）和XML。在NSTL目前已有元数据合作的来源中，多数出版社使用XML格式提供元数据。由于不同来源元数据遵循的标准不一，其对应元数据文件的逻辑结构、文件构成的元素、元素的属性以及元素和元素属性的关系存在差异。因此，来源元数据遵循标准对应的XML Schema或DTD文件，作为XML文档结构的定义和描述，是元数据管理主体对获取元数据进行格式解析和验证的主要依据。

根据来源元数据标准对应XML Schema或DTD文件，元数据管理方能够更好地理解不同来源元数据标记符的语法规则，并构建专门的元数据解析器。在加载XML文件路径及XML文件基础上，元数据解析器获取数据文件中的相关元素并进行解析。同时，XML Schema或DTD文件作为数据结构说明文件也是验证数据文件元素、属性是否完整和准确的重要工具，解析后的元数据还需要经过XML Schema格式校验，以确保获取元数据符合来源元数据标准，这是元数据管理主体对获取元数据最基本的质量要求。

3.2 基于映射转换规则的元数据格式统一

数据标准化是对多源异构元数据同构化的过程，

基于核心元数据标准对不同来源数据进行格式转换，有利于元数据规范和交换。根据来源元数据标准与核心元数据标准间的映射转换规则，对格式校验合格的元数据进行字段映射和格式转换，能够使所有元数据按照统一的标准格式描述资源。需要注意的是，原始格式数据仍然需要保留，以便后期溯源。此外，在元数据格式转换和标准化过程中，需要对每篇文献相关期刊、作者、基金项目等科研实体赋予唯一标识符，以作为后续抽取科研实体和回溯的管理依据。

目前，除大型学术出版商和二次文献提供商使用自定义元数据规范外，多数学术出版商仍是遵循JATS标准加工元数据。NSTL以《统一文献元数据标准》为中心标准，基于各元数据来源提供遵循标准的XML Schema或DTD文件，形成元数据标准之间的映射转换规则，实现各来源元数据在形式上的统一。

3.3 基于机器学习的元数据相似度计算

对同一数据渠道来源数据的重复情况，需要制定一系列规则进行查重，识别重复数据。与传统通过人工对核心字段设置权重的查重方式不同，利用各类机器学习算法进行数据查重，能够在较大程度上提升元数据相似度的计算效率。具体而言，通过输入数据重复样本，利用神经网络分类算法，生成判断数据重复的数学模型，进而对存量数据以及增量数据进行分类。对于神经网络的输入层，输入数据分别为标题、摘要、作者、关键词、作者机构等参数。对于标题、作者、关键词和作者机构等非常短的文本，可以延续系统之前使用的编辑距离来确定字段之间的相似度；对于摘要，可以通过SimHash算法来判别摘要的相似度。

以期刊论文为例，期刊自上而下包含期刊品种、卷期和论文3个层面，相应的查重规则也应该从上述3个层面依次展开。在品种层级，通过期刊名称、ISSN、出版机构等信息，利用机器算法自动筛选出疑似重复的品种，根据机器筛选的结果进行人工比对。对重复资源在品种层面进行合并、编辑，对不重复而相似资源进行人工标记，用于再次排查疑似重复的参考。在卷期层级，需要对卷期信息以及卷期下设论文进行查重，以判断是否为重复卷期。在论文层级，依据DOI、论文题名、起始页码、总页数等对同一品种、同一卷期下论文元数据进行查重。机器自动查重结束后，为保证结果的准确性，仍需要以人工方式再次进行比较判断，确定是否为同一论文元数据并进行冲突解决。

3.4 基于逻辑验证的元数据增强

除格式规范性外，作者与机构的对应关系、关键词拆分准确性、元数据内容与来源网页内容相符性等内容层面的数据质量问题也需要重视。因此，为了进一步确保元数据质量，提升元数据在后期服务中的可用性，对符合来源元数据标准的元数据需要进行逻辑验证（部分逻辑验证规则见表1）和媒介对象损坏验证，任何一步验证失败的数据都将被系统退回，经过机器和人工修正后再次进行校验，直至完全通过校验。例如，施普林格·自然姊妹公司Digital Science开发的科研创新引擎Dimensions主要是基于自定义的元数据标准，对各来源数据补充额外的元数据字段以及元数据信息，以及深度标引实现全文层级的出版信息元数据增强。

