中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

# 面向AI4S的科学数据开放共享思考

中国科学院计算机网络信息中心
汇报人：杜一
2023年8月11日

# 汇报提纲

一、科学数据开放共享的发展态势

二、科学数据开放共享面临的挑战

三、科学数据开放共享的有益探索

# 科学数据“井喷式”增长



“中国天眼”

500米口径球面射电望远镜（FAST）

每秒采集数量最高可达38GB

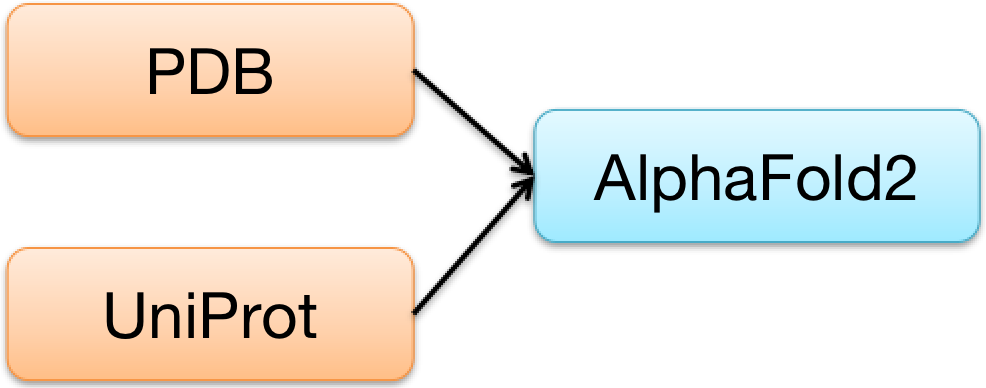
每天新增数据2000TB

https://news.sciencenet.cn/sbhtmlnews/2021/12/366880.shtm

# 科学数据在科研发现中发挥巨大作用

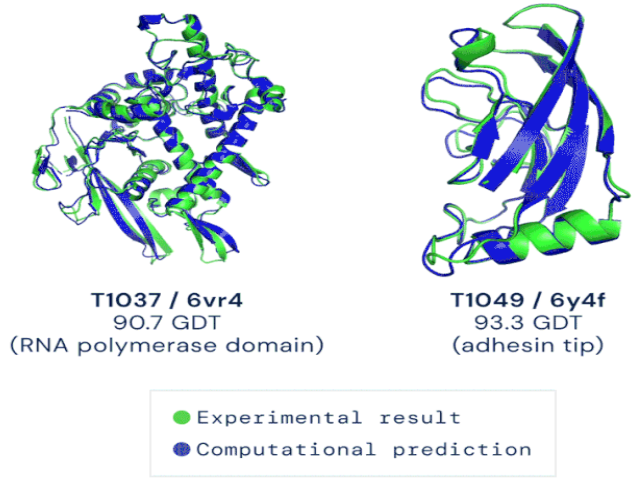
中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

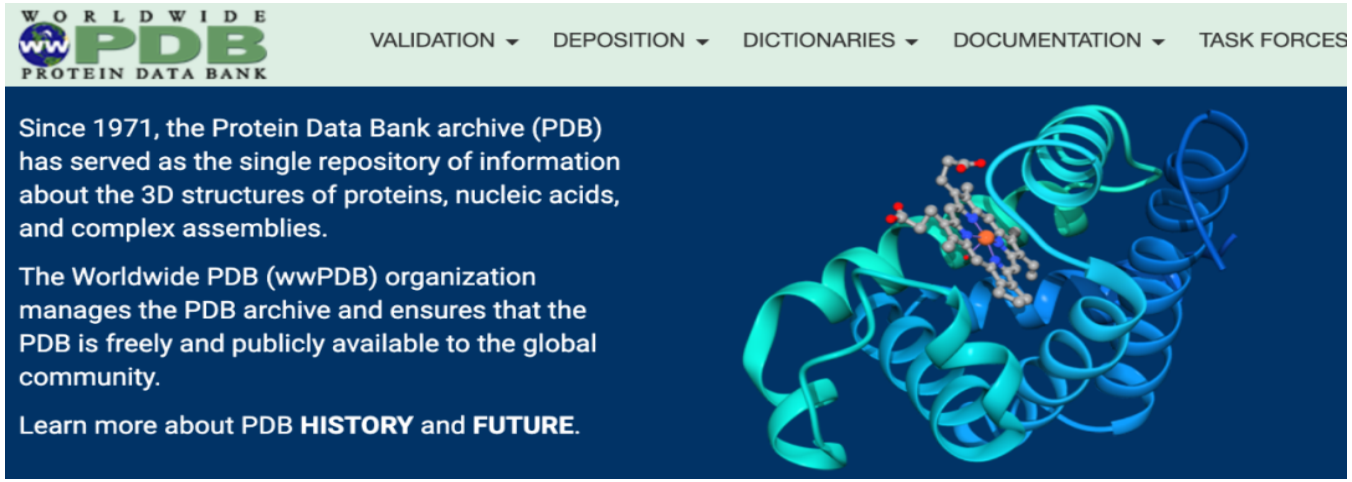
## 科研范式深刻变革，科学数据成为科研创新发展的重要驱动力




AlphaFold用于训练的数据来自蛋白质结构数据集 PDB和包含未知结构蛋白质序列的大型数据库 UniProt
共包括约 170,000 个蛋白质结构。

其中，PDB 是一个专门收录蛋白质及核酸的三维结构资料的数据集，拥有十分悠久的历史，可以追溯到 1971 年。




1. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021). doi:10.1038/s41586-021-03819-2
Tunyasuvunakool, K., Adler, J., Wu, Z. et al. Highly accurate protein structure prediction for the human proteome. Nature 596, 590–596 (2021).

