第七届中美高校图书馆合作发展论坛

开放数据新常态下增强科学数据服务体系的探讨

Discussion on Enhancing Scientific Data Service System under the New Normal of Open Data

CASHL管理中心/北京大学图书馆　吴亚平

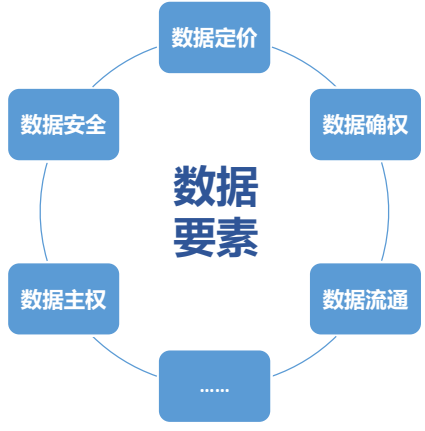
2023年7月28日

目录

OUTLINE

数据文明下的数据要素

- 大量数据的产生、聚集和要素化是数字文明的主要特征之一 。
- 数据与土地、劳动力、资本、技术等传统要素并列，成为新的生产要素。

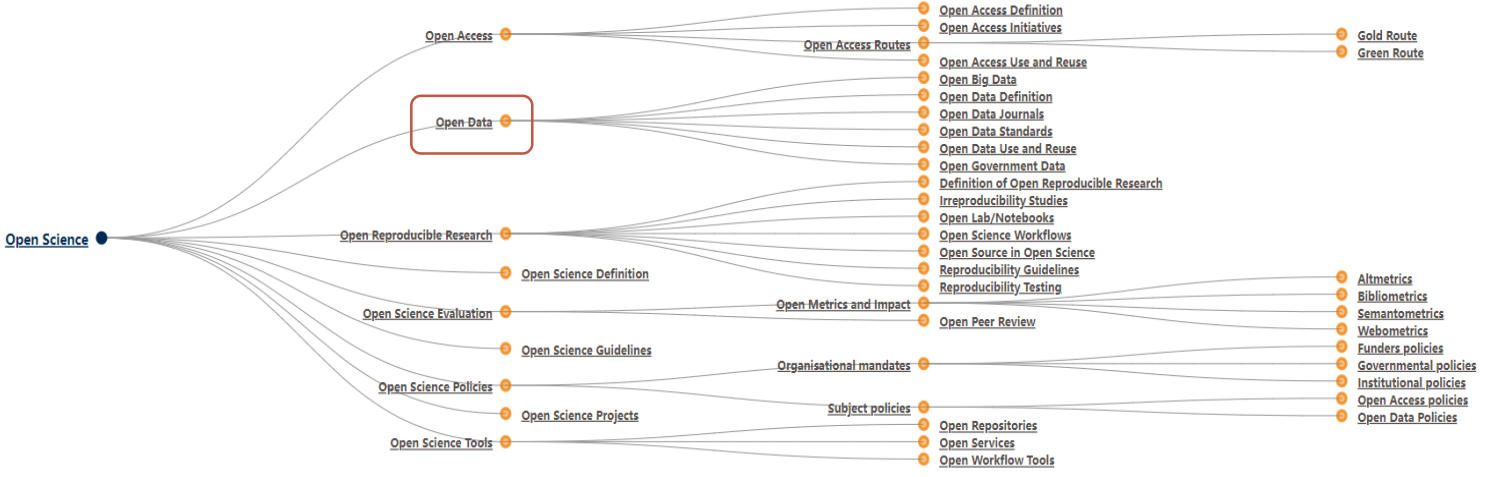


- 开放数据是促进数据要素价值释放的前提和重要路径。

- 开放数据（open data）是指数据可以被任何人自由**免费地访问、获取、利用和分享**。——世界银行
- 开放数据是任何人都可以在不受技术或法律限制的情况下访问和**可再用（Reuse）**的数据，使用者通常无需承担任何费用。——经济合作与发展组织。
- **具备必要的技术和法律特性**，从而能被任何人、在任何时间和任何地点进行自由利用、再利用和分发的电子数据。 ——《开放数据宪章》

