# 预印本资源元数据标准设计*

王朝阳　刘华　王丽婷
（中国科学技术信息研究所，北京 100038）

摘要：预印本是高端交流平台的重要文献资源之一，目前我国还未形成统一的预印本元数据标准，各个预印本平台的元数据结构存在较大差异，无法对国内预印本资源进行高质量整合，预印本平台与期刊编审系统的数据交互也存在问题。依据国内外预印本平台实践经验与预印本资源特点，并结合我国预印本元数据标准应用场景，从通用描述、预印本描述、质量控制以及版权信息4个方面构建预印本元数据标准，以期推动预印本资源的共享和整合，促进预印本服务的可持续发展。

关键词：预印本；元数据；高端交流平台；标准
中图分类号：TP392；T-65；G255　DOI：10.3772/j.issn.1673-2286.2024.03.007
引文格式：王朝阳，刘华，王丽婷. 预印本资源元数据标准设计[J]. 数字图书馆论坛，2024，20（3）：64-72.

预印本是指科研工作者出于和同行交流目的，自愿先在学术会议上或通过互联网发布的科研论文、科技报告等。预印本以快速、透明、开放、迭代而著称，凭借其快速发表、版本迭代、免费获取等优势，可发挥国家科技支柱效应。《中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议》明确提出将“构建国家科研论文和科技信息高端交流平台”作为“强化国家战略科技力量”的任务之一[1]。预印本作为高端交流平台的重要窗口，是适应全球开放科学发展潮流，遏制西方掠夺式占有中国学术资源的重要举措，是构建科研论文国内国际双循环交流体系的有效措施[2]。预印本平台顺应了开放获取的趋势，展现出良好的发展势头。在资助机构、公益基金、科学团体等多方支持下，预印本的影响力迅速扩大。预印本元数据标准描述预印本所遵循的一系列规范和准则，旨在为预印本的描述和组织提供统一的标准，在资源整合、资源检索、资源发现、资源评估及资源管理等应用中发挥着非常重要的作用[3]，因而预印本的元数据标准直接影响预印本资源建设水平与服务质量。然而，目前我国还未形成统一的预印本元数据标准，各个平台的元数据结构存在较大差异，无法对国内预印本资源进行高质量整合，预印本平台与期刊编审系统的数据交互也存在问题。因此，本文依据国内外预印本平台实践经验并结合预印本资源特点，试构建预印本元数据标准，为推动预印本资源建设贡献力量。

## 1　文献元数据标准发展概况

### 1.1　文献元数据概念

元数据是一种用于描述其他数据的数据[4]，也可以被称为“关于数据的数据”。它被广泛应用于各种专业领域，如信息系统的设计、构建、描述、保存和使用。元数据的核心意义就是描述数据的内容、覆盖范围、质

收稿日期：2023-12-22
*本研究得到中国科学技术信息研究所创新研究基金青年项目“面向国家级公益文献服务平台的数据治理策略研究”（编号：QN2023-04）资助。

量、管理方式以及数据的所有者、数据的提供方式等有关的信息。元数据作为一种重要的“数据基础设施”，在资源整合、资源检索、资源发现、资源评估及资源管理等方面发挥重要作用。文献资源元数据是指关于文献资源的描述信息，它提供了有关文献资源内容、特点、来源、使用权限等的基本信息。文献元数据主要包括标题、作者、出版日期、出版社、ISBN、ISSN、DOI等基本信息，同时也可能包含关键词、摘要、分类号、主题词等更详细的描述信息。文献元数据对于文献资源管理和检索非常重要[5]，可以帮助用户快速找到所需的文献资源，并了解其主要内容和特点。

## 1.2　文献元数据标准研究现状

文献元数据标准是描述文献类资源具体对象的所有规则的集合[6]。文献元数据标准的发展不仅促进了数据的共享和互操作，也提高了文献资源的可访问性和可用性。目前国际上已经出台的文献元数据标准包括Dublin Core[7-8]、Text Encoding Intitative、Encoded Archival Description[9]等。这些元数据标准已被用于文献资源联机目录和索引的编制，为学术研究者提供更为有效的检索方法。在国内，文献元数据标准的建设工作也得到了广泛关注，目前已经制定并出台了《科技平台资源核心元数据》（GB/T 30523—2014）、《科技平台服务核心元数据》（GB/T 31073—2014）、《土壤科学数据元数据》（GB/T 32739—2016）、《NSTL统一文献元数据标准3.0》、《信息与文献 数据交换和查询书目数据元目录》[10]、《中文元数据标准框架方案》[11]等规范。此外，国内学者亦在元数据领域开展了深入研究，例如：贾欢等[12]针对国外5个科学数据仓储的元数据基础信息、元数据元素、元数据应用举例以及元数据标准映射进行调查及比较研究；秦毓泽等[13]提出一种基于特征分类树与复合网络的元数据模型，该模型能够在一定程度上解决多种类元数据信息难以集成和融合的问题；庞娜等[14]对国外图书馆开放元数据服务的政策进行整理，详细调研其发展情况，为国内图书馆开展开放元数据服务提供借鉴；满芮等[15]就大数据时代科学数据元数据的开放共享进行探究，并对我国现行的科学数据元数据开放共享的相关政策法规进行归纳与分析；丁遒劲等[16]对国家科技图书文献中心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研究；赵阳等[17]基于Dublin Core对中医文献元数据标准进行研究；赵杰[18]针对基于OAI协议的预印本元数据交互框架和功能模块框架进行了研究，并提出优化建议。目前国内外的学者对文献资源元数据的相关研究主要聚焦图书馆开放元数据与科学数据等领域，针对预印本平台元数据的相关研究还比较少。