# 科学数据为人工智能发展提供“养料”

中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

## 高质量数据是提高人工智能性能的关键


GPT-4 Technical Report

OpenAI*

Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

GPT-4致谢提到的共815人，分预训练、长上下文、视觉、强化学习与对齐、评估与分析、部署、额外支持七个小组。

其中，各小组中的数据团队，总人数100多人

- 预训练组中，数据团队35/125
- 视觉组数据团队10/89
- 强化学习组65/182
- 评估组10/257

通义千问　文心一言　科大讯飞 iFLYTEK　讯飞星火

**ImageNet: A Large-Scale Hierarchical Image Database**

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei
Dept. of Computer Science, Princeton University, USA
{jiadeng, wdong, rsocher, jial, li, feifeili}@cs.princeton.edu


14,197,122 张图片

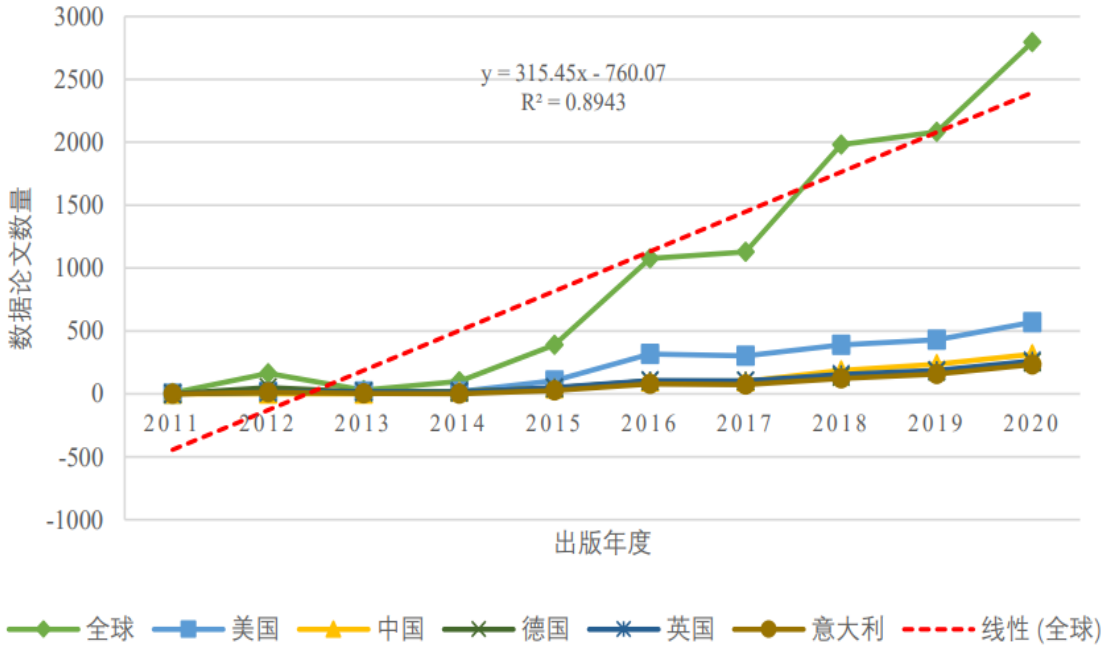
自2010年以来，每年ImageNet大规模视觉识别挑战赛，研究团队在给定的数据集上评估其算法，并在几项视觉识别任务中争夺更高的准确性。2012年，深度神经网络方法达到前所未有精度，被认为是深度学习新一轮研究热潮开始的标志事件之一

CVPR,2009,被引超55000

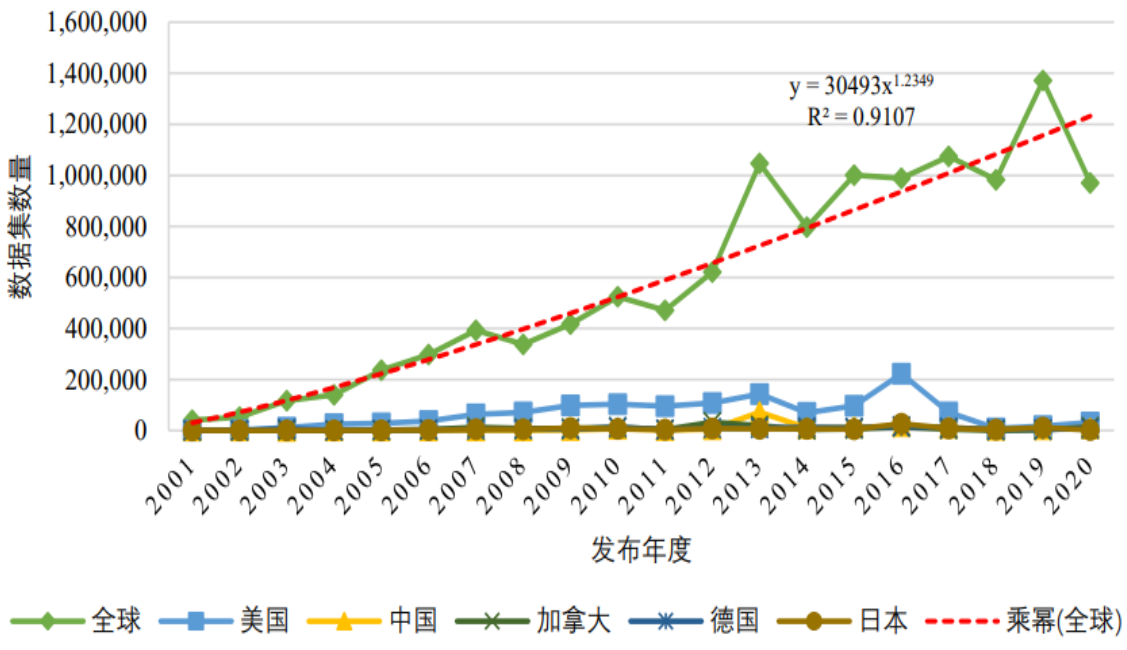
# 科学数据开放共享呈现快速上涨态势

中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

以科学数据出版为例，发现全球数据论文出版从2011年开始整体呈现快速上涨态势。数据论文出版数量排名前5的国家，其数量也基本保持逐年增长的态势。




数据论文年度分布情况




数据集年度分布情况

王卫军, 李成赞, 郑晓欢, 等. 全球科学数据出版发展态势分析——基于 Web of Science 数据库的调研[J/OL]. 中国科学数据, 2021, 6(3). (2021-09-29). DOI: 10.11922/csdata.2021.0019.zh.