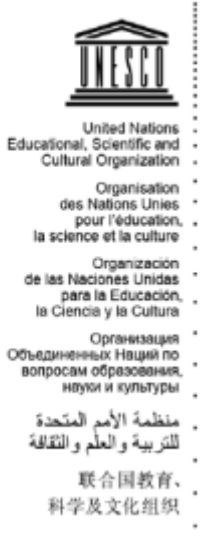
开放数据是开放科学体系的重要分支

- 科学界中的开放倡议、开放运动由来已久，目前已发展、成长为开放科学体系，旨在使科学研究更加开放、可及、有效、民主和透明，主要研究主题分类如下1：



1：foster开放科学学习门户：开放科学https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition

开放数据是开放科学体系的重要分支

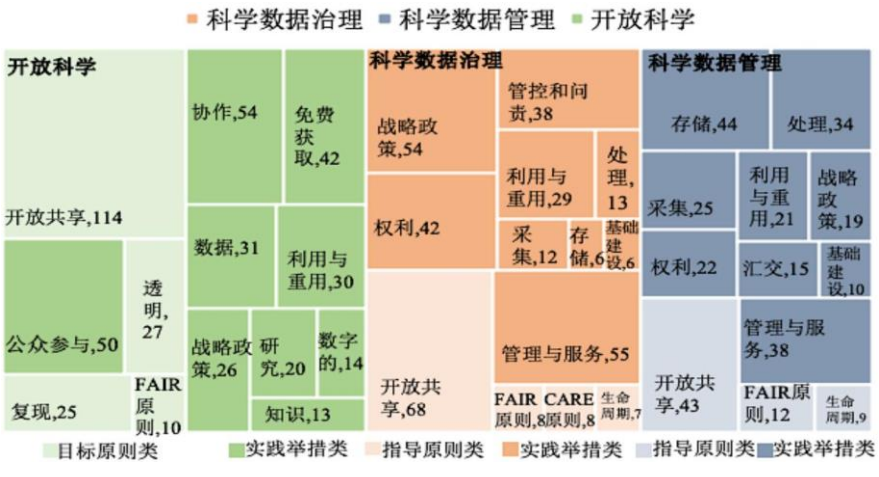


- 2021年，联合国教科文组织（UNESCO）审议通过《开放科学建议书》[1]开放科学迈入全球共识新阶段。
  - 指出开放科学涵盖所有学科以及科学研究的各个方面，包括开放科学知识、开放科学基础设施、科学传播、开放式参与以及与其他知识体系的开放式对话等。
  - 其中，开放科学知识包括科学出版物、研究数据、元数据、开放式教育资源、软件、源代码及硬件等 。

1：https://unesdoc.unesco.org/ark:/48223/pf0000376893/PDF/376893eng.pdf.multi

开放科学数据的多重价值

概念内涵不断发展1



1.司莉,刘先瑞.面向开放科学的科学数据治理：概念、框架与展望[J/OL].情报科学:1-13[2023-07-21].http://kns.cnki.net /kcms/detail/ 22.1264.G2.20230516.1458.010.html

多重价值达成共识，认同数据应该是开放的、可访问的和可重用的

- 重现、验证研究成果
- 促进公众科学
- 促进跨学科合作共享与再研究
- 支撑政府决策与社会问题的解决
- 加速科技创新与发展
- 促进智能化与人工智能

多方促进科学数据的开放

《开放科学建议书》

- 确保科学相关信息、数据、源代码和硬件等研究产出得到长期保存、共享和重复使用
- 遵循数据管理FAIR（可查找、可获取、可互操作和可重复使用）原则和CARE（集体利益、控制权、责任和道德操守）原则