# 2　预印本元数据现状分析

随着互联网的普及和科技的发展，国内外预印本平台得到了广泛的应用。从最早的arXiv平台开始，bioRxiv、ChemRxiv、medRxiv等特定领域预印本平台相继涌现，随后预印本资源整合平台OSF（Open Science Framework）出现[19]，这些平台已经具备相对完善的预印本元数据标准，元数据应用也较为成熟，可以帮助用户搜寻、选择、发现、利用和集成预印本资源。我国的预印本发展起步于2003年，目前已建成中国科学院科技论文预发布平台、中国科技论文在线、生物医学科技论文预印本系统、中国预印本服务系统等预印本平台[20]。预印本平台要求作者在论文投稿阶段填写论文的基础信息，从而形成相应的元数据，但目前国内外仍未出台针对预印本元数据的统一标准。本文对arXiv、bioRxiv以及中国科学院科技论文预发布平台等国内外8个预印本平台的元数据进行随机抽样，按照通用描述以及特征描述两个维度对各个预印本平台的元数据描述项进行聚类，并结合预印本文献资源自身特点对聚类结果进行分析，为预印本元数据标准设计提供支撑。

## 2.1　数据来源

数据来源包括国外预印本元数据与国内预印本元数据两部分。国外预印本数据集来源于arXiv、bioRxiv、ChemRxiv、OSF中2021—2022年的预印本元数据，通过应用程序编程接口（Application Programming Interface，API）进行采集，并通过网络爬虫对数据进行二次采集，以求获取最齐备的元数据描述信息，每个平台采集1 000组数据，共计4 000组。国内预印本数据集来源于我国预印本平台2021—2022年的预印本元数据，包含中国科学院科技论文预发布平台、中国科技论文在线、生物医学科技论文预印本系统

以及中国预印本服务系统4个预印本平台的元数据，通过API进行采集。由于生物医学科技论文预印本系统以及中国预印本服务系统数据总量相对较少，为确保国内外数据样本总数相同，对中国科学院科技论文预发布平台、中国科技论文在线分别随机抽取1 500组数据，对生物医学科技论文预印本系统以及中国预印本服务系统随机抽取500组数据，共计4 000组数据。

## 2.2 数据分析

去除无效和异常的数据后，对相同时期的8个来源的国内外预印本元数据集进行横向比较分析，并对相同含义的描述项进行整合，实验数据中各个平台元数据齐备性超过50%的描述项包含标题、关键词、摘要、分类、发表时间、作者、资源链接、版权管理、评论等，共计29项。预印本作为一种新兴的文献形式，在元数据描述主体部分与期刊论文、会议论文等文献类型存在共性，同时也存在大量针对预印本资源特性的描述信息。因此，将描述项划分为“通用信息”“预印本信息”两部分。通过比较不同预印本平台元数据描述项的设计，分析不同平台在预印本元数据方面的差异和相似之处，以及平台侧重的应用方向。

（1）预印本元数据描述。不同平台在描述预印本元数据时采用的标准和规范存在差异，元数据描述各有侧重，元数据描述项数量从12项至25项不等。针对同一含义的描述项各平台所使用的名称有所不同，导致预印本元数据之间存在不兼容的问题，这不利于异构分散的预印本资源的整合。预印本元数据描述差异示例见表1。

