清華大學 Tsinghua University

Building and Connecting Institutional Repositories Based on Scholarly Output and Data Aggregation: Possibilities and Challenges

基于论文和数据汇交的大学知识库建设及其互联互通：可能性与挑战

Jianbin JIN， Tsinghua University Library

金兼斌，清华大学图书馆

July 14-15, Hong Kong, China

清華大學 Tsinghua University

01 Institutional Repositories: A Research Infrastructure Amid De-globalization
机构知识库：逆全球化变局的科研支撑平台

02 University Institutional Repository Development: Challenges and Breakthroughs
大学机构知识库建设：困难与突破

03 Enabling Broader Data Interoperability in the Age of Artificial Intelligence
在人工智能时代促进更广泛的数据互通

Institutional repositories

are digital storage platforms developed and maintained by universities and research institutions to collect and preserve the scholarly outputs of their affiliated researchers.

机构知识库是由高校、科研机构等主导建设的数字存储平台，用来收集、保存本机构的学术资源

serve as a key infrastructure in Green Open Access (Green OA) practices, enabling researchers to self-archive their scientific publications and disseminate them to the public.

机构知识库是绿色开放获取实践中重要的基础设施，用于帮助研究者实现科研文章的自存储及面向公众传播

Scholarly Outputs Archived 存储的学术成果	**Theses and dissertations, Journal articles, Conference papers, Books and monographs, Research data...** 学位论文、期刊论文、会议论文、书籍著作、研究数据等
Key Objective 建立目的	**Promoting open access, Increasing research impact, Preserving research outputs, Supporting research assessment** 促进开放获取、提高研究影响力、保存机构重要科研成果、服务科研评估工作

National-level repositories 国家级知识库

Mainland China 中国内地

国家科研论文和科技信息高端交流平台
STIC National Advanced Platform for S&T Information and Communication

- National Advanced Platform for S&T Information and Communication 国家科技论文和科技信息高端交流平台

National Science Data Discovery Platform
国家科学数据发现平台

- ✓ National open platforms for research data
  国家级科研数据分享平台
- ✓ Promotes data reuse by following FAIR principles
  遵循FAIR原则，促进数据的重复使用

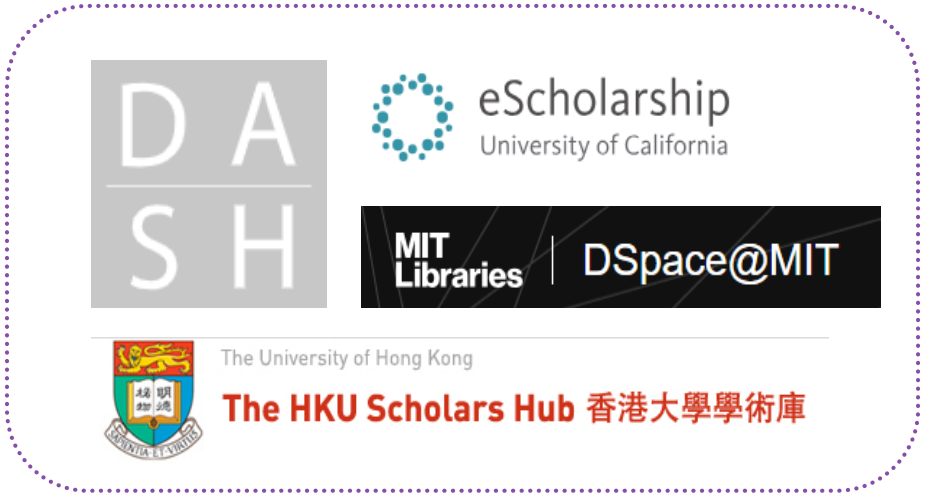
Knowledge Repository
国家知识仓储

- ✓ Stores knowledge outputs from publicly funded research
  存储由公共资金支持的科研产出
- ✓ Hosts 752,175 research projects and 976,449 full-text publications
  目前平台共汇集逾75万研究项目，逾97万全文论文

National-level repositories in other countries 其他国家的国家级知识库

HAL — France
FRDR — Canada
Plataforma Lattes | CNPq — Brazil
PMC PubMed Central® — United States
SwePub — Sweden

University-level Institutional Repositories 大学机构知识库



Representative University Institutional Repositories

代表性大学机构知识库案例

- DSpace (MIT)
- eScholarship (University of California)
- DASH (Harvard University)
- The HKU Scholar HUB (The University of Hong Kong)

Mainland China 中国内地

University Institutional Repository development remains at an early stage.