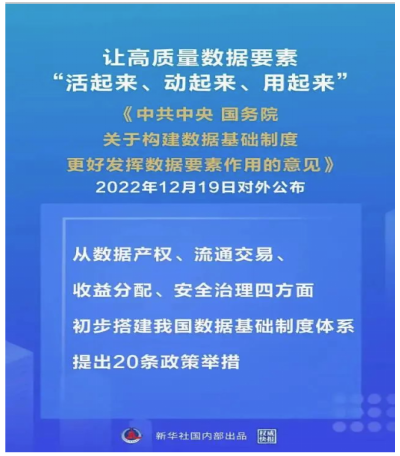
# 科学数据开放共享也成为世界各国关注的焦点

中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

## 数据纳入国家战略，科学数据成为基础性战略资源

- 2019年12月，美国《联邦数据战略与2020 年行动计划》明确将数据界定为战略资产
- 2020年2月，欧盟发布《欧洲数据战略》，开启“欧盟单一数据市场”进程
- 2020年9月，英国发布《国家数据战略》，为英国处理和投资数据以促进经济发展构建框架
- 2020年3月30日，《中共中央、国务院关于构建更加完善的要素市场化配置体制机制的意见》正式发布，明确将数据界定为第五大生产要素
- 2022年12月19日，《中共中央、国务院关于构建数据基础制度更好发挥数据要素作用的意见》正式发布，提出构建**数据产权、流通交易、收益分配、安全治理**等制度，初步形成我国数据基础制度的“四梁八柱”。



# 汇报提纲

一、科学数据开放共享的发展态势

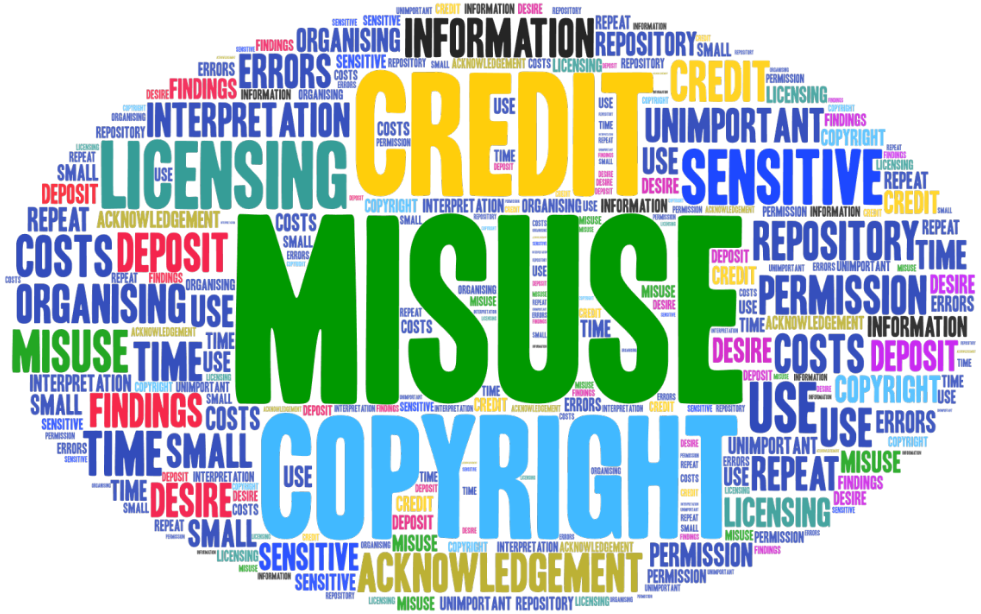
二、科学数据开放共享面临的挑战

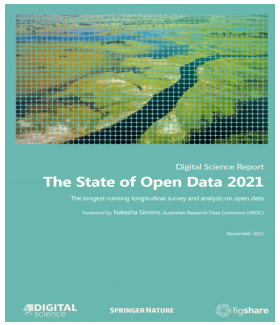
三、科学数据开放共享的有益探索

# 科研人员对于数据开放共享还存在很多疑虑



## 国内外对数据共享相关问题做过多次调研，很多共性问题阻碍了数据的开放共享




- 担心数据会被不当使用
- 不确定数据版权和许可
- 担心敏感信息泄露
- 数据共享后没有得到足够认可
- 共享数据花费高
- 没有时间发布数据
- 担心其他人抢先自己发布成果
- 数据量太大
- 觉得数据不重要
- ……