（2）通用信息。预印本与期刊论文、会议论文等文献资源都是学术论文的形式，对该资源本体的描述已经相对成熟且趋于统一，如标题、作者、关键词等元素均为各个预印本平台的必备字段，且都与Dublin Core的描述基本一致。除此之外，多个预印本平台在元数据描述中还对资助情况进行了详细标注。由于部分资助与政府部门或专业机构相关，这些机构通常会对资助的项目进行严格的审核和评估，通过在论文中注明这些资助来源也可以增加研究的可信度和权威性。针对预印本唯一标识，arXiv等多个预印本平台选择使用DOI，DOI的兼容性和互操作性使得预印本文档在不同的系统和平台上都能被正确识别和处理。在整合多个预印本平台和数据库之间的信息时，DOI可以作为一个共同的标准，使得不同来源的信息能够相互关联。通用信息元数据字段如表2所示。

表1 元数据描述差异示例

| 元数据字段 | arXiv | bioRxiv | OSF | ChemRxiv | 中国科学院科技论文预发布平台 |
|---|---|---|---|---|---|
| 标题 | Title | Title | Title | title | Title |
| 作者 | Authors | Authors | Creator | authors | Authors |
| 摘要 | Abstract | Abstract | Description | abstract | Abstract |
| 分类 | Category | Category | Subject | Category | Category |
| 发表时间 | date | date | Date | publishedDate | date |

表2 通用信息元数据字段

| 元数据字段 | arXiv | bioRxiv | OSF | ChemRxiv | 中国科学院科技论文预发布平台 | 中国科技论文在线 | 生物医学科技论文预印本系统 | 中国预印本服务系统 |
|---|---|---|---|---|---|---|---|---|
| 标题 | √ | √ | √ | √ | √ | √ | √ | √ |
| 关键词 | √ | √ | √ | √ | √ | √ | √ | √ |
| 摘要 | √ | √ | √ | √ | √ | √ | √ | √ |
| 分类 | √ | √ | √ | √ | √ | √ | √ | √ |
| 发表时间 | √ | √ | √ | √ | √ | √ | √ | √ |
| 作者 | √ | √ | √ | √ | √ | √ | √ | √ |
| 联系方式 | √ | √ | √ | √ | √ | √ | √ | √ |
| 作者机构 | √ | √ | √ | √ | √ | √ | √ | √ |
| DOI | √ | √ | √ | √ | √ | | | |
| ORCID | √ | √ | √ | √ | √ | | | |
| 基金 | | | | √ | √ | √ | √ | |

（3）预印本信息。由于预印本具备可迭代、开放评论、开放协议可选择等与传统学术文献不同的特点，各预印本平台结合预印本资源特点与平台功能，通过元数据项对预印本版本管理信息、出版管理信息、评论管理信息、版权管理信息等进行全面的描述。①版本管理信息。各个预印本平台均结合预印本可更新的特点，增加了对版本与版本更新时间的描述，从而更好地对预印本的版本迭代进行管理与追踪，辅助用户更好地跟踪研究进展。此外，随着研究的推进，作者可能会发现并修正预印本中的错误或遗漏。通过记录预印本版本更新，可以确保读者获取到最新、最准确的信息。②出版管理信息。目前在自然科学领域，多数期刊和出版机构对预印本持开放态度，允许作者在论文出版前在相关预印本平台发布论文，因此多数预印本平台对论文的出版信息进行了追踪与描述。当预印本在正式期刊上发表后，研究内容的科学性与创新性得到证实。③评论管理信息。由于预印本尚未经过正式的同行评议，bioRxiv等预印本平台提供了评论和反馈功能，使研究人员能够收到其他学者的意见以改进他们的研究。当论文被推荐到期刊编辑部时，预印本平台的评论数据也将成为重要的评审参考项。④版权管理信息。各个预印本平台都通过元数据设计来保护其知识产权。通过在预印本的元数据中嵌入版权声明，标明版权所有人和授权范围，可以防止未经授权的复制和分发。目前多数预印本平台通过使用CC协议允许创作者将他们的作品免费分发给公众，并规定了使用者可以做什么、不可以做什么。中国科学院科技论文预发布平台与中国科技论文在线则设定了自己的开放获取协议。ChemRxiv在预印本的元数据中添加版权声明标签，帮助搜索引擎和其他工具识别预印本的版权声明，从而进一步加强预印本的知识产权保护。

预印本平台针对各自运营模式与学科侧重还提供了其他个性化的元数据字段。OSF除本站资源外还提供27个预印本托管库的资源检索服务，包括AgriXiv、BITSS、Earth ArXiv、engrXiv、FocUS Archive等，因此OSF在知识产权描述中增设了数据来源这一字段，保证文献信息的准确性、一致性和完整性，并帮助用户找到原始文献，以便查阅更多详细信息。arXiv对学科分类进行进一步补充，论文的分类不仅有Spec和Categories，还有MSC Class（数学分类）和ACM Class（计算机分类），便于数学领域和计算机领域的用户进行查找和分析。预印本信息元数据字段如表3所示。