中国大学机构知识库建设仍在发展初期

Currently, most university repositories serve only as publication directories and do not store the full-text papers

当前许多大学的"知识库" 仅提供论文导航功能，并不存储全文

Preprint repositories 预印本平台

Preprint platforms can also be considered a type of institutional repository, particularly when operated by universities or research institutions. They serve to preserve and disseminate early-stage research outputs before formal peer review.

预印本平台也可以视为一种机构知识库，尤其是在其由大学或科研机构主办的情况下。它们用于保存和传播尚未正式发表的早期研究成果，在同行评审之前实现快速公开。

Key features of preprint repositories include:
预印本知识库的主要特点包括：

- **Storage and long-term preservation of unpublished manuscripts**
  存储和长期保存尚未出版的研究手稿
- **Free and open access to early research findings**
  向公众免费开放获取早期研究成果
- **Support for rapid dissemination and scholarly feedback**
  支持快速传播与学术反馈

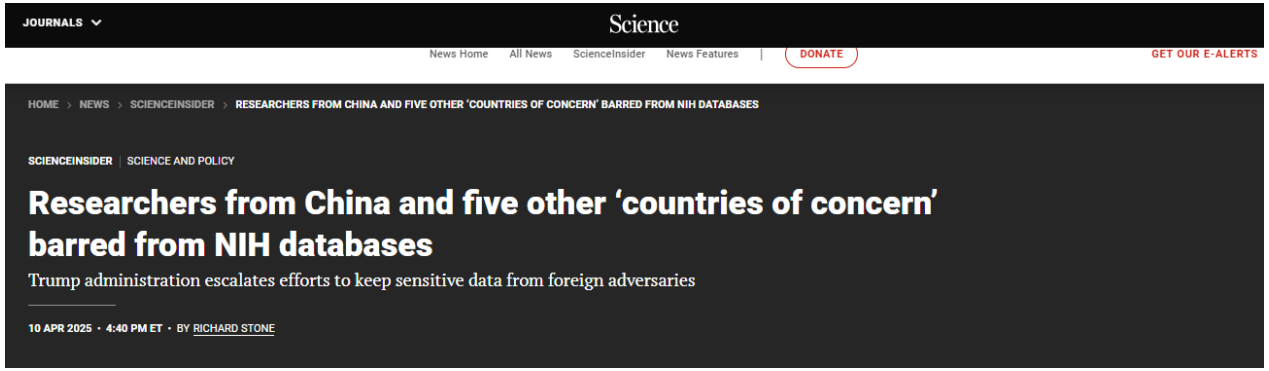
Examples of Preprint Platforms
中外预印本平台案例

ChinaXiv
BiomedRxiv 生物医学科技论文预印本系统
arXiv
bioRxiv THE PREPRINT SERVER FOR BIOLOGY
medRxiv THE PREPRINT SERVER FOR HEALTH SCIENCES
ChemRxiv®

De-globalization and Scholarly Communication 逆全球化与学术交流

Geopolitical tensions and technological decoupling are disrupting global research collaboration
全球政治与科技脱钩趋势增强，国际科研合作受限

A series of measures led by the United States — including technology blockades, export controls, and restrictions on database access — have posed significant risks to the accessibility of academic resources.
美国主导的一系列举措，包括技术封锁、出口管制以及数据库访问限制，对学术资源的可访问性构成了重大风险


JOURNALS Science
News Home All News ScienceInsider News Features DONATE GET OUR E-ALERTS
HOME › NEWS › SCIENCEINSIDER › RESEARCHERS FROM CHINA AND FIVE OTHER 'COUNTRIES OF CONCERN' BARRED FROM NIH DATABASES
SCIENCEINSIDER | SCIENCE AND POLICY
Researchers from China and five other 'countries of concern' barred from NIH databases
Trump administration escalates efforts to keep sensitive data from foreign adversaries
10 APR 2025 · 4:40 PM ET · BY RICHARD STONE

In April 2025, the NIH banned research institutions from China and other "countries of concern" from accessing its related data
2025年4月，NIH禁止中国等“受关注国”科研机构访问其相关数据

Strategic Importance of Institutional Repositories 机构数据库的战略重要性

- Reduces dependence on commercial platforms and enhances national visibility
  减少对商业平台依赖，提升国家科研可见度
- Ensures long-term accessibility of publicly funded research outputs and safeguards research outputs as national knowledge assets
  保障国家资助科研成果的长期可访问性，保护作为国家知识资产的科研成果
- Supports self-owned infrastructure for knowledge dissemination
  建设自主可控的科研传播基础设施
- Supports national Science, Technology & Innovation evaluation and planning systems
  服务国家科研评价与科技规划

01 Institutional Repositories: A Research Infrastructure Amid De-globalization
机构知识库：逆全球化变局的科研支撑平台

02 University Institutional Repository Development: Challenges and Breakthroughs
大学机构知识库建设：困难与突破

03 Enabling Broader Data Interoperability in the Age of Artificial Intelligence
在人工智能时代促进更广泛的数据互通

Directory-style Institutional Repositories
导航式“机构知识库”

Some university institutional repositories store only metadata of publications without archiving the full-text content. They provide external links that redirect users to the publisher’s website for access.

