# 基于主题视角的数字保存研究综述及发展趋势

张乃帅* 王继民

**摘要** 数字保存作为保障数字资源长期可访问性的系统性管理活动，其研究动态与技术演进对数字文明传承具有战略意义。文章突破传统文献计量范式，基于 Web of Science 和 Scopus 数据库构建国际研究数据集，基于 CNKI 数据库构建国内研究数据集，引入 BERTopic 动态主题建模技术，通过主题强度分析、时间序列演化图谱和跨域主题网络关联，对数字保存领域的研究主题、发展趋势及知识结构进行多维度解析。研究发现，国内研究与国际研究相比，在研究全面性、研究侧重点、研究趋势和主题特征等方面存在一定差异。基于此，文章从融入信息安全机制、突破学科壁垒、拥抱人工智能、保存沉浸式体验等方面提出发展建议，为优化数字保存研究生态、制定差异化发展战略提供数据驱动的决策支持。

**关键词** 数字保存 长期保存 BERTopic 主题建模
**分类号** G250
**DOI** 10.16603/j.issn1002－1027.2025.06.012
**引用本文格式** 张乃帅，王继民.基于主题视角的数字保存研究综述及发展趋势[J].大学图书馆学报，2025，43(6)：110－122.

## 1 引言

数字对象以二进制编码形式存在，其内容解析高度依赖特定软硬件环境，在可读性、完整性与真实性层面面临根本性挑战。与物理载体逐渐损毁不同，数字对象的脆弱性体现为系统性失效：技术迭代导致格式淘汰、存储介质退化造成数据丢失、软硬件依赖引发访问障碍、元数据缺失带来语义断裂。例如，FLV(Flash Video)等曾经主流的视频格式已因技术过时而丧失浏览器支持，甚至濒临彻底不可读。此类现象揭示了数字文明易逝性与人类知识长期传承之间的深刻矛盾。

为应对这一挑战，“数字保存”(Digital Preservation)作为一个系统性的研究领域应运而生，其核心是通过持续的管理与干预，保障数字资源在技术变迁中的可持续访问与理解。国际层面已形成开放档案信息系统(Open Archival Information System，OAIS)[1]、保存元数据实施策略(PREservation Metadata Implementation Strategies，PREMIS)[2]、可信赖数字仓储库(Trustworthy Digital Repositories，TDR)[3]等重要标准，研究对象也从早期文本图像扩展至复杂数据类型。然而，数字资源的长期保存仍面临多重现实困境：数字技术快速迭代持续提出新问题，信息安全风险长期存在，物质与非物质文化遗产的数字化抢救与保存任务迫切。与此同时，数智技术的发展也为大规模、多模态数字资源的可持续保存与管理带来新的可能。

我国在数字保存领域已开展多方面探索，但在研究主题分布、方法创新与技术应用等方面仍存在提升空间。本文基于多源学术数据，采用动态主题建模方法，系统梳理国际数字保存研究的知识体系与发展脉络，比较分析国内外研究热点差异，以回顾促展望，为我国该领域的未来研究与发展提供参考。

## 2 研究综述

数字保存并非单纯的技术修复行为，而是通过策略性干预确保数字资料的持续可理解与可访问，即跨越时空与技术变迁的知识传递能力。数字保存的对象，既包括原生数字对象，也包括以各种技术手段由其他载体(如各类文化遗产)转换而来的数字对

* 通讯作者：张乃帅，邮箱：zhangns@lib.pku.edu.cn。

象。本文采用数字保存联盟（Digital Preservation Coalition，DPC）的定义，“数字保存指一系列管理活动，旨在确保在必要长的时间内对数字资料的持续访问……指的是所有必要的行动，以确保在保存介质故障或技术与组织变化时，对数字资料的访问得以维持。[4]”本文将不区分“数字保存”和“数字内容长期保存（长期保存）”，同时采用这两种表述。

数字保存作为应对文化遗产传承与数据资产长期可访问性挑战的关键活动，其研究与实践已在全球范围内形成多维度、差异化的发展路径。在过去三十余年间，国内外学者围绕数字保存的理论、技术、策略与实践产出了海量的研究成果。为了系统把握该领域的发展脉络，识别研究热点与前沿，许多学者已经开展了综述性研究。

### 2.1 早期描述性与引介性综述

数字保存研究的早期综述，主要以描述和引介为主，旨在向国内学界普及基本概念、介绍国际先进经验。例如，有研究将数字信息保存视为网络信息时代的四大棘手问题之一，并对其研究状况进行了初步描述[5]。同时，大量研究聚焦于国外进展，系统介绍国际上参考模型（如 OAIS）、标准政策、实践项目等方面的成果[6-8]，以及在数字资源长期保存的管理与技术策略上的探索[9-10]。

这些早期的综述为我国数字保存研究的起步奠定了坚实的基础，起到了重要的知识引介和启蒙作用。然而，这一阶段的综述多为定性描述，缺乏对大规模文献的系统性量化分析，视角较为宏观，难以揭示研究主题内部的精细结构和动态演变。

### 2.2 基于文献计量的宏观态势分析

随着研究文献的快速增长，学者们开始运用文献计量学（Bibliometrics）和科学知识图谱（Science Mapping）等定量方法分析领域发展态势。此类研究通过对特定时间范围内的文献进行统计，分析其年度分布、核心期刊、核心作者、高产机构以及关键词共现等，勾勒出研究的宏观轮廓[11-12]。

近年来，以 CiteSpace、VOSviewer 等可视化工具为代表的知识图谱分析方法被广泛应用。学者们利用这些工具绘制了国内外数字保存研究的合作网络、共被引网络和关键词共现网络，识别出如“电子文件长期保存”“OAIS 模型”“数字策展”“可信赖数字仓库”等研究热点和前沿主题[13-16]。这些研究极大地提升了综述的系统性和客观性，能够从宏观上揭示这一研究领域的知识结构和演进路径。

尽管如此，这类方法也存在固有的局限性。首先，它们主要依赖于关键词共现或引文关系，分析的粒度较粗，难以深入到摘要文本中捕捉主题的深层语义内涵。其次，聚类出的“热点”往往是宽泛的概念，无法精确揭示特定主题下的具体研究内容和视角差异。例如，同为“保存策略”这一热点，其内部可能包含了成本效益、风险管理、技术选型等多个细分主题，而传统文献计量方法难以有效区分。

### 2.3 聚焦特定子领域或维度的专题综述

除了宏观态势分析，大量综述聚焦于数字保存的特定子领域或具体维度，进行更为深入的梳理。依据其核心视角，此类研究大致可归纳为以下 4 类。

技术与标准视角：部分研究围绕数字保存的技术策略展开述评，如迁移（Migration）和仿真（Emulation）等方法论的比较与分析[17]；另有研究系统梳理了 OAIS 参考模型[18]、可信赖认证标准[19]等核心标准的研究进展与实际应用情况。

管理与政策视角：诸多综述聚焦于管理与政策层面的关键议题，涵盖数字资源保存中的知识产权问题[20-21]、成本模型[22]、法定缴存制度[23-24]及国家层面的政策与战略[25-26]等。

保存对象视角：针对不同类型数字资源的保存挑战与实践经验，学界涌现出大量的专题综述，研究对象广泛涉及 Web 资源[27]、电子文件[28]、科学数据[29]、社交媒体[30]、个人数字遗产[31]以及音视频资料[32]等。