明晰 → 科学数据权属

提升 → 科学数据质量

明确 → 科学数据价值

确保 → 科学数据安全

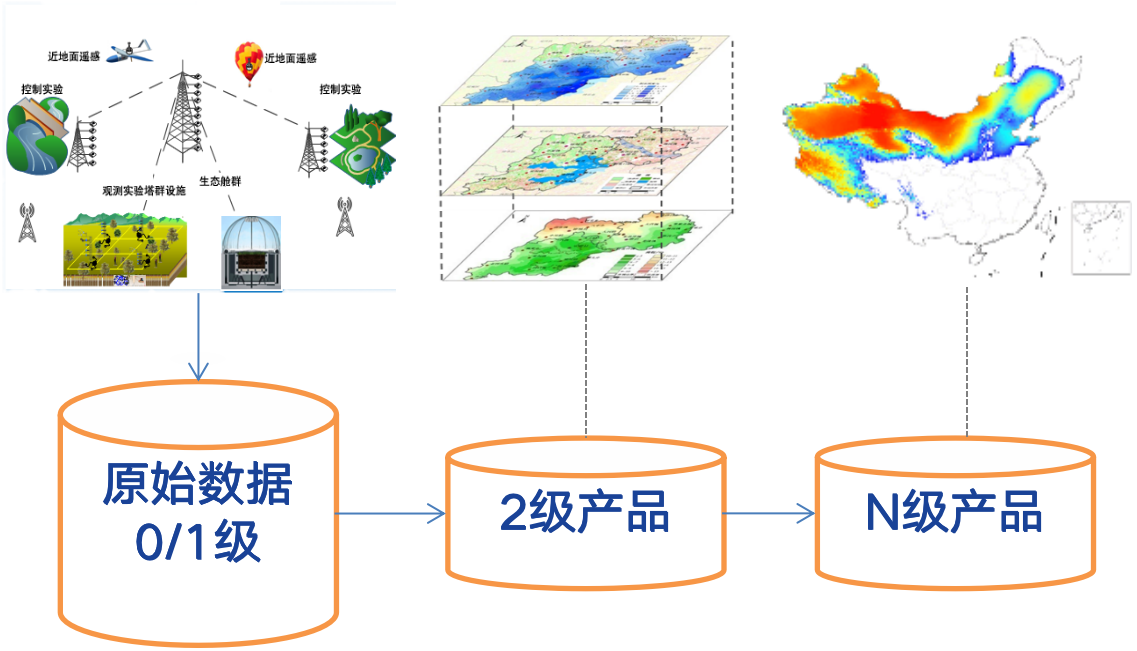
https://digitalscience.figshare.com/articles/report/The_State_of_Open_Data_2021/17061347

# 权属：科学数据的权属管理尚未形成成熟机制

中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

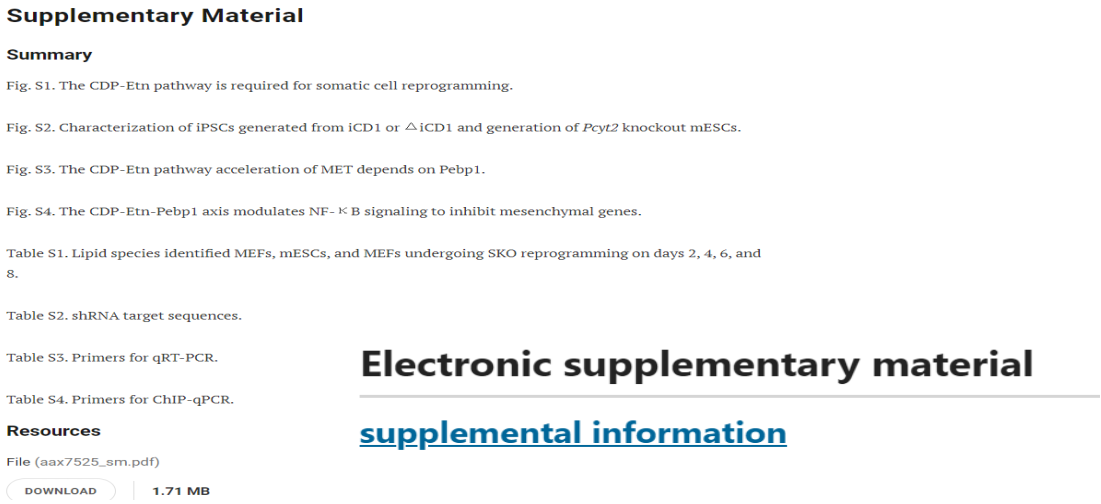
## 科学数据高价值、多主体、权属构成复杂

- 科学数据的产生、获取等往往需要大量时间和智慧的投入，其生命周期不同环节涉及多元主体的贡献，在衍生和传播过程中，权属变得更加复杂




科学数据的采集、加工、再加工等衍生过程**涉及多主体贡献**

- 论文版权管理模式已基本成熟，而科学数据权属的管理还没有形成成熟的机制
- 需要协调好知识产权的专有性和科学数据的共享性



**Supplementary Material**

**Summary**

Fig. S1. The CDP-Etn pathway is required for somatic cell reprogramming.

Fig. S2. Characterization of iPSCs generated from iCD1 or △iCD1 and generation of *Pcyt2* knockout mESCs.

Fig. S3. The CDP-Etn pathway acceleration of MET depends on Pebp1.

Fig. S4. The CDP-Etn-Pebp1 axis modulates NF-κB signaling to inhibit mesenchymal genes.

Table S1. Lipid species identified MEFs, mESCs, and MEFs undergoing SKO reprogramming on days 2, 4, 6, and 8.

Table S2. shRNA target sequences.

Table S3. Primers for qRT-PCR.

Table S4. Primers for ChIP-qPCR.

**Resources**

File (aax7525_sm.pdf)