大学的“机构知识库”只存储文章的元数据，不存储全文信息，在“机构知识库”中提供DOI导向出版商网站

Drawbacks of Directory-style IRs
导航式“机构知识库”的弊端

- No long-term preservation of research outputs 无法真正长期保存科研成果
- The access dependence on external publisher platforms’ policies 获取科研成果依赖于外部出版平台的政策
- Incompatible with research assessment and data governance needs 难以对接科研评估和数据治理

Lack of Mandatory Policies: Data Cannot Be Reliably Collected
缺乏制度保障，数据难以汇集

Researchers are not required to deposit their publications into the institutional repository
研究人员没有被要求将论文上传至机构知识库
Results in incomplete coverage and fragmented research outputs
导致知识库数据不足，研究人员的成果覆盖率低且分散

Unclear Copyright Rules: Institutions Are Unsure What They Can Host
缺乏版权知识，机构不知道可以持有何种数据

The administrators of university institutional repository often misunderstand publisher copyright policies
机构知识库的管理员对出版商的版权政策存在误解
Uncertain about which versions (e.g., accepted manuscripts) are legally depositable
机构人员不知道何种版本可以被合法存储

Limited Resources: No dedicated team and Financial Support
缺乏资源投入，没有专职团队和财政支持

Repository tasks are often handled as side duties, leading to limited expertise and low priority
机构知识库相关工作通常被当作附带任务来处理，导致专业性不足、优先级较低
Without funding, repositories become symbolic systems — online but inactive
没有资金，知识库就变成了一个符号性的系统：虽上线但不可运作

Development of Open Science Policy Frameworks
开放科学制度建设

Mandatory Deposit 强制提交规定

Institutions establish the open access policies and request that the obligated researchers deposit the accepted author manuscript (AAM) of the research paper into the university institutional repository after publication

机构要求相关研究人员在论文发表后，将作者接受稿存入大学的机构知识库

Grant a non-exclusive license to institutions
授予非排他性出版许可

Researchers were requested to grant the institution a non-exclusive license for open dissemination

研究者被要求授予机构非排他性出版协议以开放传播

Each Faculty member grants to the Massachusetts Institute of Technology **nonexclusive permission** to make available his or her scholarly articles and to exercise the copyright in those articles for the purpose of open dissemination. In legal terms, each Faculty member grants to MIT a **nonexclusive, irrevocable, paid-up, worldwide license** to exercise any and all rights under copyright relating to each of his or her scholarly articles, in any medium, provided that the articles are not sold for a profit, and to authorize others to do the same. **The policy will apply to all scholarly articles written while the person is a member of the Faculty** except for any articles completed before the adoption of this policy and any articles for which the Faculty member entered into an incompatible licensing or assignment agreement before the adoption of this policy.

MIT Faculty Open Access Policy

Development of Open Science Policy Frameworks
开放科学制度建设

Mandatory Deposit 强制提交规定

Institutions establish the open access policies and request that the obligated researchers deposit the accepted author manuscript (AAM) of the research paper into the university institutional repository after publication

机构要求相关研究人员在论文发表后，将作者接受稿存入大学的机构知识库

Grant a non-exclusive license to institutions
授予非排他性出版许可

Researchers were requested to grant the institution a non-exclusive license for open dissemination

研究者被要求授予机构非排他性出版协议以开放传播

Each Faculty member grants to the Massachusetts Institute of Technology **nonexclusive permission** to make available his or her scholarly articles and to exercise the copyright in those articles for the purpose of open dissemination. In legal terms, each Faculty member grants to MIT a **nonexclusive, irrevocable, paid-up, worldwide license** to exercise any and all rights under copyright relating to each of his or her scholarly articles, in any medium, provided that the articles are not sold for a profit, and to authorize others to do the same. **The policy will apply to all scholarly articles written while the person is a member of the Faculty** except for any articles completed before the adoption of this policy and any articles for which the Faculty member entered into an incompatible licensing or assignment agreement before the adoption of this policy.