表3　预印本信息元数据字段

| 元数据字段 | arXiv | bioRxiv | OSF | ChemRxiv | 中国科学院科技论文预发布平台 | 中国科技论文在线 | 生物医学科技论文预印本系统 | 中国预印本服务系统 |
|---|---|---|---|---|---|---|---|---|
| 资源链接 | √ | √ | √ | √ | √ | √ | √ | √ |
| 平台唯一标识 | √ | √ | √ | √ | √ | √ | √ | √ |
| 推荐引用 | √ | √ | √ | √ | √ | √ | √ | √ |
| 版权管理 | √ | √ | √ | √ | √ | √ | √ | √ |
| 版本 | √ | √ | √ | √ | √ | √ | √ | |
| 更新时间 | √ | √ | √ | √ | √ | √ | √ | |
| 是否出版 | √ | √ | √ | √ | √ | √ | | |
| 出版年 | √ | √ | √ | √ | √ | √ | | |
| 出版商 | √ | √ | √ | √ | √ | √ | | |
| 评论 | | √ | | √ | √ | √ | √ | |
| 使用统计 | | √ | √ | √ | √ | √ | √ | |
| OAI协议 | √ | √ | √ | √ | √ | | | |
| 资源类型 | | √ | | √ | | | | |
| 语言 | | √ | | | √ | | | |
| 来源 | | | √ | | √ | | | |
| 主题 | | | √ | | | | | |
| 国家/地区 | | | √ | | | | | |
| 其他分类 | √ | | | | | | | |

### 2.3 预印本元数据资源的特点

（1）发展速度快。随着科研人员数量的增加和科研活动的日益活跃，预印本平台数量以及载文量都达到了新的高度。预印本平台建设的高峰期是2016—2022年，平均每年有近15个新预印本平台问世，2017年更是多达21个，平台累计总数也在几年间迅速攀升至100个以上。2022年，全球主要预印本平台arXiv、bioRxiv、ChemRxiv、Preprints载文总量超过17万篇，伴随着预印本生态发展与平台扩建，有7个预印本平台的载文量超过50万篇，其中5个预印本平台的载文量超过100万篇[21]。伴随着预印本元数据来源与资源总量的不断增加，预印本元数据的描述方式与资源类型也在快速发展。

（2）标准不统一。一方面，各个机构或平台在定义和记录预印本元数据时，采用了不同的规范和标准，这种不一致性导致预印本元数据之间难以实现无缝兼容。除元数据描述存在较大不同外，预印本元数据的结构也存在差异。不同的平台和数据库在构建预印本元数据时，可能会采用不同的层次结构或分类方式，这种结构上的差异可能导致元数据的互操作性和共享性受到影响。另一方面，预印本元数据的开放共享方式也有所不同。国内外主要预印本平台主要有以下几种元数据服务方式。①API。许多预印本平台提供了API，用户可以通过编程获取预印本元数据。例如，通过arXiv的API可以获取预印本的标题、作者、摘要、关键词、DOI等信息。②OAI-PMH协议。OAI-PMH是一种标准协议，用于在不同的档案馆之间交换元数据。一些预印本平台支持OAI-PMH协议，用户可以通过它来获取预印本元数据。③RSS/Atom订阅。一些预印本平台提供RSS/Atom订阅服务，用户可以通过订阅获取新发布的预印本元数据。④数据抓取。用户可以通过爬虫程序或其他自动化工具抓取预印本平台上的元数据。通过这种方式可以获取大量的预印本元数据，但需要遵守各平台robots.txt文件的规定，以免违反相关规定。这些不同的数据与应用标准导致预印本元数据之间存在不兼容的问题，不利于异构分散预印本资源的集成。

（3）多源异构。预印本资源分散存储在不同的学术平台和数据库中，且运营主体的属性也存在较大差异。法国开放科学交流中心（Centre pour la Communication Scientifique Directe，CCSD）和开放存取知识库联盟（Confederation of Open Access Repositories，COAR）合作推出了开放获取的预印本存储库目录，从平台数量来看，以科研或资助机构为运营主体的预印本平台数量最多，占平台总数的40%。科研机构与资助机构具备更多的科研资源，无论是在论文产出还是在合作模式中均存在优势。以商业组织为运营主体的平台占平台总数的25%，其中绝大多数平台由出版商运营，并与期刊平台保持紧密联系。非营利性组织在人员以及经费支持上与前两种运营主体存在一定差距，由其运营的平台占平台总数的18%。由公众或社会团体运营的预印本平台数量最少，占平台总数的17%，这些预印本平台的文献内容更加分散，文献数量也相对较少。由于不同的平台和数据库在数据格式、存储方式等方面存在差异，而且同一平台的不同论文的元数据也不尽相同，预印本元数据也呈现出多源异构的特点。