区域与项目视角：不少文献对特定国家或地区在数字保存领域的研究进展、实践成果及代表性项目进行了系统总结与评述，以揭示区域特征与项目经验[33-34]。

现有专题综述为特定研究方向提供了详尽的文献指引，深化了对具体问题的理解。但其固有的“分割式”视角，也使得研究领域整体图景变得碎片化。研究者难以通过这些综述建立起不同子领域之间的关联，也难以从一个统一的、数据驱动的视角审视国际与国内研究在各个主题上的异同与互动。

### 2.4 研究述评与本研究切入点

综上所述，现有综述研究描绘了数字保存领域的大致轮廓，但仍存在以下几点不足，为本研究提供了切入点。

首先，已有研究存在方法论的局限性。传统定

性综述主观性强，覆盖面有限，而基于关键词或引文的定量分析方法，虽具客观性，但在揭示深层语义主题方面能力不足，分析粒度较粗。这些方法难以满足对大规模文献进行细粒度、深层次主题结构挖掘的需求。

其次，已有研究存在视角的割裂性。现有综述或聚焦宏观态势，或深耕特定专题，或简单地将“国内”与“国际”研究割裂开来分别论述。目前尚缺乏一项研究，能够在一个统一的分析框架下，同时对国际与国内海量文献进行主题建模，并从主题内容、主题强度和主题演化等多个维度进行系统性的比较与关联分析。

第三，已有研究缺乏数据的时效性与全面性。鉴于数字保存领域的快速发展，定期的、基于最新和更全面数据的综述是必要的。

为了弥补上述不足，本研究将采用具备强大语义理解能力的 BERTopic 主题建模技术，从 5000 余篇论文摘要的全文语境中发现和抽取更为连贯、细致和精准的主题。通过采用数据驱动的研究路径，本研究旨在克服传统方法的局限，构建一个能够跨越国界的高精度数字保存研究主题图谱，进而打破现有研究中的视角分割壁垒，在统一的“主题视角”下系统比较国际与国内研究的共同关注点、特色领域及发展趋势，最终为全面深入地理解全球数字保存研究全貌提供一个全新且更具洞察力的分析框架。

## 3 数据与方法

### 3.1 数据收集与预处理

本文采用多源异构数据融合策略，构建覆盖全球数字保存研究的知识谱系。国际研究数据选取 Web of Science 和 Scopus 数据库（涵盖计算机科学、社会科学与人文艺术等领域的权威文献），国内研究数据选取 CNKI 数据库（覆盖图书情报、档案学与计算机应用等学科），时间跨度为 1991 年（首个“Digital Preservation”术语出现年份）至 2024 年，完整覆盖数字保存研究的技术演进周期。

国际文献检索采用复合布尔逻辑表达式：TS=（“Digital Preservation” OR “Digital Curation” OR “Digital Archiving” OR “Long－term Preservation”），其中 Web of Science 执行主题字段检索，Scopus 限定于标题、摘要与关键词字段，避免全文检索带来的语义泛化风险；国内文献通过 CNKI 专业检索式 SU=（“数字保存”＋“长期保存”）实现主题精确匹配，规避单纯关键词检索导致的文献过载问题，文献类型限定为期刊论文、学位论文与会议文献，确保学术成果的代表性。

针对跨学科术语干扰问题，本文实施双重过滤机制。首先通过学科门类限定排除无关领域，国际数据库（Web of Science 和 Scopus）聚焦计算机科学、信息科学、社会科学交叉学科及人类与艺术等学科，国内数据库（CNKI）锁定图书情报与数字图书馆、计算机硬件技术、计算机软件及计算机应用、档案及博物馆、互联网技术等相关学科。其次构建语义校准规则，利用正则表达式识别并剔除生物医学、农业等领域的误检文献（如含“Cryopreservation”的低温保存研究），经人工抽检验证控制误检率。在数据清洗阶段，通过文献管理软件 Endnote 进行去重，剔除跨库重复记录。

为确保数据质量，研究团队建立三级人工校验机制。首先排除征稿启事、新闻报道等非学术文献；其次过滤摘要缺失或内容不完整的记录；最终由团队成员用双盲审阅法对文献相关性进行判定，争议文献通过共识会议裁决。经严格筛选后，共获得有效文献 5214 篇，其中国际文献 3191 篇、国内文献 2023 篇。需要说明的是，本文将中国作者和团队发表的英文论文归入国际研究进展，未严格按照作者国别区分属于国内研究进展还是国际研究进展，仅以论文语种作为区分。

### 3.2 数据分析方法

本文采用基于深度学习的动态主题建模框架，突破传统主题模型对词频统计的路径依赖，通过论文摘要语义嵌入与密度聚类相结合的创新方法，实现数字保存领域知识结构的跨时空解析与发展演化分析。完整的数据处理工作流程如图 1 所示。

传统主题模型挖掘的常用方法包括 Latent Dirichlet Allocation（LDA）[35]和 Dynamic Topic Models（DTM）[36]等，都是基于词袋假设构建概率分布，难以捕捉数字保存领域特有的跨学科术语关联。相较而言，BERTopic 模型通过预训练语言模型构建文本嵌入来表征文本的语义信息，在主题连贯性和跨文档区分度上表现更优[37]。不同于词嵌入，长文本的嵌入能够更好的理解文本中单词的语义并形成整体文本的语义表示。

BERTopic 可以使用多种基于 Transformer 的预训练语言模型生成文本嵌入，语义嵌入生成采用针对中英文分别优化的模型组合，以获得各自语言环境下的高质量嵌入。本文对英文语料使用 Bge-large-en-v1.5 模型追求单语言下的高精度，中文语料采用 Paraphrase-Multilingual-MiniLM-L12-v2 模型利用其多语言适应性，力求在异构空间中提取有效语义特征。文本嵌入过程保留完整摘要文本，通过动态长度适配算法解决长文本截断问题。

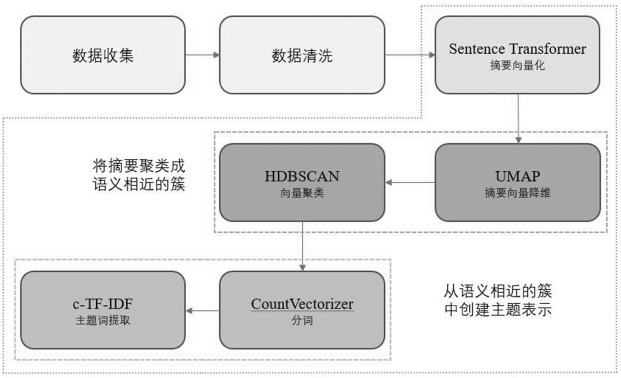


**图 1　数据处理工作流**

本研究将主题识别构建为密度聚类任务，针对高维语义嵌入空间稀疏且复杂的特性，构建“降维—聚类”两阶段主题提取框架。首先，采用 UMAP（Uniform Manifold Approximation and Projection）算法对高维语义向量进行非线性降维，通过拟合数据内在流形结构并将其映射至低维空间，在降维过程中借助局部连接性参数保持专业术语及相关文本间的局部拓扑关系。随后，基于降维结果使用 HDBSCAN（Hierarchical Density-Based Spatial Clustering of Applications with Noise）算法进行自适应密度聚类，该算法能够依据数据分布自动识别不同密度簇，并生成层次化主题结构，从而有效区分易混淆主题与核心语义簇，实现稳定可靠且层次清晰的主题划分。