DOWNLOAD | 1.71 MB

**Electronic supplementary material**

supplemental information

常见的论文关联数据共享，往往以论文附件的形式出现，**无法得到引用**，其产权往往没有得到认可和保障

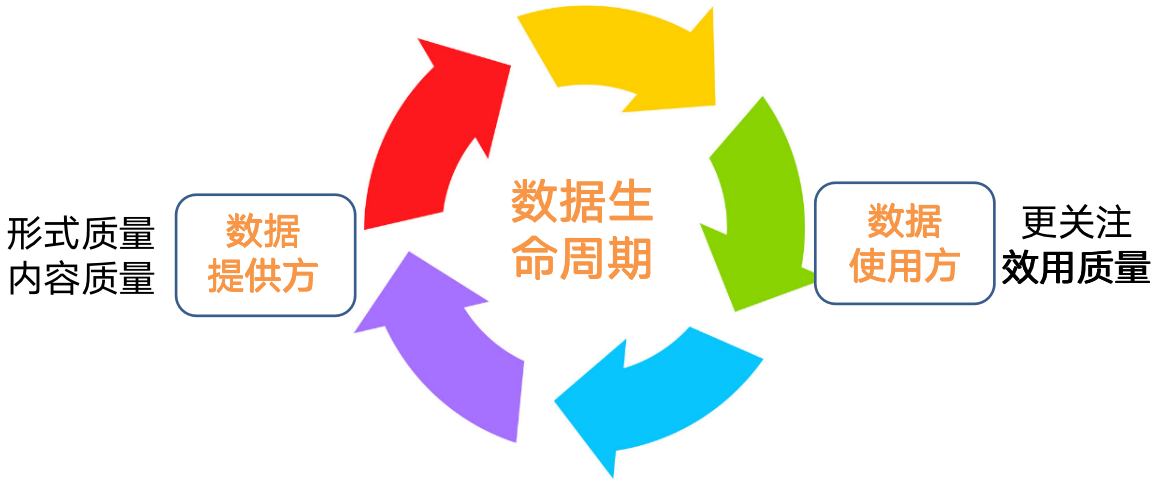
# 质量：数据质量治理是影响开放共享的重要因素

中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

质量是数据的生命。科学数据质量控制是一个复杂过程，数据采集、组织、加工、存储、开放与应用等每个环节都会影响数据质量；ChatGPT等新兴技术的使用也会产生大量噪声数据，进一步加剧数据质量问题。

**Data, Data, Everywhere, Nor Any Drop to Drink**

--Borgman, Christine L.，2014




数据产品从采集开始涉及多种处理方法和环节，开放科学场景下，关注数据的复用性，用户更关心数据的效用质量以及数据处理全过程所采用的手段和方法

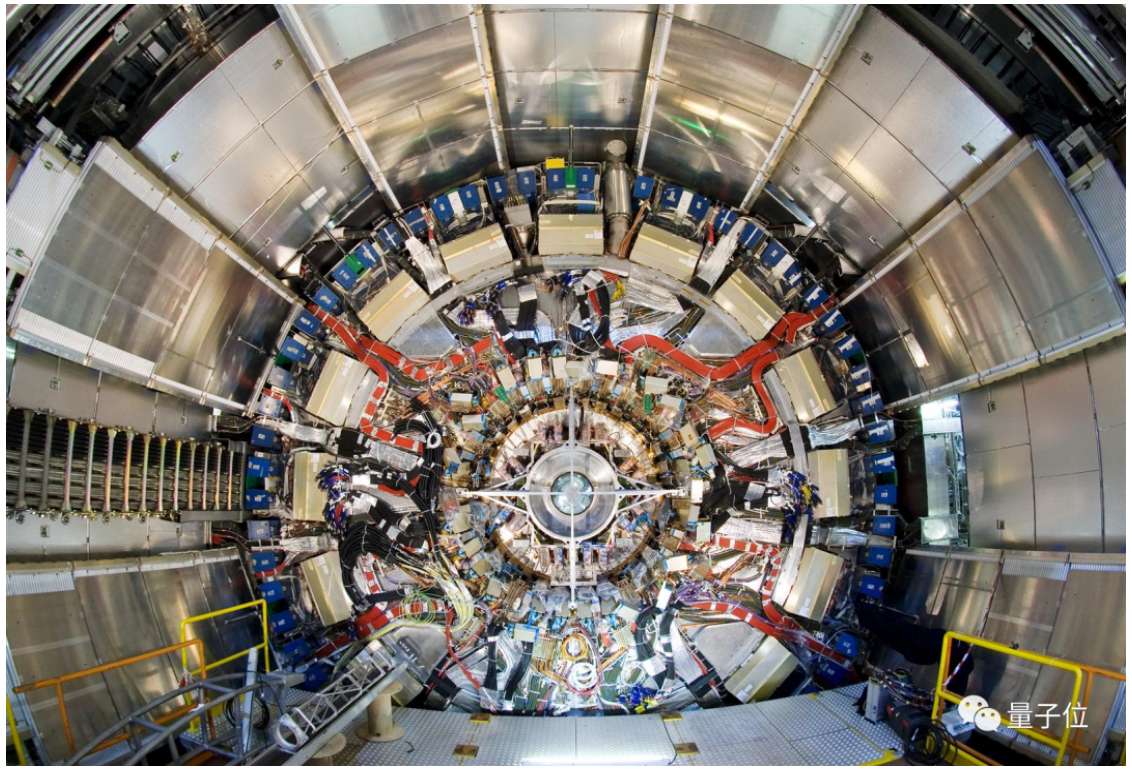
| *Computer vision task (steel sheet inspection)* | Accuracy |
|---|---|
| Baseline | 76.2% |
| Model-Centric | +0% |
| Data-Centric | +16.9% (93.1%) |

数据质量提升能大幅提升模型效果

# 价值：科学数据的价值难以有效衡量

中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

**科学数据研制过程中涉及的种类多、环节多，以及全球内流通共享等问题，导致各环节科研人员贡献值的测算目前没有统一方法**





欧洲核子研究中心的ATLAS对撞机每秒产生的数据量相当于**地球上每个人同时进行20次电话交谈**，却只有**不到百万分之一**是有研究价值的，实验中筛选、处理数据需要由分布在全球的超过130个超级计算机支持。论文的8000个署名作者中，除了物理学家，还包括大量做**数据筛选和分析的软件工作者**。

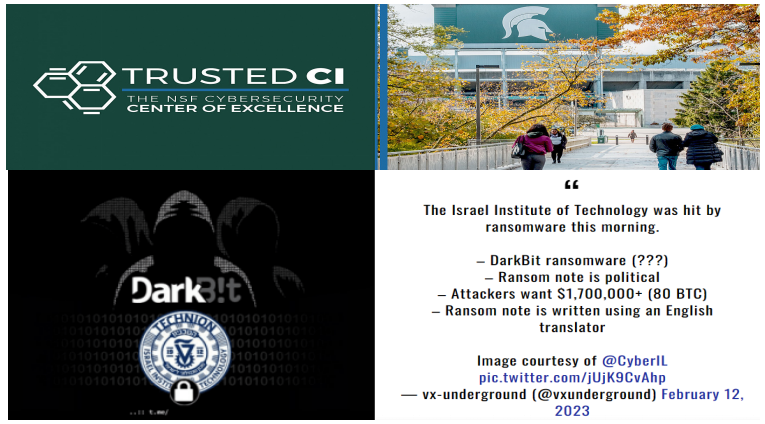
https://mp.weixin.qq.com/s/rpsratmQTzdXaP-Y0DW33g