多方促进科学数据的开放

- 2018年国务院办公厅发布《科学数据管理办法》要求“政府预算资金资助形成的科学数据应当按照开放为常态、不开放为例外的原则，接入国家数据共享交换平台，面向社会和相关部门开放共享 。
- 欧盟的开放科学政策（2020-2024）也指出FAIR原则和开放数据共享应该成为欧盟资助的科学研究结果的默认标准 。
- 2022年8月美国白宫科技政策办公室（OSTP）发布了最新的政策指南，要求到2026年，联邦资助的研究产生的出版物及其支撑数据可以立即开放获取，没有时滞期 。
- 美国心脏协会要求独立验证所需的数据 在资助结束后 12 个月内公开。

多方促进科学数据的开放

- 提供研究数据的保存、管理与发布服务，促进研究数据的传播、再利用和规范引用。

- 哈佛大学开放存取科研数据平台Harvard Dataverse
- 北京大学开放研究数据平台
- 中国科学院ScienceDB:打造科学数据长期共享与出版的国际化通用存储库

数据出版是开放科学模式的典型特征

- 期刊文章/论文是科研成果的重要体现，也是开放运动关注的重点
- 在柏林宣言 布达佩斯开放获取倡议 PlanS 等运动下推出了绿色OA、金色OA、钻石OA等模式。
- 逐步细化到某个部分，包括2020年启动的开放摘要倡议（I4OA）和2017年启动开放引文倡议（I4OC）
- 当前支撑论文研究直接相关的数据也备受关注

数据出版是开放科学模式的典型特征之一[1]

模式	环节					
	发现	分析	创作	出版	扩展	评价
传统模式	权威文摘索引系统(如 WoS)	商用套装软件(如 SPSS 和 SAS 等)	文书软件和文献管理(如 MSWord 和 Endnote)	顶级期刊(如 *Nature* 和 *Science* 等)	作者唯一识别符(如 Research ID)	期刊影响因子
现代模式	搜索引擎(Google Scholar)	开源软件(如 R 和 Python)	云端办公系统(如 Google Docs 和 RefWorks 等)	预印本和完全开放获取期刊(如 ArXiv 和 PLoS)	机构知识库(如 ir. las. ac. cn)	期刊级别计量、论文级别计量(如 Eigenfactor 和 ALM 等)
开放科学模式	开放资源汇聚池(如 Paperity)	开放分析平台(如 ROpenSci Science Exchange Zooniverse)	云群组协作和共享权限(如 Hypothes. is 和 Google Drive 和 Zotero 等)	数据、代码、软件共享(如 F1000Res、PLoSONE 和、Figshare、Dryad 等)	唯一标识符和文献关联推荐(如 ORCID、Research Gate、SlideShare 等)	公众评估(如 Publons 和 ImpactStory 等)
创新模式	科研群组共享模式	科研众包(如 Zooniverse 和 Hivebench)	群体协作(如 Authorea 和 Colwiz)	课件共享与论文草稿注册(如 Figshare 和 PeerJ 等)	唯一标识符的开放接口运用(如 ResearchGate 和 ORCID 等)	学术与社会影响力评估(如 Peerage of Science 和 Altmetric)

1.顾立平.科研模式变革中的数据管理服务:实现开放获取、开放数据、开放科学的途径[J].中国图书馆学报,2018,44(06):43-58.