主题表征优化通过 c-TF-IDF（Class-based TF-IDF）算法，增强跨语言主题表征的可解释性，对聚类后的主题进行主题词提取，形成主题标签。c-TF-IDF 是 TF-IDF 的变体，通过将 TF-IDF 调整至分类/主题上工作，并考虑使一个主题区别于另一个主题的因素，识别主题中的重要概念来更准确地表示文本的特征。

## 4　数据处理与分析

### 4.1　研究曲线

通过对 1991 年至 2024 年国内外相关数据库的统计分析，本文梳理了数字保存相关研究文献的发表情况。图 2 显示国际和国内学界对数字保存议题的关注度自 2000 年左右开始快速增长，并在 2012 至 2014 年达到高峰。其后，国际研究的发表数量虽有波动但整体保持在较高水平，而国内研究则在高峰后呈现出较为明显的下降趋势，且波动幅度较大。

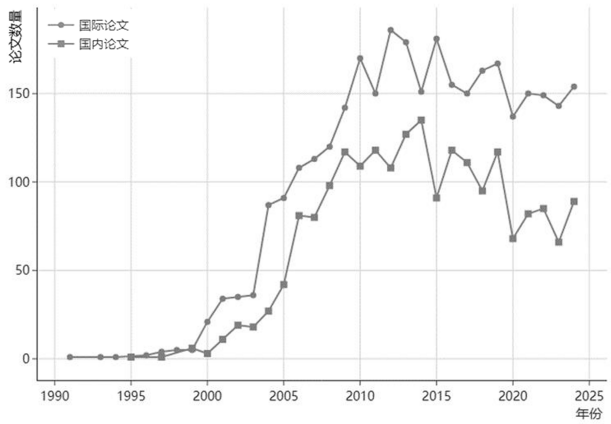


**图 2　国际国内数字保存研究论文发文量趋势对比分析**

### 4.2　热点主题分析

本文使用 BERTopic 对中英文论文摘要进行主题建模。为提高主题的集中性与解释性，在模型构建过程中，将 min_topic_size 参数设置为 20，确保每个生成的主题簇至少包含 20 篇论文，以避免产生稀疏或语义不明确的主题。

在文本嵌入阶段，考虑到停用词在中文语义构建中可能承载一定的语义特征，本文在进行中文文本嵌入时未去除停用词，而是直接使用完整的摘要文本进行嵌入生成。在嵌入后阶段，利用 HDBSCAN 对文本进行聚类后，再对各主题簇的文本进行停用词处理并提取主题词，以提高主题词的代表性和辨识度。有关停用词及分词等的设置如表 1 所示。

在完成前述预处理与模型参数设定的基础上，本文分别对国际与国内数字保存研究数据集进行了主题建模分析。通过主体建模，国际数据集共识别出 26 个有效主题，国内数据集识别出 8 个有效主题。鉴于国际研究主题数量较多，本文在全面获取主题分布的基础上，进一步从主题内容的显著性出发，筛选出每组中排名前八的核心主题作为数字保

存研究的代表性侧面，展开深入分析。随后，针对每一代表主题，提取其前五个关键词及其权重进行可视化呈现，从更精细的层面和更立体的角度来揭示国际与国内研究在长期保存相关理论、技术路径与实践关注方面的热点特征和关注差异。

**表 1　停用词、分词及语义单元设置**

| 处理项目 | 取值/设置 | 说明 |
|---|---|---|
| 中文停用词 | GitHub 常用停用词表＋补充停用词 | 合并“哈尔滨工业大学停用词表”“百度停用词表”等通用停用词表，补充“数字保存”“长期保存”“长期”“保存”等高频出现但是对区分主题作用有限的关键词作为停用词 |
| 英文停用词 | Python nltk 模块＋补充停用词 | 补充“digital”“preservation”“long” “term”“long-term”等关键词作为停用词 |
| 英文语义单元 | 设置 ngram_range 为(2,3) | 设置为 2－3，以保留语义更完整的词组 |
| 中文分词 | Jieba 分词＋论文关键词 | 使分词更准确且保证领域核心概念完整性 |

4.2.1　国际热点主题分析

国际研究主题聚类的前 8 个主题的主题词分布结果如图 3 所示，图中的数值代表主题词在该主题中的 c-TF-IDF 分数，分数越大表示该词对该主题越有代表性。

(1)主题 0：实体遗产的高精度三维建模与数字重建

该主题聚焦于实体文化遗产(建筑、考古遗址、文物等)的几何记录与数字化重建，强调利用多模态传感技术构建高精度的“数字孪生”以应对物理实体的消亡风险，共涉及 209 篇论文。随着技术手段的演进，研究重点已从早期的单一激光扫描或摄影测量，转向融合地面激光扫描、无人机摄影测量、多光谱成像及计算机断层扫描等的综合采集工作流。该方向的核心在于如何将复杂的几何数据转化为具备语义信息的档案，以支持文物的全生命周期管理、虚拟修复及预防性保护。该主题涵盖了从数据采集、点云处理、语义丰富到最终归档保存的全过程，虽然在三维重建的精度和自动化程度上取得了显著进展，但在海量三维数据的长期存储标准、元数据互操作性以及复杂数字对象的格式过时风险管理方面仍面临挑战。

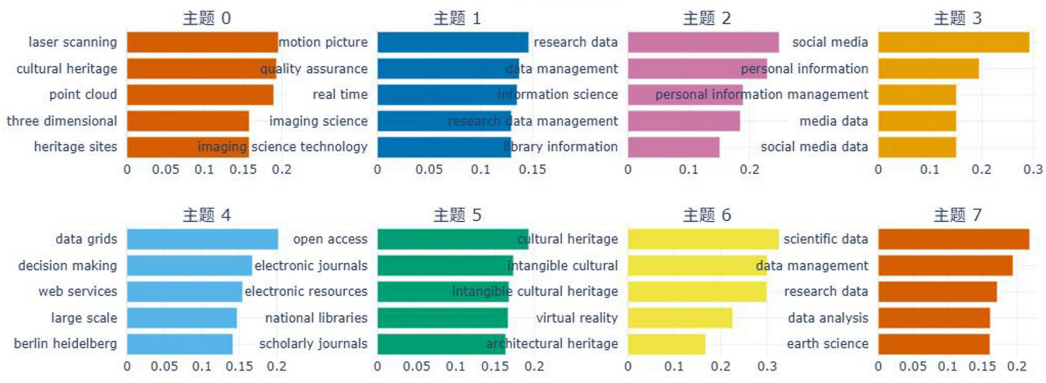


**图 3　国际数字保存各研究主题的主题词分布(8 个主题)**

(2)主题 1：多模态文档的数字保存技术及应用

该主题聚焦于将模拟形态的物理载体(文本、图像、视听资料、建筑实体等)转化为可信数字资产的技术过程与管理策略，共涉及 206 篇论文。研究内容从传统的二维文档扫描与光学字符识别，扩展至利用多光谱成像揭示受损文物的潜在信息，以及针对濒危视听载体的信号抢救与数字修复。该主题强调在数字化迁移过程中对原件内容完整性与真实性的维护，深入探讨了格式标准化、高清数字化采集、元数据管理等关键技术环节。尽管数据采集技术已趋于成熟，但在应对海量异构数据的存储成本及自动化语义提取的准确性方面仍存在显著挑战。