表1　元数据逻辑验证规则示例

字段名称	规则类型	规则说明
DOI	长度验证	小于70
	有效性验证	“10.”开头
ISSN和EISSN	格式验证	8位阿拉伯数字，最后一位可为x，前4位与后4位之间可有“-”
出版年	必备性验证	不为空
	长度验证	等于4
卷号和期号	长度验证	不大于25
题名	必备性验证	不为空
	长度验证	1~450
	有效性验证	筛查出包含front cover或table of content或blank page的字符串
摘要	长度验证	大于30
作者	长度验证	2~5 000
	对应性验证	作者与机构存在对应关系
作者机构	长度验证	大于2
页码	长度验证	小于50
刊名	必备性验证	不为空
	长度验证	3~160

数据逻辑验证一般可通过程序自动实现，例如PubMed Central的在线校验工具可基于自身设定的规则对数据进行验证，对与规范不符的元数据发出警告或报错[6]。随着元数据来源的不断增加，数据逻辑验证规则并非固定不变，在日常自动校验外还需要继续采用人工质检方法，按照一定比例对数据进行人工抽检，

以人工质检结果结合自动验证结果作为有监督学习训练集，定期进行最优模型训练。其中，预警模型用来检验和优化数据逻辑验证规则的合理性，预测模型用来评估各渠道元数据质量，以支持确定服务利用的优先顺序。

4 NSTL元数据管理系统建设实践

2019年，NSTL基于自身业务流程再造需要，开始设计开发元数据管理系统。目前，该系统已初步完成开发，并实际应用于NSTL业务中。

4.1 系统结构

NSTL元数据管理系统覆盖数据采集获取、格式校验、映射转换、查重规范、集成归一的元数据管理全生命周期。系统以管理规则发现构造及维护为核心，将机器学习计算与人工训练核查相结合，通过多重迭代优化，实现多源异构元数据集成工作流的高效流转运行。图4为NSTL元数据管理系统的基本结构。