MIT Faculty Open Access Policy

Strategic Management at the University Level
大学层面的战略管理

Faculty-Led Policy Adoption 教师主导政策

Faculty adoption ensures that the policy reflects their own interests, not administrative imposition. Since the policy affects faculty the most and involves the grant of non-exclusive rights, it must be grounded in informed faculty consent
由教师主导采纳政策有助于确保该政策源于教师群体的真实意愿，而非行政强加。由于政策涉及将非排他性权利授予机构时，必须建立在教师知情同意的基础上。

Leadership-led Coordination 大学主管协调

The institutional repository should be regarded as an integral part of the university's strategic development, led by university leadership, managed and operated by a professional library team, and supported with appropriate policies and funding
将机构图书馆作为大学战略发展的重要部分，由大学主管主导协调、图书馆专业团队负责管理执行，并给予政策和资金支持

MIT	**Leader: Provost → Implementing Units: University Library System** 主管：教务长 → 执行机构：图书馆系统
Harvard	**Leader: Provost /Dean → Implementing Units: University Library** 主管：教务长或各学院院长→ 执行机构：大学图书馆
HKU	**Leader: Liberian→ Implementing Units: University Library** 主管：图书馆馆长→ 执行机构：大学图书馆

Copyright Compliance
版权合规

Waive & Dark Deposit 豁免和暗存储

If a researcher has signed an exclusive copyright agreement with a publisher and is unwilling or unable to grant the institution a non-exclusive license, they may obtain a waiver from the policy
研究者如已签署排他性版权协议，无法授予机构非排他性授权时，可获得豁免

The waiver only exempts the paper from Green Open Access requirements. Researchers are still required to deposit the paper into the university's institutional repository after publication
豁免仅适用于论文的开放获取，研究者仍需要在出版后提交论文到大学机构知识库进行存储

Waived articles are stored in the institutional repository under a dark deposit model: only the metadata is publicly visible, while the full text is retained in the repository without public access. Depending on the terms of the copyright agreement, the waived article may be made openly accessible (Green Open Access) after an embargo period
豁免的文章在机构知识库中采用暗存储模式：仅对外展示文章的元数据信息，全文则保存在机构知识库中但不公开访问。根据版权协议的具体条款，豁免论文可在禁运期结束后转为绿色开放获取。

Copyright Compliance
版权合规

Embargo & Publishers' Green OA Policy
时滞期和出版机构的绿色开放获取政策

In response to the growing trend of open science, publishers have established policies to support the development of Green OA. The use of an embargo period is a common feature of such policies.

在开放科学的发展趋势下，出版机构制定了支持绿色开放获取发展的政策，“时滞期”规定是这些政策的特征。

	Elsevier	Springer Nature	Wiley	Taylor & Francis
Release Timing 开放时机	Open access is permitted after an embargo period, in non-profit institutional repositories. 时滞期结束后，文章可以在非营利性机构知识库中以开放获取方式发布			
Depositable Version 存储版本	*Author Accepted Manuscript* 作者接受稿			
Embargo 时滞期	*12-36 months* Depending on the journal	*6-12 months* Depending on the Portfolio	*STM: 12 months HSS: 24 months*	*STM: 12 months HSS: 18 months*
Other Requirements 其他要求	• Provide a link to the published version 链接到原文 • Use the CC BY-NC-ND license 使用指定CC协议	• Provide a link to the published version 链接到原文 • Use the CC BY-NC-ND license 使用指定CC协议	• Include a notice on the first page 首页包含特定声明	• Include a notice on the first page 首页包含特定声明 • Recommend using CC BY-NC or CC-BY-NC-ND License 推荐使用CC协议

The Green OA Policies of Leading Academic Publishers
主要出版机构的绿色OA政策

During the embargo period, the publisher retains exclusive rights to distribute the article. After the embargo expires, authors are allowed to share the AAM via non-commercial institutional repositories under certain conditions.
在时滞期内，出版方保留文章的独家传播权。在时滞期结束后，作者可以在满足一定条件的前提下，将作者接受稿通过非商业机构知识库进行传播。

2.4 How a University Institutional Repository Works
大学机构知识库的运转模型

清華大學 Tsinghua University

Self-archiving by researchers
研究人员自存储

Researchers provide a deposit after publication
研究者成果发表后提供存档

学术产出
Scholarly Outputs

Researchers
研究人员

Institutional repository
机构存储库

Publicly deposit
公开存档

OPEN ACCESS Green OA

Internal deposit
内部存档

Deposit for institutional internal use
(e.g., academic evaluation)
机构内部用途
（如科研评估使用）