(3)主题 2：科研数据管理与策展

该主题包括 178 篇论文，反映了科研数据生命

周期管理在数字保存中的重要地位。从采集、存储、组织到长期保存与再利用，科学数据管理正在成为开放科学框架下的研究重点。特别是“数据策展”概念的提出，使科研数据得以系统化、结构化地存档，为科学研究的复现和知识积累提供保障。该主题已取得跨学科共识，但是在数据质量与再利用的评估、实践与政策之间的差距等方面有待进一步探讨。

(4)主题 3：个人数字档案与社会记忆保存

该主题聚焦于个人数字档案及社交媒体内容的长期保存，探讨了在数字化时代个人记忆如何转化为社会历史记录的问题，共涉及 150 篇论文。随着移动互联网和社交网络的发展，研究重点从传统的机构档案转向了个人数字遗产、网络生活记录的保存。该主题深入探讨了针对海量、碎片化且依赖于商业平台的数字内容的采集与归档策略，强调了保存过程中的伦理问题。尽管学界已提出多种理论模型与素养教育框架，但实践应用仍面临多重阻碍，主要挑战包括社交媒体应用程序编程接口(API)的技术限制、公众保存意识的薄弱，以及非机构主导档案在技术迭代下的可持续性危机。

(5)主题 4：数字保存的基础设施架构、自动化策略与可扩展性

该主题聚焦于数字保存的技术基础设施与系统架构设计，重点探讨如何通过网格计算、云存储及面向服务的架构等技术方案实现海量数字对象的可持续管理，共涉及 136 篇论文。该方向强调保存流程的自动化与智能化，涉及利用智能代理和微服务自动执行格式迁移、完整性校验及保存规划。此外，研究还深入探讨了“自我保存对象”及保存成本模型等前沿概念。尽管在技术架构的可扩展性和互操作性方面取得了显著进展，但如何将复杂的科研原型转化为低成本、易维护的生产级系统，以及应对“大数据”环境下实时保存的性能瓶颈，仍存在持续的挑战。

(6)主题 5：学术资源的机构管理与电子期刊的长期保存

该主题聚焦于学术记录的机构化管护与长期获取，特别是图书馆和记忆机构在应对电子期刊、电子书及开放获取出版物时的保存策略，共涉及 132 篇论文。随着学术交流模式从“实体拥有”向“数字许可”的范式转变，研究重点探讨了如何解决“访问权与保存权分离”的危机，以及如何构建可信的第三方保存网络。该方向强调了版权法、许可协议与保存技术的深度纠缠，关注点从早期的载体寿命延伸至内容的可持续性与永久访问权。尽管在商业出版物的保存机制上已建立了一定规范，但在应对长尾内容、开放获取资源的保存缺口以及处理动态学术成果的复杂性方面，仍缺乏系统的保存机制，存在严重的资源配置失衡问题。

(7)主题 6：文化遗产的数字保存与呈现

该主题聚焦于文化遗产的数字化保护与呈现，强调虚拟现实(VR)技术在文化遗产再现中的应用，共涉及 127 篇论文。随着技术的发展，虚拟现实和媒体艺术正在成为传承与传播文化遗产的重要手段。该方向强调保存手段的创新性和互动性，通过沉浸式互动增强公众对传统文化的理解和记忆。该主题涵盖众多文化遗产类型，技术方法多元，应用场景丰富，但是多数聚焦于技术实现，在可持续性模式探讨及效果等方面存在缺位现象。

(8)主题 7：特定领域的科学保存基础设施与软件复用

与通用的科研数据管理不同，该主题聚焦于数据密集型学科(如天文学、地球科学等)的专用保存基础设施构建与技术实现，共涉及 116 篇论文。这些研究多源于大型科研机构的工程实践，重点探讨如何针对海量观测数据构建符合 OAIS 参考模型的领域存储库。该方向超越了单纯的数据归档，深入到了科研软件与计算环境的保存，以确保科学发现的可复现性。尽管在构建大规模分布式存储系统方面积累了丰富经验，但在解决科研软件的长期维护、跨学科数据融合的技术壁垒以及旧任务数据的“知识流失”方面仍面临技术与资金的双重挑战。

整体来看，国际研究的各主题涉及的论文规模差距不大，主题涵盖了从文化遗产的数字化保存到科研数据的管理机制、从法律政策框架到技术实现路径等多个维度，充分反映了数字保存作为跨学科领域的复杂性与多样性。未来的研究会在技术、制度与实践之间做进一步协同，以实现数字资源的真正“永续传承”。

4.2.2 国内热点主题分析

国内研究主题聚类的前 8 个主题的主题词分布结果如图 4 所示。

(1)主题 0：电子文件管理与数字档案

该主题聚焦于电子文件与数字档案管理，反映出国内数字保存研究中对电子文件全生命周期管理的

持续关注，由575篇论文构成。研究通常围绕电子文件与数字档案的生成、归档、长期保存和利用展开，强调制度、技术及标准在保障数字档案真实性、完整性等方面的关键作用。该主题所涉及的论文数量多，覆盖内容广，场景多样，但是理论研究多于实践，研究同质化现象明显，跨学科融合有待进一步加强。

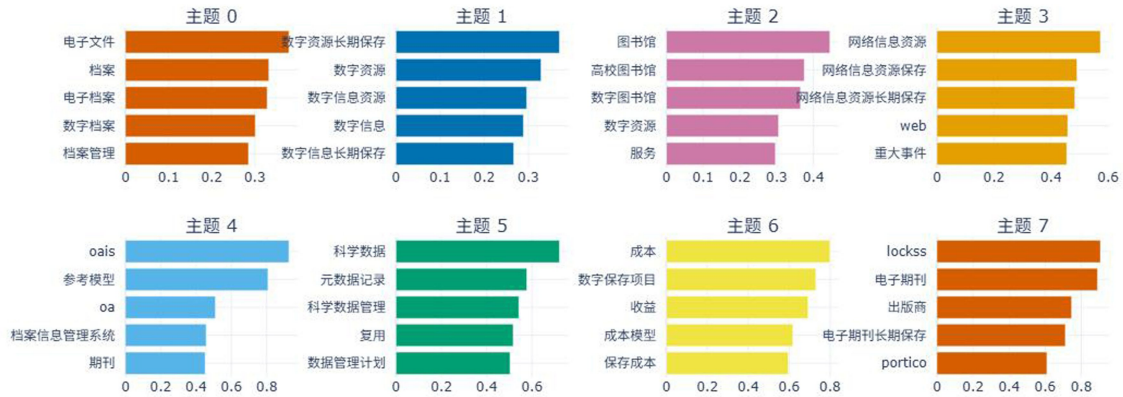


**图4　国内数字保存各研究主题的主题词分布(8个主题)**

(2)主题1：数字资源长期保存的体系架构、本土化策略与多维治理

该主题聚焦于数字资源长期保存的体系架构、本土化策略与多维治理，全面覆盖了从国家宏观政策、法律法规到具体保存技术的理论与实践，共涉及373篇论文。研究内容呈现出鲜明的“引进—吸收—本土化”特征，强调保存活动的系统性与社会性。尽管理论体系日趋完善，但在建立可持续的资金保障机制、跨机构协同保存网络以及应对复杂版权环境下的合法保存权等方面，仍面临显著的实践挑战。