出版商推动开放数据形成新常态

□ 开放科学中心https://www.cos.io/

2015年《透明度和开放性促进指南（TOP）》包括八个模块化标准，每个标准都有三个严格程度不断提高的级别

	未实现	一级	二级	三级
引用标准	没有提到数据引用。	期刊通过明确的规则和示例向作者描述了指南中数据的引用。	文章为符合期刊作者指南的数据和材料提供了适当的引用。	在按照期刊的作者指南为数据和材料提供适当的引用之前，文章不会发表。
数据透明度	Journal 鼓励数据共享，或者什么都不说。	文章说明数据是否可用，如果可用，在哪里访问它们。	数据必须发布到受信任的存储库。必须在文章提交时确定例外情况。	数据必须发布到受信任的存储库，报告的分析将在发布前独立复制。
分析方法（代码）透明度	Journal 鼓励代码共享，或者什么都不说。	文章说明代码是否可用，如果可用，在哪里访问它。	代码必须发布到受信任的存储库。必须在文章提交时确定例外情况。	代码必须发布到受信任的存储库，报告的分析将在发布之前独立复制。
研究材料透明度	Journal 鼓励材料共享，或者什么都不说。	文章说明材料是否可用，如果可用，在哪里可以访问它们。	材料必须发布到受信任的存储库。必须在文章提交时确定例外情况。	材料必须发布到受信任的存储库，报告的分析将在发布前独立复制。
设计和分析透明度	Journal 鼓励设计和分析的透明度，或者什么都不说。	期刊阐明了设计透明度标准。	期刊要求遵守审查和出版的设计透明度标准。	期刊要求并强制遵守审查和出版的设计透明度标准。
研究预注册	日记什么也没说。	文章说明是否存在研究预注册，如果存在，在哪里可以访问它。	文章说明是否存在研究预注册，如果存在，允许在同行评审期间访问期刊以进行验证。	期刊需要预先注册研究，并在文章中提供链接和徽章以满足要求。
分析计划预注册	日记什么也没说。	文章说明是否存在研究预注册，如果存在，在哪里可以访问它。	文章说明是否存在分析计划预注册，如果存在，允许在同行评审期间访问期刊以进行验证。	期刊要求使用分析计划预先注册研究，并在文章中提供链接和徽章以满足要求。
复制	期刊不鼓励提交复制研究，或者什么都不说。	期刊鼓励提交复制研究。	期刊鼓励提交复制研究并对结果进行盲审。	在观察研究结果之前，期刊使用注册报告作为复制研究的提交选项，并进行同行评审。

□ 同行评审专家的倡议和承诺https://opennessinitiative.org

- 2017年，承诺作为审稿人，不对任何没有明确理由不公开数据的手稿进行全面审阅。

出版商推动开放数据形成新常态

- 数据引用原则联合声明和FORCE11 软件引用实施小组，倡议研究数据和软件是合法的、可引用的研究产品，其引用应与出版物引用具有相同的地位。
- 国际科学，技术和医学出版商协会（STM协会）将2020定为研究数据年，旨在增加实行数据政策、施行存储数据链接的期刊数量，增加了对数据集引用量的特别行动计划(Cambridge University Press, Elsevier, DeGruyter, Karger Publishers, Oxford University Press, Sage Publishing, Springer Nature, Taylor & Francis and Wiley均参与了该计划)

  Share 共享：快速增加实行数据政策和带有数据可用性声明（DAS）的文章和期刊数量；

  Link 链接：加快将数据链接存放到SCHOLIX框架以链接数据集和出版物的期刊数量；

  Cite 引用：加快对数据集的引用。

出版社推动开放数据形成新常态

总体来看

- 达成共识，认同数据的可重用、可发现、可解释和可引用等特性，致力于促进更快、更有效、更透明的研究发现。
- 差异化政策，不同出版社针对不同期刊和不同领域的数据有着不同的政策，兼顾学科差异和隐私保护。
- 努力推动新常态，开拓开放数据的新方法、支持工具，助力使数据共享成为新常态。