不同预印本平台的元数据之间很难兼容，这不仅会影响预印本资源的整合和共享，也会影响用户对预印本资源的定位和使用。因此，需要对预印本资源元数据进行标准化设计，以便于系统化整合预印本资源，帮助用户快速准确定位符合其需求的预印本资源，进而实现物理分散、跨领域预印本资源的聚合共享以及协同服务。

## 3 预印本元数据标准设计

### 3.1 描述实体与关联关系

通过对预印本资源描述所涉及的概念、概念间关系及其包含属性进行梳理，并结合上述分析结果，构建预印本资源描述关系（见图1）。预印本资源描述关系涉及预印本、期刊、个人作者、科研项目以及评议者等多个实体，这些实体间存在着更新关系、创建关系、资助关系、衍生关系以及评价关系。

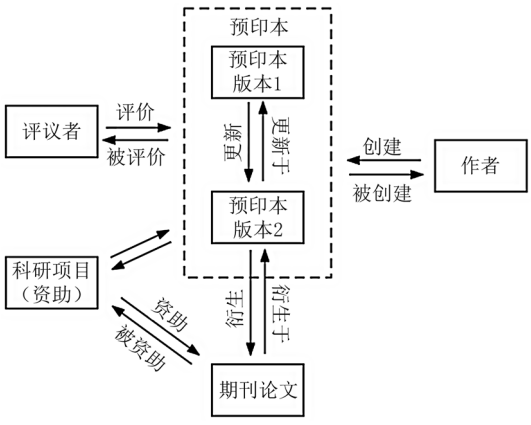


图1 预印本资源描述关系

（1）更新关系，不同版本的预印本之间存在更新关系。存在更新关系的资源可用“更新”或“更新于”关联方式连接。

（2）创建关系，在作者和预印本论文之间存在创建关系。存在创建关系的资源可用“创建”或“被创建”关联方式连接。

（3）资助关系，资助作者从事科学研究、发表学术论文的科研项目与所发表的研究成果之间存在资助关系。存在资助关系的资源可用“资助”或“被资助”关联方式连接。

（4）衍生关系，在预印本和正式发表的论文之间存在衍生关系。存在衍生关系的资源可用“衍生”或“被衍生”关联方式连接。

（5）评价关系，在评议者和预印本之间存在评价关系。存在评价关系的资源可用“评价”或“被评价”关联方式连接。

## 3.2　预印本元数据标准框架

综合考虑国内外预印本平台实践经验，设计预印本元数据标准框架（见图2）。预印本元数据标准应包括通用描述元数据、预印本描述元数据、质量控制元数据、版权信息元数据4个描述对象，而这4个描述对象又包含若干个元素集，其中每个元素集由元数据元素组成。预印本元数据元素由9个属性描述，包括中文名称、英文名称、标识、定义、类型、值域、可选性、最大出现次数以及注释。该元数据标准通过定义系列描述元素集，说明预印本论文内容主题、提供查找和定位特定论文所需信息，以及版本迭代、公开评论、论文出版等展示预印本论文特点的信息，预印本平台也能够通过该元数据标准与外部系统，如期刊编审系统等传输预印本元数据。