(3)主题2：图书馆数字资源建设与保存

该主题共聚合363篇论文，关注图书馆在数字化转型过程中的数字资源建设与保存问题。数字资源对教学、科研、文化传播等方面的重要性日益增长，促使数字资源的建设、组织、管理与保存成为图书馆尤其是高校图书馆的关注重点，也反映了图书馆在数字保存中的实践经验与挑战。主题研究深度较为充分，对策深入，模式多样，但是同样存在偏重理论与策略研究的问题，量化分析与本土化方案有待进一步加强。

(4)主题3：网络信息资源保存

该主题关注网络信息资源的长期保存，共涉及130篇论文。研究聚焦于网络信息资源的保存意义、保存策略、责任体系和合作模式等，探讨网络信息资源保存面临的技术、经费、版权保护等方面的挑战。该主题研究内容侧重宏观策略与理论探讨，但是实践操作研究、可持续性机制研究等的深度与广度尚显不足，有待进一步拓展。

(5)主题4：数字保存的标准化体系

该主题关注数字保存的标准化体系建设，探讨数字保存系统的标准化设计与建设，由47篇论文构成。OAIS已成为数字保存领域的核心模型之一，相关研究致力于以OAIS为参照或基础，开展档案信息系统等数字保存系统的设计与建设，推动保存实践的标准化。该主题对标准化体系进行了充分的引介，但是内容同质化较为严重，缺乏本土标准体系的探索及探讨。

(6)主题5：科学数据管理与保存

该主题聚焦于科学数据的管理与数字保存，由47篇论文构成。研究强调科研过程中及科研项目结束后科学数据的可追溯性、可重复性及持续可用性，关注科学数据的描述、格式、保存周期与复用策略。该主题理论探讨与国外经验借鉴较多，实证研究相对匮乏，技术细节探讨较少。

(7)主题6：电子期刊的第三方保存系统与永久获取保障

该主题聚焦于数字学术资源的第三方长期保存，旨在应对从资源“拥有”到“许可”的商业模式转变所带来的永久访问风险，共涉及32篇论文。研究核心在于剖析两种主流保存范式，分别是以LOCKSS为代表的分布式、图书馆主导的社区网络，及以Portico为代表的集中式、专业托管的归档服务。主题进一步延伸至开放获取资源的保存策略。尽管国际方案成熟，但其本土化应用仍面临资

源适配、权责界定与资金可持续性等核心挑战。

(8)主题 7：数字保存的成本效益评估

该主题聚焦于数字保存的经济学分析，重点探讨如何通过成本模型、收益评估及投资决策来实现数字资源保存的经济可持续性，共涉及 30 篇论文。研究内容涵盖了从保存成本的构成分析、保存项目的财务评价指标，到数字作为经济产品的供需属性与外部性问题。尽管在理论模型构建上已较为成熟，但在模型的通用性、实际应用中的数据获取难度以及跨机构成本比较方面，仍存在显著的实证缺口与工具化挑战。

整体来看，国内研究的主题规模差异较大，研究主题重点集中在数字档案与图书馆数字资源保存这两个主题，涵盖了从数字档案管理到科研数据保存，从元数据规范、参考模型等宏观框架到成本效益评估等实践导向的分析维度，反映了数字保存关注视角的多样性。

### 4.3 主题趋势分析

本文利用折线图对各个主题进行趋势分析，通过颜色和标记来区分不同的数据系列，便于进一步分析。本文主要选取两个数据集中研究数量最多的前 8 个主题，开展趋势分析。

#### 4.3.1 国际主题趋势分析

基于 BERTopic 主题强度时序分析，本文识别出国际领域八大核心主题的演进轨迹，其变化趋势如图 5 所示，各主题的发展并非线性同步，而是呈现出“萌芽起步—爆发震荡—分化演进”的复杂动态特征。

1991 年至 1999 年可视为研究的萌芽期。这一时期各主题的论文数量均维持在低位，曲线平缓。这表明数字保存研究处于探索起步阶段，学术界对此领域的关注尚未形成规模效应，主要集中在基础理论和早期数字化尝试上。

2000 年至 2015 年为爆发震荡期。随着信息技术发展及数字时代的到来，特定主题呈现爆发式增长。主题 2 与主题 4 在这一阶段集中爆发，反映了研究数据的有效保存与基础设施建设在短期内受到学界的高度关注。主题 1 研究数量稳步攀升，反映了多模态文档的数字保存技术在该阶段持续积累。主题 2 在 2015 年呈现出“尖峰”特征，年发文量达到峰值，约为 30 篇，随后关注度迅速回落，可能与当时重要科研数据管理政策发布有关。

2016 年至 2024 年为成熟与分化期，各主题的发展路径出现了明显的分化，反映了研究重点从“底层构建”向“技术深化”和“应用拓展”的转移。传统热点主题(如主题 2、主题 4)呈下降或低位徘徊趋势，表明相关基础理论与架构研究已趋于饱和或成熟。新兴前沿主题(如主题 3、主题 6)关注度逐渐上升，表明数字保存的研究重心聚焦于更前沿的挑战，从对文化遗产本体的“静态保存”，迈向构建可永久留存并可持续活化的“数字生命体”。

综上所述，尽管各主题均在一定时期内受到学术界广泛关注，但从时间脉络看，不同主题受到政策导向、社会需求和技术进步等多重因素影响，呈现出明显差异性。这种“共时性与异步性并存”的演化特征，揭示了数字保存领域内的复杂机构与多元驱动机制。

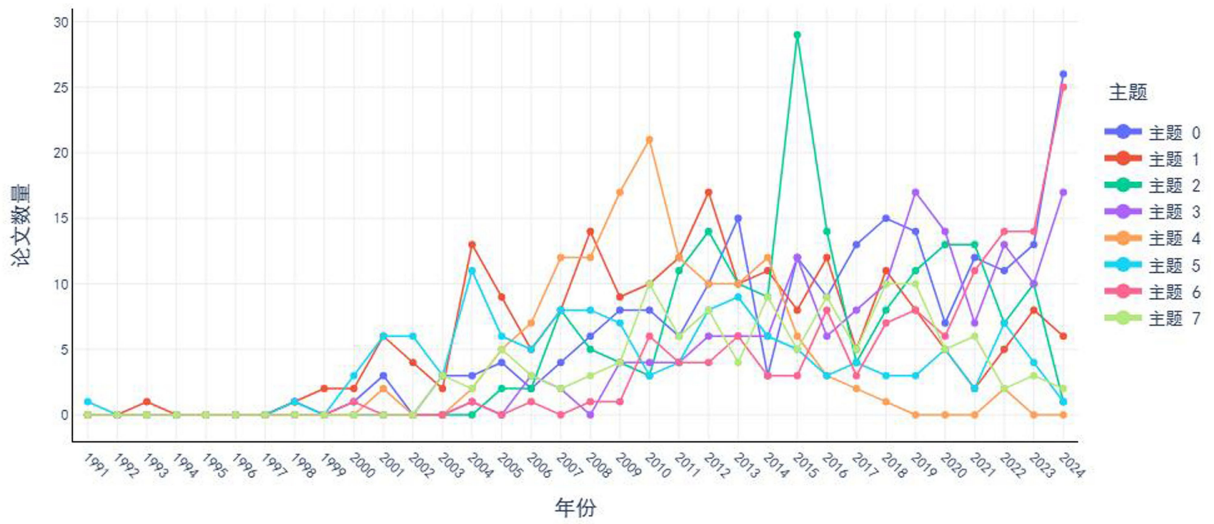


**图 5 国际研究主题趋势变化(8 个主题)**

4.3.2　国内主题趋势分析

国内数字保存研究主题的变化趋势如图 6 所示，呈现出与国际主题相似的“萌芽起步—爆发震荡—分化演进”的趋势特征。

整体而言，学术界对数字保存相关主题的研究热度在 2002 年后显著上升并保持稳定，在 2005 年前后出现多个主题的集中增长现象，反映了数字化转型背景下对数字保存相关主题的持续关注与深入探索。