出版社开放数据政策举例

□ 数据可用性声明（Data Availability Statement，DAS）

确认共享数据的存在。应包括在哪里可以找到支持文章中报告的结果的数据的信息和公开存档数据集的超链接，以及在适当的情况下，不共享数据。

- 数据可用性声明模板(如需多个数据集，可采用多形式组合（Wiley）

数据的可用性	数据可用性声明的模板
在发布带有 DOI 的数据集的公共存储库中公开提供的数据	支持本研究结果的数据可在[存储库名称，例如"figshare"]中公开获得，网址为 http://doi.org/[doi]，参考编号[参考编号]。
在不发布数字对象标识符的公共存储库中公开提供的数据	支持本研究结果的数据可在[存储库名称]的[URL]，参考编号[参考编号]中公开获得。
源自公共领域资源的数据	支持本研究结果的数据可在 [URL/DOI] 的 [存储库名称] 中获取，参考编号 [参考编号]。这些数据派生自公共领域中可用的以下资源：[列出资源和 URL]
由于商业限制而禁止数据	支持研究结果的数据将在 [存储库名称] 的 [URL / DOI 链接] 中提供，自发布之日起禁运，以便将研究结果商业化。
由于隐私/道德限制，可根据要求提供数据	支持本研究结果的数据可向通讯作者索取。由于隐私或道德限制，数据不公开。
受第三方限制的数据	支持本研究结果的数据可从[第三方]获得。限制适用于这些数据的可用性，这些数据是在本研究的许可下使用的。数据可在[第三方]的许可下[从作者/网址]获得。
可应作者要求提供数据	支持本研究结果的数据可根据合理要求从通讯作者处获得。
数据共享不适用 – 未生成新数据	数据共享不适用于本文，因为本研究中没有创建或分析新数据。
作者选择不共享数据	研究数据不共享。
文章补充材料中提供的数据	支持本研究结果的数据可在本文的补充材料中找到
数据共享不适用 – 没有生成新数据，或者文章完全描述了理论研究	数据共享不适用于本文，因为在当前研究期间没有生成或分析数据集

出版社开放数据政策举例

□ 数据共享及其证据（Data has been shared）：

对数据可用性声明中的数据链接进行检查。如果数据在数据存储库中共享，则数据可用性声明包括到该数据的永久链接，和共享数据的引用信息。

□ 数据同行评审（Data has been peer reviewed）：

包括作者是否遵守了期刊关于研究数据可用性的政策，以及是否做出了合理的努力；对共享数据的质量和可复制性同行评审，评审论文结果和数据存储库中的数据来评审数据质量（如样本量大小和变量匹配），数据的可复制性等。

同行审稿人在需要对稿件进行评估时，有权要求访问底层数据(和代码)。(Springer Nature)

出版社开放数据政策举例

- **数据同行评审指南:**

数据可用性声明

是否提供了适当的DAS
读者如何访问数据是否清楚
在DAS中提供链接的地方，它们是否有效
在数据访问受到限制的情况下，是否有适当的访问控制
如果数据被描述为包含在手稿和/或补充信息文件中，是准确的吗

可用的数据文件

数据是否在最合适的存储库中
数据是否以严谨和方法合理的方式产生
数据和任何元数据是否符合研究团体的文件格式和报告标准
作者存放的数据文件是否完整，是否与稿件中的描述相符
它们是否包含个人身份、敏感或不适当的信息

出版社开放数据政策举例

口 开放数据政策（Wiley）

	数据可用性声明已发布 1	数据已共享 2	数据已经过同行评审 3	威利期刊示例
鼓励数据共享	自选	自选	自选	
期望数据共享	必填	自选	自选	英国社会心理学杂志
强制数据共享	必填	必填	自选	生态与进化
授权数据共享和同行评审数据	必填	必填	必填	地球科学数据杂志 美国政治学杂志

出版社开放数据政策举例

开放数据政策（Springer Nature）

Type	Policy summary	Example Journal
Type 1st	鼓励数据共享和提供引用 Data sharing and data citation is encouraged	Photosynthesis Research
Type 2nd	鼓励数据共享和数据共享的证据 Data sharing and evidence of data sharing encouraged	Plant and Soil
Type 3rd	鼓励数据共享，并要求提供数据可用性声明 Data sharing encouraged and statements of data availability required	Humanities and Social Science Communications
Type 4th	数据共享，数据共享的证据和所需数据的同行评审 Data sharing, evidence of data sharing and peer review of data required	Scientific Data

出版社开放数据政策举例

开放数据政策（Springer Nature）

对于生物科学领域下的数据集（部分），必须提交到社区认可的公共存储库

Mandatory deposition	Suitable repositories
Protein sequences	Uniprot
DNA and RNA sequences	Genbank DNA DataBank of Japan (DDBJ) EMBL Nucleotide Sequence Database (ENA)
DNA and RNA sequencing data	NCBI Trace Archive NCBI Sequence Read Archive (SRA)
Genetic polymorphisms	dbSNP dbVar European Variation Archive (EVA)
Linked genotype and phenotype data	dbGAP The European Genome-phenome Archive (EGA)