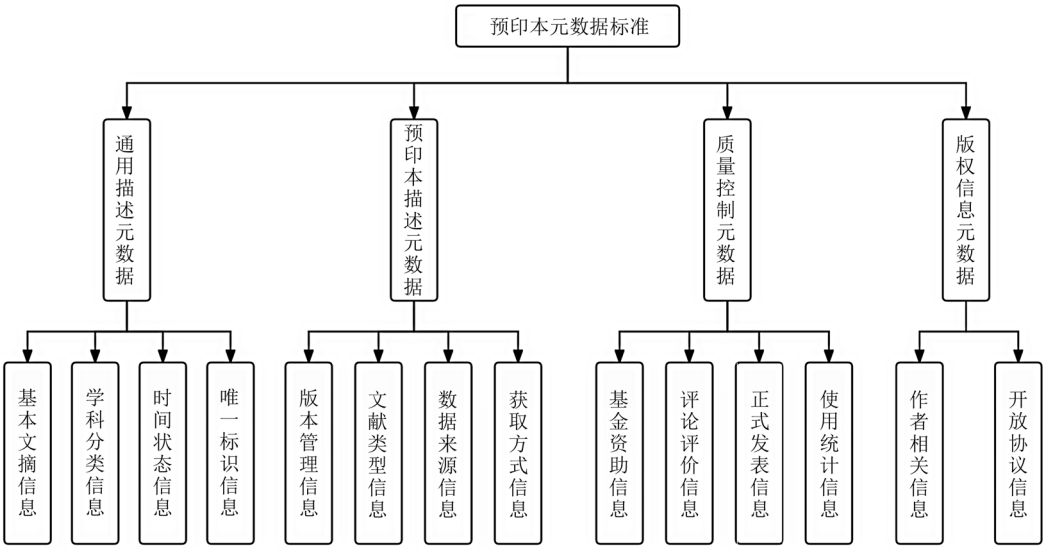


图2　预印本元数据标准框架

### 3.2.1　通用描述元数据

通用描述元数据是预印本元数据标准的主体部分，是对文献资源基本信息的通用描述，与期刊论文、会议论文等文献类型元数据标准的主体部分基本相同，是各个预印本平台的必备字段，通用描述元数据主要包括以下4类信息。

（1）基本文摘信息：包括论文标题、论文摘要、关键词、发表时间等元素，能够帮助读者快速了解预印本的主要内容。

（2）学科分类信息：包括论文所属的一级学科分类、二级学科分类以及文章主题等元素，且学科分类具备分类标准这一属性，可以针对不同的学科分类标准进行映射。

（3）时间状态信息：包括提交时间、首发时间以及下撤时间等元素，这些时间信息是学术论文生命周期的重要节点，需要被准确记录。

（4）唯一标识信息：包括平台唯一标识、仓储唯一

标识以及DOI等元素。这些标识应用于预印本资源管理，同时可以帮助用户识别和区分不同的预印本。

### 3.2.2 预印本描述元数据

预印本描述元数据主要针对预印本资源开放、迭代、类型丰富等特点设计，能够进一步突出预印本资源优势，促进预印本的传播和共享。预印本描述元数据主要包括以下4类信息。

（1）版本管理信息：包括版本标识、更新时间、修改记录等元素，这些信息有助于用户了解学术研究的发展路径、学术价值和可信度。

（2）文献类型信息：预印本存在不同的文献类型，如学术论文、研究报告、评论评介等，对预印本进行准确的资源分类有助于更好地管理和利用预印本资源。

（3）数据来源信息：通过准确标注预印本来源信息，可以提高元数据质量、促进元数据共享与资源整合、提升检索效率以及促进知识发现。

（4）获取方式信息：不同的预印本平台具备不同的预印本元数据开放获取方式，这些方式可能因平台的特性、政策和技术架构等而有所不同。

### 3.2.3 质量控制元数据

预印本质量控制元数据是用于描述和评估预印本质量的元数据，对预印本进行遴选与分级。由于预印本在发布前未经过同行评议，需要后置的评价体系与机制判断其在学术上的价值和贡献。这有助于学术界了解和认可该论文，推动相关领域的研究进展。预印本质量控制元数据主要包括以下4类信息。

（1）基金资助信息：包括基金类型、项目编号、项目名称等元素。国家级以及省部级的基金资助是对论文研究价值的认可，通常只有具有重要性和创新性的研究才能获得这类资助。因此有国家级或省部级基金资助的预印本通常代表了该领域的前沿和热点内容，可能具有较高的研究价值。

（2）评论评价信息：包括评论类型、评论时间、评论标题、评论内容、评论分值等元素。预印本平台的公开评论功能允许读者在预印本发布后立即对其进行评价和反馈。这有助于作者及时了解读者对论文的看法和建议，同时也为预印本的质量控制与评价分级提供了依据。

（3）正式发表信息：包括期刊名称、期刊论文名称、期刊论文唯一标识、年、卷、期等信息。由于预印本不属于正式发表，可以进一步向期刊投稿。学术期刊已经具备了相对成熟的评价体系，预印本正式发表的期刊信息也从侧面对预印本的质量进行了验证。

（4）使用统计信息：包括被引频次、浏览量、下载量、收藏量等元素，通过预印本的使用情况可以推断其质量，特别是被引频次。如果一篇预印本被其他论文频繁引用，该预印本就在学术界具有一定的价值和影响力。

### 3.2.4 版权信息元数据

预印本版权信息元数据是预印本版权信息的集合，可以记录和追踪预印本的版权归属和使用权限，确保作者的知识产权得到充分保护，有助于防止未经授权的复制、传播和使用，维护学术界的公正和秩序。预印本版权信息元数据主要包括以下两类信息。

（1）作者相关信息：包括作者姓名、作者邮箱、作者单位、作者唯一标识等元素，为学者主页建设提供元数据支撑，有助于读者了解作者的研究背景和其他学术成果，从而更加信任和关注其预印本。

（2）开放协议信息：包括协议类型、协议时间、协议内容等元素，标准化的开放协议信息元数据有助于促进学术交流、增加论文可见度、推动学术创新以及保障学术诚信，对于推动学术研究和进步具有积极的作用。