从单个主题视角看，某些主题在整体稳定态势下呈现出显著不同的变化轨迹。主题 0（电子文件管理与数字档案）、主题 1（数字资源长期保存的体系架构、本土化策略与多维治理）和主题 2（图书馆数字资源建设与保存）受关注程度远高于其他主题。尤其是主题 0，自 2003 年起受关注程度持续攀升，成为了国内数字保存领域的主导力量，体现了档案部门在国家数字记忆保存中的核心地位日益凸显。

国内数字保存研究呈现出整体关注度保持稳定、部分主题关注度突显的特点。关注度平稳的主题存在进一步深入拓展的空间，是未来可以探索的方向。

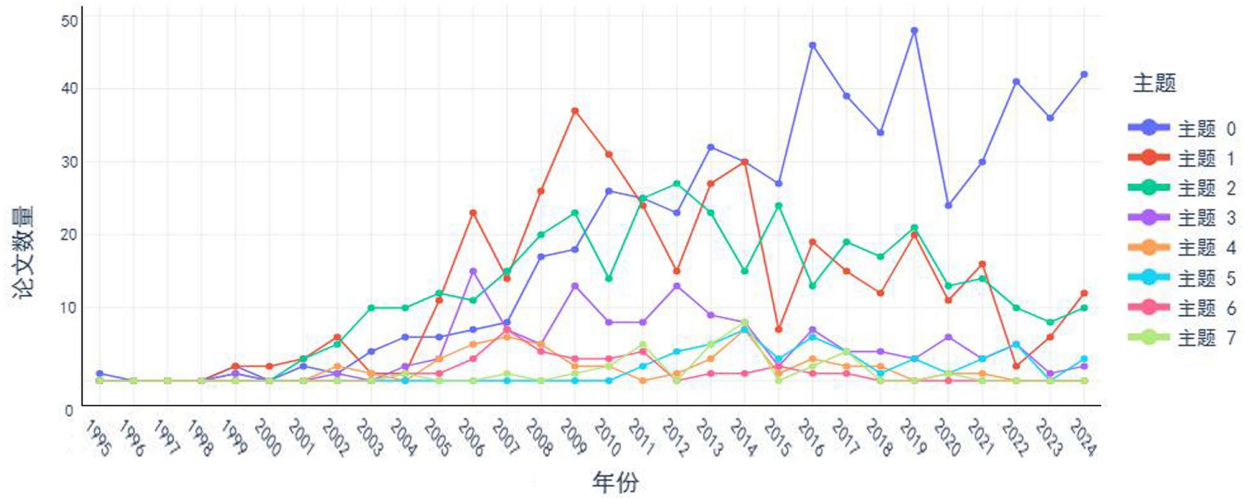


**图 6　国内研究主题趋势变化（8 个主题）**

4.4　主题可视化图谱分析

为了更直观地展示 BERTopic 识别的主题结构，本文对识别结果进行了可视化处理，生成了交互式主题图谱。其中，每个圆圈代表一个主题，圆圈的大小反映了该主题在整个数据集中的出现频率；圆圈之间的距离则表示主题之间的语义相似性，距离越近表明相似度越高。

如图 7 所示，尽管国际研究主题数量较多，但整体呈现出明显的聚类特征，26 个主题大致聚集形成了 4 个规模较大的主题簇，表明研究内容在一定程度上存在集中趋势。相比之下，国内研究主题数量相对较少，8 个主题分布较为分散，仅有部分主题之间存在较近的距离关系，但尚未形成明显的密集主题簇，反映出国内研究在该领域尚处于较为零散的发展阶段。

## 5　研究结论与展望

5.1　核心研究发现

本文基于 BERTopic 主题建模与跨域对比分析，发现数字保存研究的全球知识版图与区域发展呈现一定差异。

（1）国内研究数量接近国际水平，然而主题丰富度呈现结构性落差

从论文产出规模来看，国内研究论文数量与国际研究论文数量的比值约为 2∶3，表明我国在该领域已形成相当规模的学术产出。从主题聚类结果来看，国内研究主题丰富度不足。聚类分析时，国内研究数据集仅得到 8 个研究主题，而国际研究数据集则有 26 个研究主题，二者的主题数量比值约为 1∶3。这种差距映射出国内研究的双重局限：部分前沿方向未形成独立主题簇，主题间关联强度均值较低，跨主题协同不足。

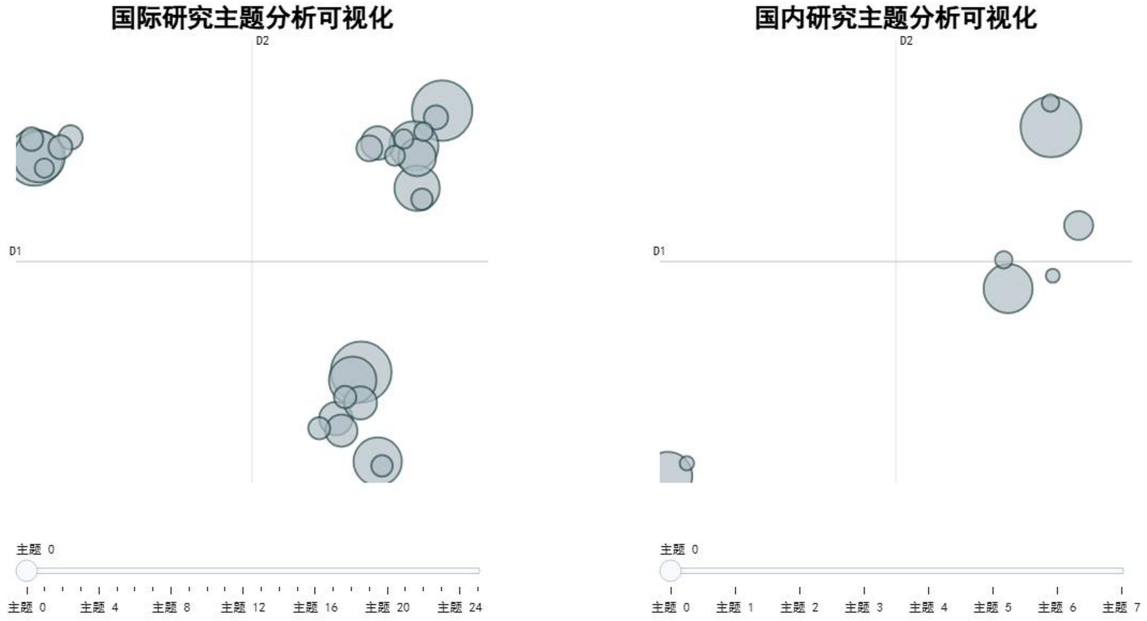


**图 7　国际国内研究主题可视化图谱对比**

(2)研究重心呈现显著的区域分化特征

国际研究形成“技术—文化—政策”三维驱动格局，文化遗产的数字保存成为全球关注的核心议题。相关研究不仅聚焦于数字保存的技术路径与实施策略，还深入探讨了其面临的挑战及应对机制，涵盖具体案例分析与跨领域协同实践。学术界高度关注数字技术在文化遗产长期保存与传承中的作用，强调在信息时代构建可持续的保存体系。此外，数据长期保存也是一大重点，其广泛性与深度也在不断增长，涵盖了数据格式的标准化、元数据的规范化等技术层面的研究内容。