出版社开放数据政策举例

开放数据政策（OUP）

- 级别 1，期刊鼓励所有作者在道德上可行的情况下公开发布任何已发表论文的所有数据。
- 级别 2，期刊鼓励所有作者在道德上可能的情况下公开发布任何已发表论文的所有数据,必须包含数据可用性声明。
- 级别3，期刊要求所有作者，在道德上可能的情况下，公开发布任何已发表论文的所有数据，作为发表的条件,必须包含数据可用性声明。
- 级别4，期刊要求所有作者在道德上可能的情况下，公开发布任何已发表论文的所有数据，作为发表的条件。作为接受过程的一部分，数据必须与手稿一起进行同行评审，必须包含数据可用性声明。

出版社开放数据政策举例

努力推动形成新常态

（1）模板文件及流程标准化

- 制定标准化的指南（数据可用性声明模板、同行评议重点内容）简化流程，改善体验。

（2）存储库推荐

- 通用数据存储库：FAIRsharing.org、re3data.org、figshare、Dryad平台列表。
- 分学科存储库：生物科学、地球 环境和空间科学、健康科学、物理学、社会科学等。
- 自己维护的研究数据存储库（Springer Nature）。

出版社开放数据政策举例

口 努力推动形成新常态

（3）集成、开发工具和服务（IEEE）

- Code Ocean：可重现性代码共享平台，用户可以将代码和相关数据上传到站点，其他用户可以在其中运行和/或修改它们。
- IEEE DataPort ™，集数据存储、检索、共享、合作为一体的平台。

（4）咨询、支持、帮助台

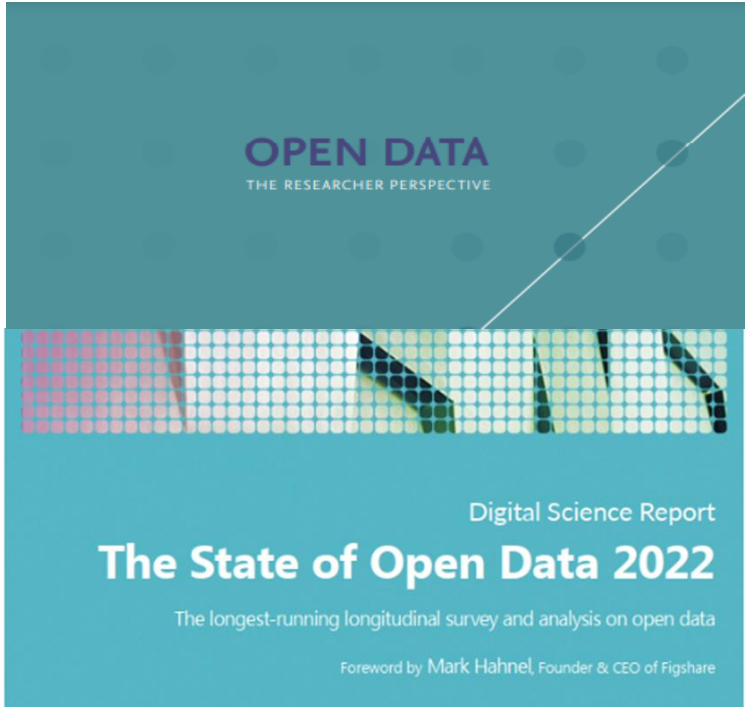
- Wiley作者合规工具
- Springer Nature 研究数据帮助台

出版社开放数据政策举例

口 努力推动形成新常态

（5）开展调研与追踪

- Elsevier和荷兰莱顿大学科学技术研究中心(CWTS)对1200名研究人员的全球调查和三个案例研究。
- Digital Science、Figshare和Springer Nature出版社联合发布发布《2022年开放数据状况》公布：超过 70% 的受访者在一定程度上认同最近的一项研究需要遵循数据共享政策。