## 3.3 预印本元数据标准应用

提出的预印本元数据标准主要强调了预印本的资源特点与应用场景，解决了元数据标准空缺时存在的数据格式不统一、数据内容不完整、数据来源不明以及数据互操作性差等问题。预印本元数据标准应用场景如图3所示。基于预印本元数据标准，可以推进各个预印本平台通过合作联盟的形式展开合作，通过OAI-PMH协议将元数据汇缴到预印本资源发现平台，并提供原始全文链接。在这一过程中，预印本联盟实现元数据共享、全文互通，进一步整合预印本资源。合作联盟中的成员共同建立预印本评价体系，并对预印本联盟资源

进行评价，将优质预印本论文推荐到合作期刊进行发表，从而助力国家科研论文和科技信息高端交流平台建设。此外，预印本联盟可以直接与期刊编辑部的采编系统进行数据传输，作者向预印本平台投稿的同时，同步向期刊编辑部提交稿件。作者在向期刊投稿时，若同意期刊以预印本的形式预发布稿件，初审通过后采编系统会自动将稿件元数据及全文数据发送给预印本平台，预印本平台接收稿件后对其进行预发布。预印本元数据标准支撑了预印本资源元数据的规范、描述与汇交，从根本上实现我国预印本资源的汇聚融合和互联互通，为实现海量资源一站式、全方位搜索和发现服务奠定了坚实的基础。

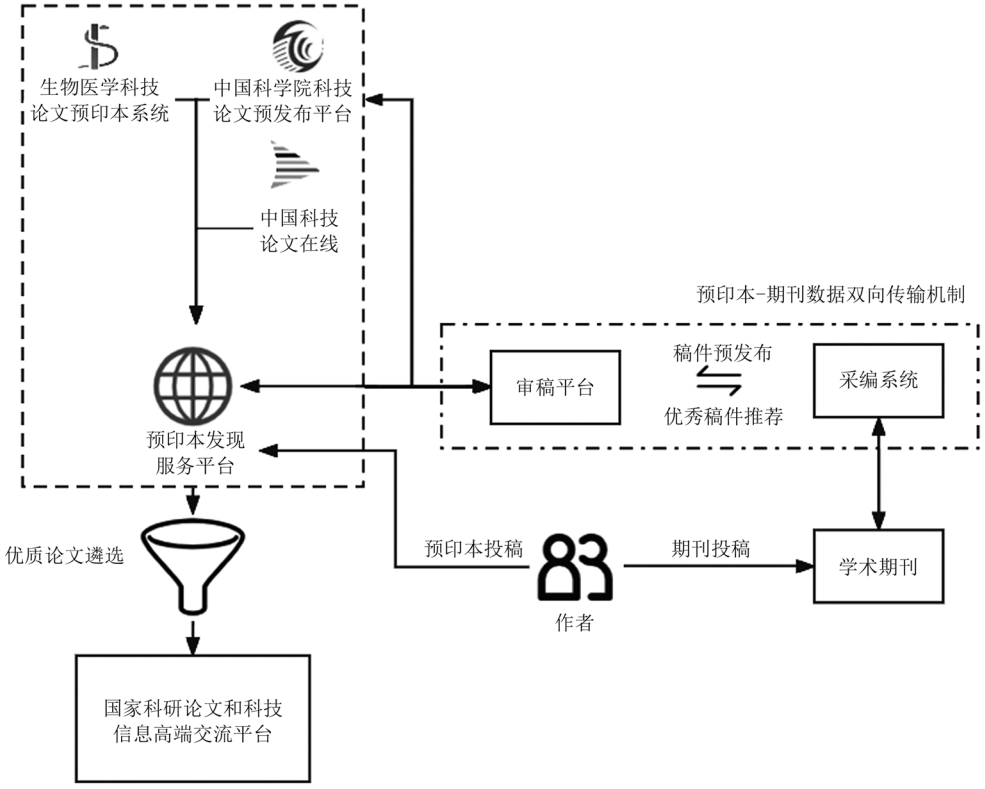


图3　预印本元数据标准应用场景

## 4　总结

建立统一规范的预印本元数据体系是当前科研领域发展的重要趋势之一。通过规范和整合预印本元数据，可以更好地促进预印本资源的共享和利用，提高科研工作的效率和成果的质量。同时，预印本元数据体系也能提高数据的可追溯性和可靠性，满足日益增长的数据安全需求。本文依据国内外预印本平台实践经验与预印本资源特点，并结合我国预印本元数据标准应用场景，从通用描述、预印本描述、质量控制以及版权信息4个方面构建预印本元数据标准。目前，设计的预印本元数据标准已经应用于国家预印本平台以及国家预印本发展联盟的资源建设。通过该标准，国家预印本发展联盟实现元数据共享、全文互通，进一步整合预印本资源。合作联盟中的成员共同建立预印本评价体系，并对预印本联盟资源进行评价，将优质预印本论文推荐到合作期刊进行发表，助力国家科研论文和科技信息高端交流平台建设。

### 参考文献

[1]　赵志耘. 构建国家科研论文和科技信息高端交流平台[J]. 数字图书馆论坛，2020（11）：1.