国内研究则呈现出“档案中心化”特征。从研究内容来看，国内从数字保存研究起步阶段就开始关注如何实现档案的长期保存与可靠利用，包括数字档案的政策、安全治理、可信存储与验证等方面的探索。对于在国际研究中广受关注的文化遗产和个人信息的数字保存，在国内尚未形成独立的研究主题簇。尽管有部分关于文化遗产的数字保存研究，但并未形成系统、长期深入的研究趋势。个人信息的数字保存，也未能成为国内学术界的重点话题，尤其是随着信息技术的发展与个人隐私问题的日益严峻，数字化时代个人信息的保存与安全，值得更多的学者关注和探索。

(3)时序演化分析进一步揭示发展轨迹差异

从年度发文趋势可以看出，国内外研究数量的变化趋势基本一致，但国内研究呈现出整体趋势滞后约一年的显著特征。这一定程度上表明，国内数字保存研究还处于跟随国际研究热点与趋势的阶段，尚未具备引领地位。

此外，主题聚类结果显示，国际研究与国内研究在知识结构上存在显著差异。国际研究呈现出“高内聚、低耦合”的典型簇状结构，即同一主题簇内部关联紧密、研究方向集中，而不同主题簇之间界限分明、独立性高。相比之下，国内研究的主题分布则较为松散，主题内部关联强度有限，且未能形成界限清晰的簇状结构，反映出研究体系化程度不足，存在潜在的研究空白。这一对比说明，国内研究在主题深化与跨领域协同方面尚有较大提升空间，值得进一步引导与整合。

### 5.2　战略建议

基于上述对国际前沿趋势与国内现状差距的分析，为推动我国数字保存研究与实践向智能化、生态化与安全可信方向转型，本研究从以下四个维度提出未来发展建议。

(1)融入信息安全机制，构建可信数字保存体系

当前，数字保存领域已形成较为成熟和完整的学科体系与实践范式，但其安全范式仍主要停留在应对技术退化和意外损失的层面。本文认为，未来的数字保存研究必须融入现代信息安全的前沿思想。传统的访问控制策略无法应对更长时间跨度的

组织与政策变化，校验码机制难以抵抗恶意篡改攻击，已有的可信数字仓库模型对内部网络和人员的信任容易引发可用性灾难。未来的研究应构建自我演进的动态安全框架，探索策略驱动的动态访问控制模型和必要的加密敏捷机制，借鉴分布式账本技术的不可篡改特性建立可追溯可验证的溯源链，重塑可信数字仓库的安全理念，引入“零信任架构”等机制，从被动应对技术迭代转向主动防御潜在威胁，从源头保障数字对象的真实性、完整性与持久可用性。

(2)突破学科壁垒，构建应对数字时代的生态化保存范式

当前国内数字保存研究的“跟随式”范式导致了学科壁垒分明，难以有效应对数字时代动态、关联的文化遗产保存挑战。未来的研究必须打破图书馆学、档案学、计算机科学与文化遗产等领域的界限，转向构建一种生态化的保存范式。这要求学界不再满足于引介国外的静态归档模型，而是要通过跨学科协作，主动研究如何保存社交媒体事件流、网络集体记忆等具有本土复杂性的“数字事件生态”，将大量停留在数据采集阶段的实践真正提升至长期保存的战略高度。通过聚焦数字遗产的动态性与情境关联，突破孤立、静态的保存旧模式，形成真正能够应对数字时代复杂挑战的、具有中国特色的自主研究路径。

(3)拥抱人工智能，实现智能自治的保存范式

人工智能的崛起为应对海量数字遗产的挑战提供了革命性工具，正推动数字保存从“人工干预”迈向“智能自治”。未来的保存工作应深度整合人工智能技术，利用 AI 实现大规模数字遗产的智能鉴定、自动描述与深度内容挖掘，从而提升整理效率并揭示人类难以企及的知识关联。更进一步，基于 AI 的预测性维护模型将使数字档案馆、数字图书馆等具备“自修复”能力，主动预警格式过时与介质失效风险，并自动触发保存动作。同时，随着 AI 生成内容(AIGC)成为新型文化遗产，保存的边界应扩展至 AI 模型、关键参数与训练数据等“数字基因”，确保这类新型遗产的真实性与可解释性。

(4) 保存沉浸式体验，迎接元宇宙时代的范式革命

元宇宙将文化遗产的呈现与互动方式从“客体观察”提升至“场域体验”，将引发数字保存的范式革命。未来的保存目标应不再是孤立的 3D 模型或文件，而是用户在虚拟世界中完整的“沉浸式体验”，这包括空间环境、物理规则、社会交互等所有要素的整体性记录和保存。对于与物理世界同步演化的“数字孪生”文化遗产，保存的挑战在于如何归档其持续更新的生命周期数据。此外，元宇宙中原生的数字资产(如 NFT 艺术品)和社会事件也构成了全新的文化遗产，需要结合区块链等技术记录其所有权流转历史与社会语境，从而确保这些“原生虚拟遗产”作为文化记忆的真实性与持久价值。

## 6 结语

本文通过引入 BERTopic 动态主题建模技术，对国际与国内数字保存研究进行了系统梳理与多维比较，揭示出当前研究在主题广度、重点内容与发展路径上的显著差异。这种差异不仅体现在研究主题的丰富度与聚合程度，更体现在数字保存理念的成熟度与创新性上。通过对比分析可见，国际研究在技术、文化、政策三维协同推进下，已逐步构建起一套较为成熟的理论体系与实践框架，而我国研究虽已具备一定规模，但在研究视野、主题协同及自主创新方面仍有较大发展空间。未来，随着数字化浪潮的持续深化，我国数字保存领域的研究亟需从数量增长向质量跃升转变，从跟随式模仿走向自主式创新。

## 参考文献

1 Lavoie B, OCLC Research. The Open Archival Information System (OAIS) reference model: introductory guide (2nd edition) [EB/OL]. [2025－03－14]. https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file.

2 Caplan P, Guenther R. Practical preservation: the PREMIS experience[J]. Library Trends, 2005,54(1):111－124.

3 International Organization for Standardization. Space data and information transfer systems—audit and certification of trustworthy digital repositories: ISO 16363:2025[S]. Geneva: ISO, 2025.

4 Coalition Digital Preservation. Digital preservation handbook, 2nd edition[EB/OL]. [2025－03－14]. https://www.dpconline.org/handbook.

5 粟慧. 网络信息四大棘手问题研究近况[J]. 情报资料工作, 1999(6):14－17.

6 安艳杰. 国外数字保存研究的相关进展[J]. 情报杂志, 2004(2):127－129.