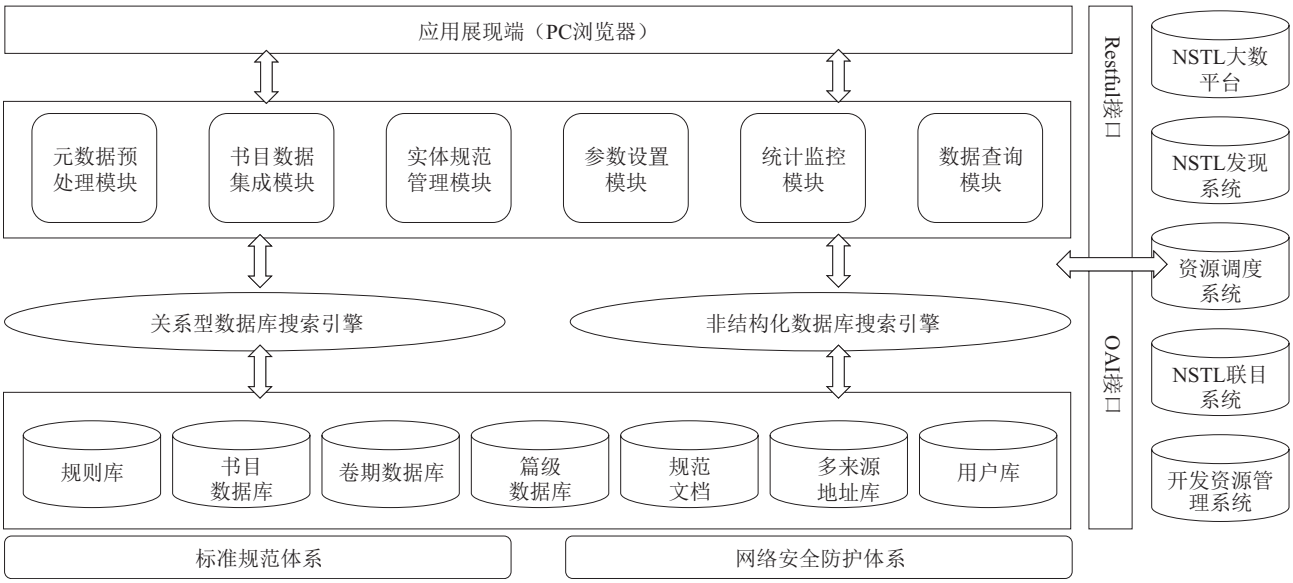

图4 NSTL元数据管理系统的基本结构

4.2 主要功能模块

NSTL元数据管理系统的主要功能模块包括元数据预处理模块、书目数据集成模块、实体规范管理模块。

4.2.1 元数据预处理模块

系统通过对第三方来源的论文元数据进行格式检验、数据解析转换、按来源查重、机器质检、人工质检，以及卷期及书目元数据的析出，形成符合NSTL元数据标准的第三方来源的元数据预处理库。基于预处理各环节，系统开发了预处理库整体流程的监控管理统计页面，可实现从数据按来源按批次经格式增强校验、批次及来源内查重、机器质检到人工质检的全流程直观的监控管理。同时，系统基于实际业务需求，开发完成了“刊频配置管理”“文献到货统计”“卷期完整性统计”“字段齐备性统计”“刊种覆盖情况统计”的数据统计功能。

4.2.2 书目数据集成模块

书目数据集成模块能够实现对多来源印本编目数据、OPAC馆藏信息以及数据库涵盖资源品种书目数据的统一管理，以及不同类型数据之间的关联耦合，通过书目、卷期元数据的集成、归一和关联，建立统一的书目、卷期集成库。系统可通过OAI方式收割NSTL联合目录书目及登到数据，同时支持从标准接口收割其他系统书目数据，也可通过Excel格式文件按规范要求导入其他来源书目数据，导入后数据将进行格式校验、内容增强校验以及人工内容抽检，并按照数据映射规则统

一规范书目及卷期数据，支持书目及卷期数据的增删改查操作，最终形成书目、卷期规范库，并提供接口供其他系统调用。

4.2.3 实体规范管理模块

系统根据内嵌规则对书目数据进行自动归并，也可进行编辑与规范，以补全规范实体的相关信息。在归一中如果发现冲突则进行人工的冲突解决处理，同时对期刊沿革、别名等关系进行梳理，以形成多来源实体名称规范库，并提供接口供其他系统调用数据。另外，通过校验和修正的数据可根据制定好的查重规则，通过机器算法自动筛选出疑似重复的资源品种，根据机器筛选的结果进行人工比对，对重复资源进行合并、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。

4.3 论文元数据集成案例

约翰·威立是NSTL元数据合作出版商之一，基于元数据管理系统，NSTL对约翰·威立提供元数据实现自动化规范管理。

4.3.1 元数据格式转换

约翰·威立提供的元数据遵循的是出版商自定义规范，包括WileyML 3G、WileyML 4.0、BPA Content、EEP、JWSCHA以及WileyML 2.1，期刊论文元数据主要根据统一XML模型WileyML 3G形成[7]。NSTL元数据管理系统根据WileyML 3G Schema对获取元数据进行初步格式校验，并以《NSTL统一文献元数据标准》与WileyML 3G之间的映射规则形成数据转换程序，将数据转换为统一格式。

4.3.2 多来源数据查重与归并

在所有约翰·威立范围内完成数据查重后，各来源数据需要进行渠道交叉查重与归并。例如，*Pediatrics International*等10种期刊由约翰·威立学术出版商出版，同时被Web of Science收录，以该10种期刊2020年出版论文元数据为例，受数据提供及时性、数据收录标准因素影响，两来源在相同期刊的论文元数据量和具体内容存在部分交叉。原始数据包含214期（10 944条）约翰·威立来源数据以及309期（16 301条）科睿唯安来源数据，共计523期（27 245条）数据。经过NSTL元数据管理系统处理后，归并为379期（18 798条）数据。