[2]　和鸿鹏. 预印本可否替代学术期刊？：基于科学社会学的视角[J]. 自然辩证法研究，2021，37（7）：72-77.

[3]　王英，杨新涯. 信息仓储建设的数字资源采购规范化流程研究及ERMS系统开发[J]. 图书情报工作，2020，64（12）：67-74.

[4]　中华人民共和国国土资源部. 基本农田数据库标准：TD/T 1019—2009[S]. 北京：中国标准出版社，2009.

[5]　陈伟慧. 基于大数据的智慧图书馆信息服务研究[D]. 金华：浙

江师范大学，2017.

[6] 刘万顺. 中文公安期刊全文数据库元数据标准规范建设研究[C]//湖北省图书馆学会2007年年会论文集. 2007：350-353.

[7] 吴建中. DC元数据[M]. 上海：上海科学技术文献出版社，2000.

[8] 徐佳宁. DC元数据在网络资源学科导航体系中的应用研究[J]. 图书馆建设，2002（1）：85-87.

[9] 冯项云，肖珑，廖三三，等. 国外常用元数据标准比较研究[J]. 大学图书馆学报，2001，19（4）：15-21.

[10] 高瑜蔚，朱艳华，孔丽华，等. 数据论文及关联科学数据集出版元数据标准研究[J]. 中国科技期刊研究，2023，34（10）：1270-1282.

[11] 姚伯岳，张丽娟，于义芳，等. 古籍元数据标准的设计及其系统实现[J]. 大学图书馆学报，2003，21（1）：17-21.

[12] 贾欢，李泽锋，刘越男. 多学科科学数据仓储元数据方案比较研究[J]. 档案管理，2022（4）：61-64.

[13] 秦毓泽，张辉，张军欢. 基于特征分类树与复合网络的元数据模型在大型医疗仪器共享服务中的应用研究[J]. 医学信息学杂志，2021，42（11）：67-74.

[14] 庞娜，袁钺，李广建. 国外图书馆开放元数据服务及其特点[J]. 图书情报知识，2023，40（1）：112-123.

[15] 满芮，王健. 大数据时代科学数据元数据的开放与共享[J]. 现代情报，2016，36（3）：38-41.

[16] 丁遒劲，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.

[17] 赵阳，吴开华，郑雯译. 学位论文描述性元数据的设计[J]. 图书情报工作，2005，49（6）：49-53.

[18] 赵杰. 浅析基于OAI的预印本系统[J]. 现代情报，2006，26（11）：57-58，61.

[19] 孙异凡，陈一，蒋子可，等. 开放科学视域下预印本认可政策研究[J]. 数字图书馆论坛，2021（6）：2-12.

[20] 李浩然. 预印本平台信息服务模式研究[D]. 哈尔滨：黑龙江大学，2021.

[21] 刘敬仪，杨恒，伊惠芳，等. 国际预印本平台发展态势研究[J]. 图书情报工作，2023，67（5）：15-25.

## 作者简介

王朝阳，男，硕士，助理研究员，研究方向：预印本资源建设与文本大数据处理。
刘华，女，副研究馆员，通信作者，研究方向：信息与文献标准化、信息检索和信息资源管理，E-mail：liuhua@istic.ac.cn。
王丽婷，女，硕士研究生，研究方向：资源建设、数字图书馆与知识服务。

Design of Metadata Standard for Preprint Resources

WANG ZhaoYang LIU Hua WANG LiTing
(Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China)

Abstract: Preprints are one of the important literature resources for high-end communication platform. Currently, China has not yet formed a unified preprint metadata standard. There are significant differences in the metadata structures of various preprint platforms, which make it difficult to integrate domestic preprint resources with high quality. There are also problems with the data interaction between preprint platforms and journal editing and review systems. Based on the practical experience of preprint platforms at home and abroad and the characteristics of preprint resources, this paper constructs preprint metadata standards combining with the application scenarios of China's preprint metadata standards from four aspects: general description, preprint description, quality control, and copyright information. The aim is to promote the sharing and integration of preprint resources and promote the sustainable development of preprint services.

Keywords: Preprint; Metadata; High-End Communication Platform; Standard

（责任编辑：王玮）