# 安全：数据安全治理也是开放数据的重要考量要素

中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

**数据安全风险对科学数据开放共享构成新的挑战，数据安全问题已成为关系到内部安全、外部安全以及科学发展的重点问题。**

## 数据泄露与非法盗取问题




随着近年来大量科研工作的线上开展，黑客组织瞄准科研机构，2020年密歇根州立大学物理和天文学系遭受勒索软件攻击使研究人员无法访问他们的科学数据，导致50%-70%的研究被迫停止，2023年以色列理工学院被黑客入侵，大量内部文件遭到泄露。

## 个人数据隐私保护问题

2018年6月，美国联邦贸易委员会对多家基因检测公司进行调查，质疑其处理用户信息和基因数据的方式，以及如何将这些数据共享给第三方。

## 数据主权问题


开放科学全球化加速了科学数据跨境流动，同时也引发越来越多数据安全风险和监管挑战。

https://news.iu.edu/live/news/30899-ransomware-attackers-set-their-sights-on-research
https://securityaffairs.com/142160/hacking/israeli-technion-suffered-ransomware-attack.html

# 汇报提纲

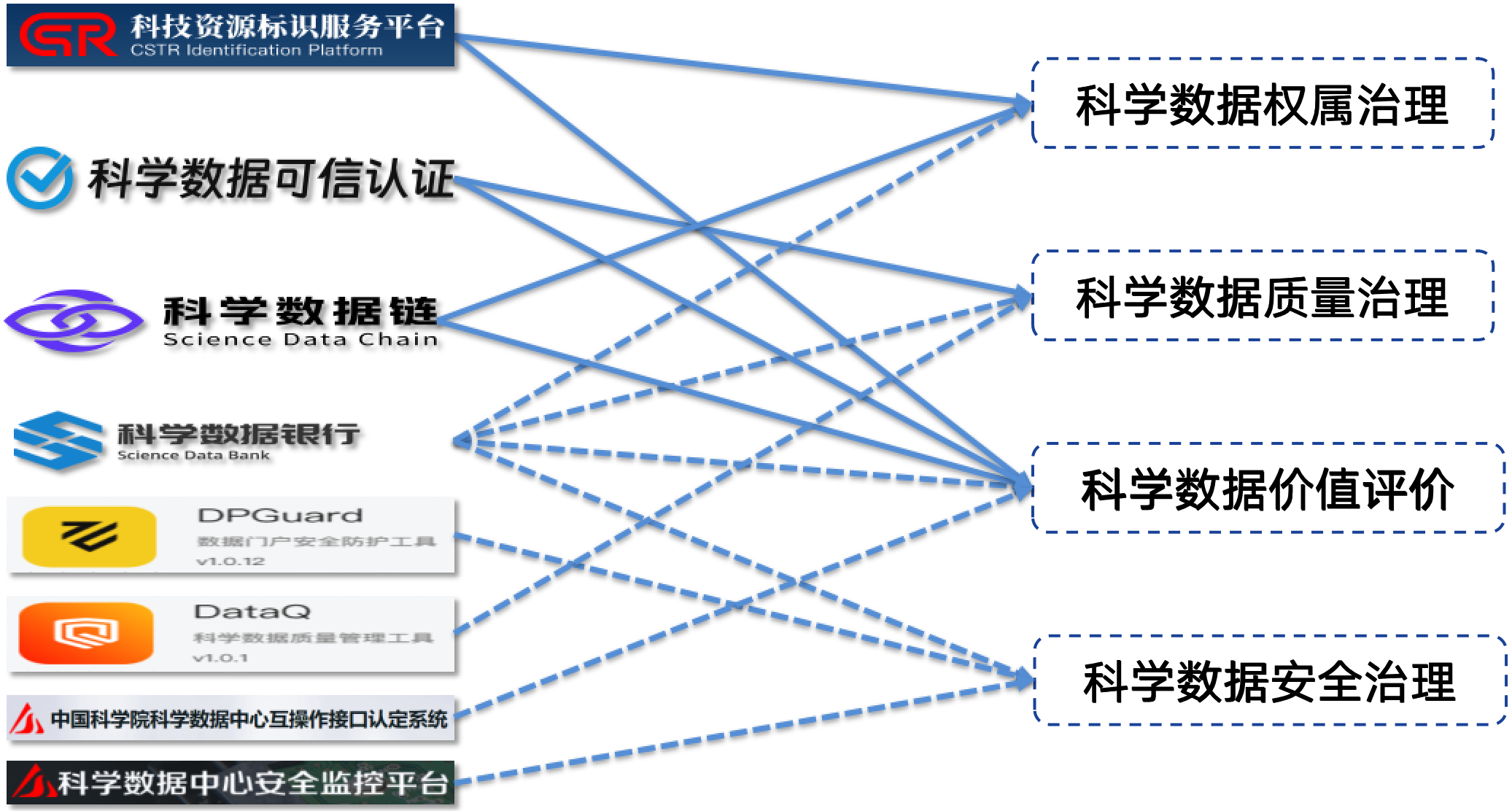
一、科学数据开放共享的发展态势

二、科学数据开放共享面临的挑战

三、科学数据开放共享的有益探索

# 推动科学数据开放共享的有益探索





# CSTR服务平台，促进科技资源信息互通

中国科学院计算机网络信息中心
Computer Network Information Center, Chinese Academy of Sciences

GRID机构标识 107,102

ORCID人员标识 12,336,757

ROR机构标识 100,467

欧盟赞助项目标识 115, 642

National Science Foundation WHERE DISCOVERIES BEGIN NSF赞助项目标识 491,339

- 科学数据、学术论文、科研人员、基金项目是贯穿科学研究全生命周期的关键科技资源
- 当前科技资源面临数据碎片化、数据外流、多头填报提交、数据壁垒等问题
- **标识服务**是为每一个科技资源分配全球唯一“身份证”，与全球主流标识互联互通的支撑服务，有助于科研成果自主管理与国际互认。