Dark deposit
暗存档

Only metadata is publicly available
Full content is stored but not accessible
仅公开文章元数据
全文仅作存储，不对外公开

Classified storage
分类存储

Classify storage based on purpose and copyright status.
按照用途和版权分类存储

Full-text documents storage
全文存储

Store the full-text version of the accepted manuscript
存储作者接受稿全文

01 Institutional Repositories: A Research Infrastructure Amid De-globalization
机构知识库：逆全球化变局的科研支撑平台

02 University Institutional Repository Development: Challenges and Breakthroughs
大学机构知识库建设：困难与突破

03 Enabling Broader Data Interoperability in the Age of Artificial Intelligence
在人工智能时代促进更广泛的数据互通

AI is reshaping research paradigms, demanding high-quality and structured research data
人工智能正在重塑科研范式，对高质量、结构化的科研数据提出更高要求

Data-driven innovation is becoming the engine of advancement across disciplines
数据驱动创新成为各学科发展的关键动力

Disrupt the monopoly over research data 打破科研数据垄断

Scientific outputs and research data are essential resources. However, they are predominantly held by publishers and sold at high costs for AI training, thereby hindering the advancement of domain-specific AI development.
科研成果和数据是重要的资源，但目前多数被商业出版社把持，以高价出售用于人工智能训练，限制了学科型人工智能的发展

Institutional repositories serve as centralized platforms for collecting research data and outputs, helping counter the data dominance of commercial publishers and fostering open science in the AI era.
机构知识库可作为数据集与科研成果的汇聚平台，帮助打破商业出版社的数据垄断，促进人工智能时代的开放科学发展

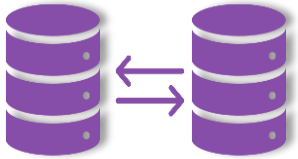
NEWS | 09 December 2024

Publishers are selling papers to train AIs – and making millions of dollars

Generative-AI models require massive amounts of data – scholarly publishers are licensing their content to train them.

"Several big publishers have cashed in on AI licensing deals this year. In May, Informa, the parent company of the UK academic publisher Taylor & Francis, announced that it made a US$10-million deal to license content to Microsoft. The next month, the US academic publisher Wiley announced to its investors that it had earned $23 million from a deal with an unnamed firm developing generative-AI models. In September, the company said that it expected to earn another $21 million from such agreements this financial year."

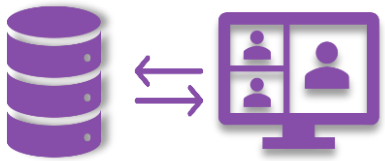
Source: Nature 636, 529-530 (2024)
https://www.nature.com/articles/d41586-024-04018-5



Interoperability among institutional repositories
机构存储库间的互联互通

- Under the protection of intellectual property rights, institutional interoperability should be promoted through policy frameworks and collaborative agreements
  在保障版权的情况下，通过制度保障和合作协议促进各机构存储库间协作互通
- Develop standardized platforms and data protocols to ensure smooth data exchange and cross-system interaction

  建立标准化平台和数据规范，保障数据交流和跨系统交互



Interoperability among institutional repositories
机构存储库间与出版平台的互联互通

- Establish collaboration agreements with publishers, data platforms, and preprint servers to expand the channels for collecting researchers' manuscripts into institutional repositories, thereby ensuring broader coverage of Green Open Access.

  与各出版商、数据平台及预印本平台签署合作协议，拓展获取机构研究者手稿的渠道，实现更广范围的存储入库，推动绿色开放获取覆盖面的持续扩大。

Advancing the FAIR principles: Findable, Accessible, Interoperable, and Reusable
最终目标：促进科研成果和科研数据的FAIR原则

- Increase investment in open science infrastructure and develop next-generation repositories with structured data and open access interfaces.
  增加设施投入，建设支持结构化数据、开放接口的下一代开放科学基础设施
- Enhance multi-stakeholder collaboration with publishers, platforms, and other partners, improve the quality and efficiency of content aggregation, reduce copyright and access barriers, and strengthen system interoperability
  强化多利益相关方合作，提升内容汇聚的质量与效率，降低版权与访问壁垒，强化系统的互操作性
- Explore the integration of institutional repositories with artificial intelligence and leverage institutional repositories for data governance and as training sources for AI models
  探索人工智能与机构知识库的整合，发挥机构知识库在科研数据治理与AI训练数据支撑中的作用

清華大學
Tsinghua University

Q&A

金兼斌

Email: jinjb@tsinghua.edu.cn