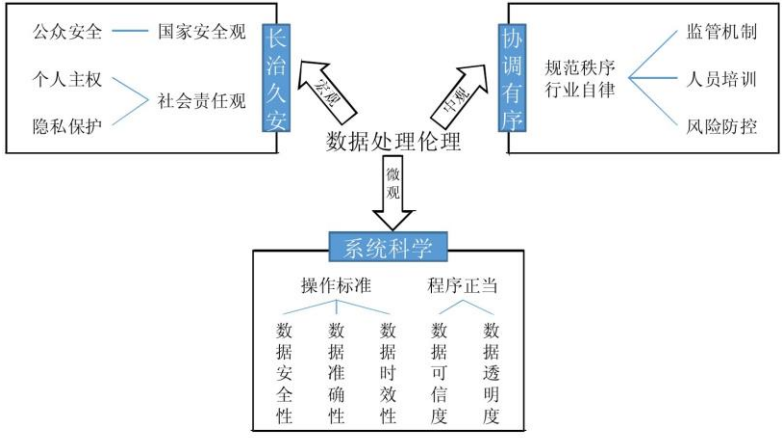
增强科学数据服务体系

如何更好地开放

是否应该开放

增强科学数据服务体系-回归数据要素

- 从生产要素的视角对科学数据确权，数据归谁所有、由谁使用、使谁获益，包括数据采集权、存储权、标记权、加工权、使用权、修改权、复制权、收益权等
- 出版社的权益边界，是否可以强制共享等。
- CARE原则下的道德准则、法律规范和利益分配等



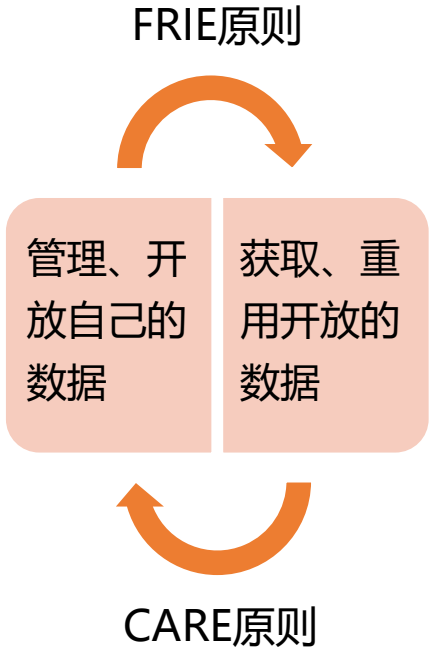
图片来源：刘静羽,贾毓洁,黄金霞,王昉.开放科学创新生态构建中的数据处理伦理准则框架研究[J].农业图书情报学报,2022,34(09):29-43.

增强科学数据服务体系-完善政策、畅通渠道

- 针对开放科学数据的滞后管理现象，在复杂的数据确权下，完善相关国家级、机构级政策规范，确保数据的机构内部妥善保存。
- 建设分级、分类的存储库，对于没有建设能力的单位，可以借助图书馆联盟的力量解决。
- 增强服务、畅通共享渠道：加大宣传，畅通机构内部的“开放数据”渠道。
- 增强、整合图书馆现有机构知识库、开放数据平台功能。
  - 2017年11月，开放存取联盟（COAR）发布了《下一代机构知识库——COAR工作组行动与技术推荐规范》
  - 定义了下一代机构知识库（Next Generation Repositories，NGR）的愿景与目标，其核心内容是保持开放性，实现科学数据重复利用和扩展增值服务

增强科学数据服务体系-增强意识、赋能用户

- 2023年2月，中共中央、国务院印发了《数字中国建设整体布局规划》，提到“构建覆盖全民、城乡融合的数字素养与技能发展培育体系”等重点内容。
- 2021年《开放科学建议书》：将一系列数据科学和数据管理技能、知识产权相关技能以及确保开放获取所需的技能，视为研究人员应具备的基础性技能，并将其纳入高等教育研究技能课程。



数字素养
数据科学
开放获取

数据监管
数据管理

谢谢！期待更多交流

wuyp@lib.pku.edu.cn