## 科技资源标识服务 https://www.cstr.cn

**自主并兼容国际的CSTR科技资源标识**
**GB/T 32843—2016《科技资源标识》**

支持8类国内国际主流标识兼容解析；汇聚5类国际开放资源

双节点24H不间断的稳定服务、为国家重大需求部署独立解析网关

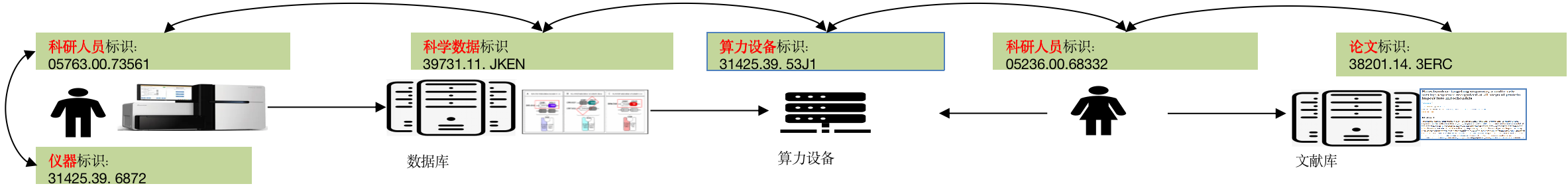
注册量 497,984条　解析量170,581次

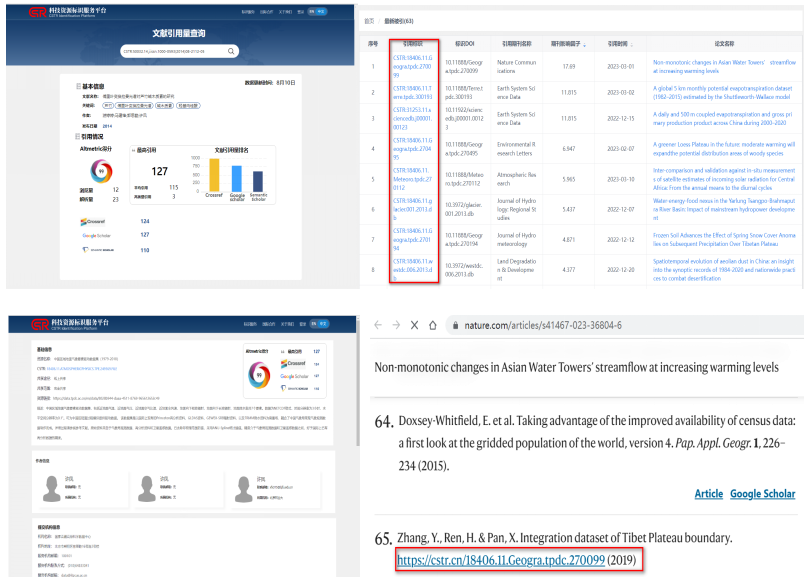
机构用户58个　汇聚量256,613,592条

# CSTR服务平台，促进科技资源信息互通

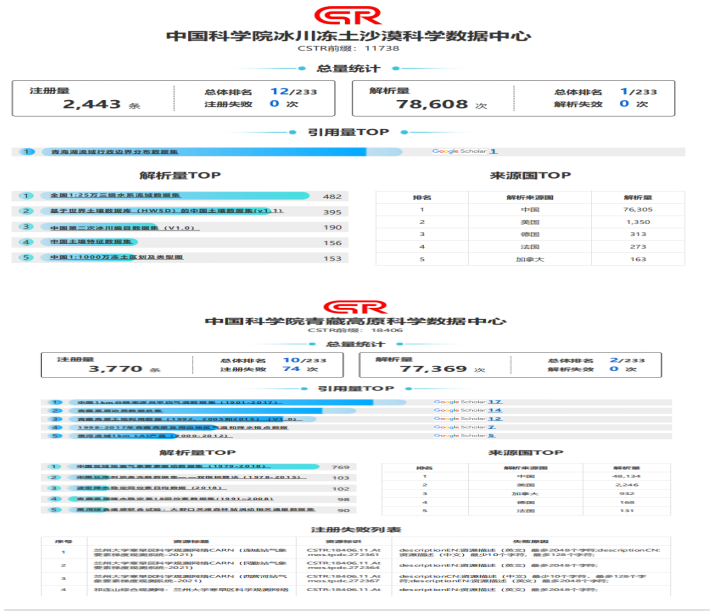
中国科学院计算机网络信息中心
Computer Network Information Center, Chinese Academy of Sciences