7 宛玲. 国外数字资源长期保存的最新发展及对我国的启示[J]. 中国图书馆学报，2004(2)：24－28.
8 孔志军. 国外数字资源长期保存活动进展[J]. 数字图书馆论坛，2006(4)：24－27.
9 郭家义，吴振新. 数字资源长期保存研究综述(1)——技术、系统、框架[J]. 图书馆杂志，2005(5)：53－58.
10 宛玲，吴振新，郭家义. 数字资源长期战略保存的管理与技术策略——中欧数字资源长期保存国际研讨会综述[J]. 现代图书情报技术，2005(1)：56－60.
11 梁妍. 近年来我国数字资源长期保存研究论文定量分析[J]. 图书馆工作与研究，2009(11)：47－50.
12 董光彩. 我国数字信息保存研究论文的计量分析[J]. 数字图书馆论坛，2006(5)：60－64.
13 聂云霞，丁家友. 复杂网络视域下数字资源长期保存研究脉络探析[J]. 信息资源管理学报，2014，4(3)：45－53.
14 冉从敬，陈一，李莎. 基于知识图谱的国外数字资源长期保存可视化研究[J]. 信息资源管理学报，2014，4(2)：106－113.
15 胡泽文，武夷山，孙建军. 数字资源保存的研究进展、热点与前沿[J]. 数字图书馆论坛，2013(2)：24－38.
16 蔡舜. 数字资源长期保存的知识图谱分析[J]. 图书馆工作与研究，2013(3)：48－51.
17 Krebs N，Borghoff U M. State-of-the-art survey of long-term archiving—strategies in the context of Geo-Data/cartographic heritage[M]//Jobst M. Preservation in digital cartography：archiving aspects. Berlin，Heidelberg：Springer Berlin Heidelberg，2010：101－127.
18 马仁杰，路思. 基于文献计量的国内 OAIS 理论研究现状分析[J]. 现代情报，2014，34(10)：50－56.
19 吴振新. 数字资源长期保存可信赖认证研究发展综述[J]. 中国图书馆学报，2015，41(3)：114－126.
20 王少辉. 国外数字信息长期保存中的知识产权问题研究进展[J]. 图书情报知识，2005(5)：95－98.
21 王磊，羊照生. 国外图书情报学界版权保护研究进展及启示[J]. 图书馆理论与实践，2009(2)：83－88.
22 肖秋会，许晓彤，赵明一. 欧美数字保存成本模型研究综述[J]. 图书馆学研究，2017(24)：2－9.
23 王爱霞，王鸿信，陶莉丽. 1998－2014 年我国数字信息资源呈缴制度研究综述[J]. 图书馆工作与研究，2015(4)：21－25.
24 Muir A. Legal deposit and preservation of digital publications：a review of research and development activity[J]. Journal of Documentation，2001，57(5)：652－682.
25 Ahmad R，Rafiq M. Global perspective on digital preservation policy：a systematic review[J]. Journal of Librarianship and Information Science，2023，55(3)：859－867.
26 华东杰. 英国图书馆数字保存战略研究[J]. 图书馆学研究，2013(19)：97－101.
27 杨道玲. Web 资源保存的热点问题管窥[J]. 图书情报工作，2005(3)：91－94.
28 宋奕宁，向禹. 电子文件长期保存研究综述[J]. 山西档案，2020(5)：169－175.
29 吴振新，陈瑶，李文燕，等. 国际 Data Curation 研究与实践发展综述[J]. 图书馆理论与实践，2016(2)：23－28.
30 黄新荣，高晨翔. 过程视角下的社交媒体存档技术研究述评[J]. 图书馆学研究，2019(2)：2－11.
31 崔旭，张若为，康璨琛. 国内外个人数字遗产保存研究综述[J]. 档案学研究，2022(6)：55－62.
32 Corrado E M，Sandy H M. Digital preservation of audiovisual-based materials：the state of the art[J]. Archiving Conference，2016，13(1)：161.
33 毛义春. 美国数字资源长期保存的研究进展及经验借鉴[J]. 北京档案，2009(7)：42－43.
34 高凡，吴振新，付鸿鹄，等. 数字资源长期保存：研究进展回顾与展望——iPRES 2019 国际会议综述[J]. 信息资源管理学报，2020，10(2)：118－127.
35 张柳，王慧，相甍甍. 基于 LDA 的突发事件应急管理主题热度与演化分析[J]. 情报科学，2023，41(6)：182－191.
36 邱均平，胡博，徐中阳，等. 基于 DTM 模型的国内外话语权研究主题挖掘及比较分析[J]. 情报理论与实践，2023，46(2)：24－34.
37 Abuzayed A，Al-Khalifa H. BERT for arabic topic modeling：an experimental study on BERTopic technique[J]. Procedia Computer Science，2021，189：191－194.

作者贡献说明：
张乃帅：确定论文选题框架，论文撰写及完善
王继民：数据集审校，论文修改与审核

作者单位：张乃帅、王继民，北京大学信息管理系，北京，100871
张乃帅，北京大学图书馆，北京，100871

收稿日期：2025 年 3 月 2 日
修回日期：2025 年 10 月 19 日
（责任编辑：支娟）

# A Thematic Review of Digital Preservation Research and Its Emerging Trends

ZHANG Naishuai　WANG Jimin

**Abstract**: Digital resources face systemic risks such as technological obsolescence, media degradation, and interpretability loss, necessitating effective digital preservation strategies. However, existing reviews often rely on qualitative descriptions or traditional bibliometrics, which fail to reveal deep semantic structures or systematically compare international and domestic research paradigms. To bridge this gap, this study utilizes a data-driven approach to reconstruct the knowledge genealogy of digital preservation and analyze evolutionary trends. Data were collected from the Web of Science, Scopus, and CNKI databases for the period 1991—2024. After rigorous screening, a total of 5214 valid papers were retained, including 3191 international and 2023 domestic publications. The study employed the BERTopic dynamic topic modeling framework and utilized deep learning-based pre-trained language models, including Bge-large-en-v1.5 and Paraphrase-Multilingual-MiniLM-L12-v2, to generate high-dimensional semantic embeddings of the abstracts. Through UMAP dimensionality reduction and HDBSCAN density clustering, fine-grained topics were extracted, followed by hierarchical clustering and trend visualization to compare research hotspots and structural evolution. The findings revealed significant structural differences between international and domestic research. International studies identified 26 distinct topics characterized by a "high-cohesion, low-coupling" structure and a comprehensive "technology-culture-policy" driving pattern. The core topics spanned multiple areas, including high-precision 3D reconstruction of cultural heritage, research data curation, and personal digital archiving. In contrast, domestic research identified only 8 topics, characterized by "Archive Centralization", focusing heavily on electronic records management and library resources. Domestic research showed lower thematic richness, looser internal associations, and a trend lag of approximately one year compared to international developments. Furthermore, while international studies shifted toward broader application expansion—such as work on social memory—domestic studies remained focused on foundational system architecture and policy development. Based on these findings, the study concludes that China's digital preservation field needs to transcend current disciplinary barriers. Future research should focus on integrating modern information security mechanisms like Zero Trust, constructing an ecological preservation paradigm that encompasses dynamic digital events, embracing Artificial Intelligence for intelligent autonomous preservation, and developing novel strategies to preserve immersive experiences in the emerging Metaverse era to move from "following" to "innovating".

**Keywords**: Digital Preservation; Long-term Preservation; BERTopic; Topic Modeling