NSTL元数据管理系统共识别出8 447条来自不同来源但是指向同一篇论文的重复元数据。重复数据之间的识别主要依靠预先设置的查重规则和大数据运算实现。以一篇论文分别来自约翰·威立和科睿唯安的数据为例，两条元数据的期刊品种信息相同，论文的DOI、关键词等信息也一致，但是论文题目书写方式存在差异，来自约翰·威立的数据为“Lead‐Free Halide Perovskite Cs3Bi2xSb2–2xI9（x≈0.3）Possessing the Photocatalytic Activity for Hydrogen Evolution Comparable to that of（CH3NH3）PbI3”，而科睿唯安的数据中将“x≈0.3”写为“x approximate to 0.3”，由此造成两条数据存在差异，但是经过元数据融合将归并为一条新的元数据。

5 结语

NSTL经过20多年的发展，始终以国家科技文献保障为发展使命，经历了从印本文献，到商业电子资源，再到开放获取资源等各类资源建设模式并存的发展格局，资源类型从传统期刊、会议录等逐步向产业报告、课件、科学数据等方向拓展。在前端科技信息资源生产传播模式变革下，NSTL积极适应外部环境变化，进行了一系列业务流程重组改造，涉及资源建设、数据管理以及系统服务各个业务模块。元数据管理同时涉及论文元数据、馆藏信息以及书目数据，是NSTL业务布局从传统订购文献向立体化资源建设转变的集中体现，同时也是图书馆未来业务流程发展的方向之一。因此，图书馆界需要重视提升元数据管理能力，不拘泥于传统编目业务，才能在不断变化的科技信息环境中获得竞争优势。

参考文献

[1] 鲜国建，罗婷婷，赵瑞雪，等. 从人工密集型到计算密集型：NSTL数据库建设模式转型之路[J]. 数字图书馆论坛，2020（7）：52-59.

[2] 元数据标准服务[EB/OL]. [2021-06-21]. http://spec.nstl.gov.cn/embed/index.htm.

[3] 沈仲祺，张建勇，曾建勋. 国家科技图书文献中心业务流程再造和系统建设方案设计 [J]. 数字图书馆论坛，2020（7）：3-11.

[4] 张勇，苏学，谢振峰. 面向科技大数据的元数据仓储建设实践探索 [J]. 情报工程，2020，6（6）：84-96.

[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] File Validation Tools [EB/OL]. [2021-06-21]. https://www.ncbi.nlm.nih.gov/pmc/pub/validation/.

[7] Schemas at Wiley [EB/OL]. [2021-06-21]. http://vendors.wiley.com/schemas/index.html.

作者简介

丁遒劲，女，1988年生，博士，馆员，研究方向：数字图书馆、数字出版，E-mail：dingqj2011@istic.ac.cn。
王星，男，1977年生，硕士，高级工程师，研究方向：图书馆自动化。
李芳菊，女，1992年生，硕士，馆员，研究方向：信息组织与资源建设、元数据集成。

Research on Integrated Metadata Management in NSTL

DING QiuJin WANG Xing LI FangJu
(Institute of Science and Technology Information of China, Beijing 100038, China)

Abstract: Driven by the trend of digital publishing and changing user demands, both the model of literature resource construction and the service needs of end users have undergone major changes from the past in NSTL. It is urgent to build an integrated metadata management model for new business development goals. In the reengineering of NSTL's overall business framework, the main function of metadata management is to integrate third-party metadata, and form a publication authority file, while supporting resource scheduling. Based on the above objectives, this paper analyzes the core process of integrated metadata management and the quality control strategy of multi-source heterogeneous metadata. Taking the integrated metadata management system in NSTL for example, the implementation of metadata management is specified in the end.

Keywords: Metadata Management; Data Integration; Library's Automatic Management System; NSTL

（收稿日期：2021-06-22）