## 构建科技资源全要素关系，形成全球影响力追踪网络




科学数据、论文影响力追踪




为科学数据中心提供基于CSTR的引用追踪服务


以人员为核心进行成果及引用追踪

# 可信认证服务平台，提升数据可用性和可靠性

中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

## 建设背景

在第S69次香山科学会议，形成了建立由我国主导的科学数据中心认证体系的建议共识。

## 服务定位

以“注册-指南-认证-监测”的机制为实践基础，面向科学数据中心、存储库、知识库等科学数据服务平台，提供注册、认证等多层级服务

可信平台
快速发现

可信平台
认证服务

可信平台
监测评价

助力数据
国际传播

科学数据可信认证 https://datatrusted.cn/

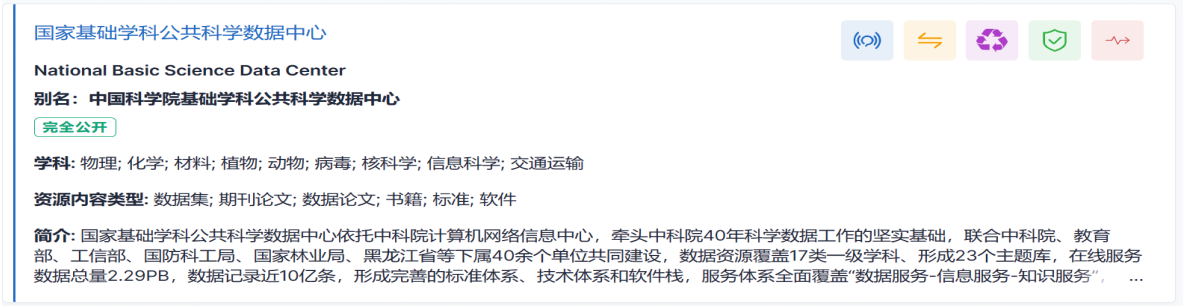
# 可信认证服务平台，提升数据可用性和可靠性

中国科学院计算机网络信息中心
Computer Network Information Center, Chinese Academy of Sciences

## 助力科研人员快速发现高质量科学数据服务平台

- 已收录国内外平台3000多个
- 包括国内外科学数据中心、存储库、知识库等
- 实现基于平台基本信息、数据服务、规范、政策、组织架构等方面的筛选分类功能


## 为服务平台提供认证、增值服务和影响力提升



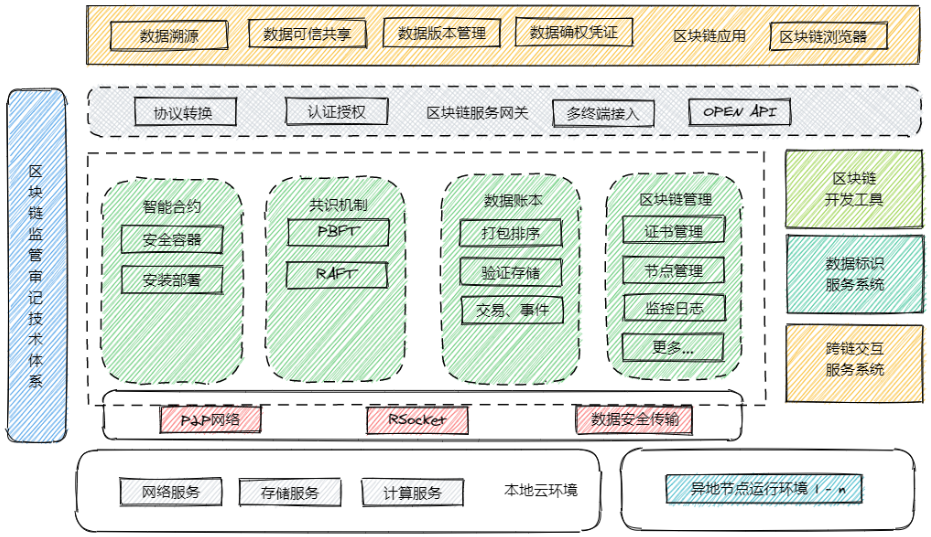


# 科学数据链-基于区块链的数据安全溯源

中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

- **科学数据链**（Science Data Chain）
  - 以**数据存证、确权、溯源**等应用为需求牵引，为科学数据提供安全、可信的区块链服务
  - 基于**自主研发**、安全可控的**ODC区块链引擎**构建
  - 加入科链科学数据链，保证上链数据**安全共享**、**可信存证**、**数据追溯**、**版本管理**、**确权凭证**、**使用记录**、**贡献量统计**等

科学数据链
Science Data Chain



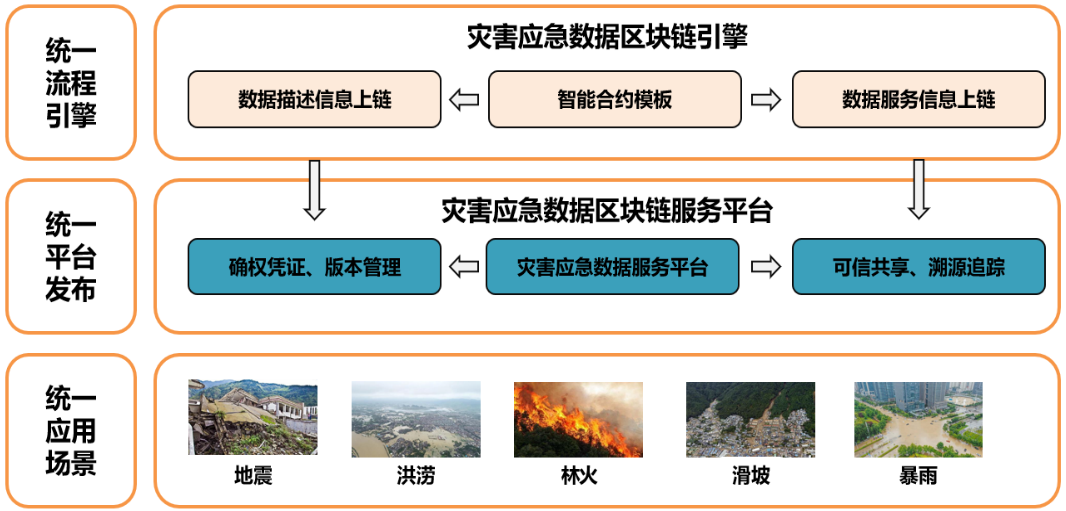


**京网信备 11010820691792890011号**
**（首个备案编号的科学数据区块链）**

# 科学数据链-基于区块链的数据安全溯源

中国科学院计算机网络信息中心
Computer Network Information Center, Chinese Academy of Sciences

## 基于区块链的灾害应急数据服务平台



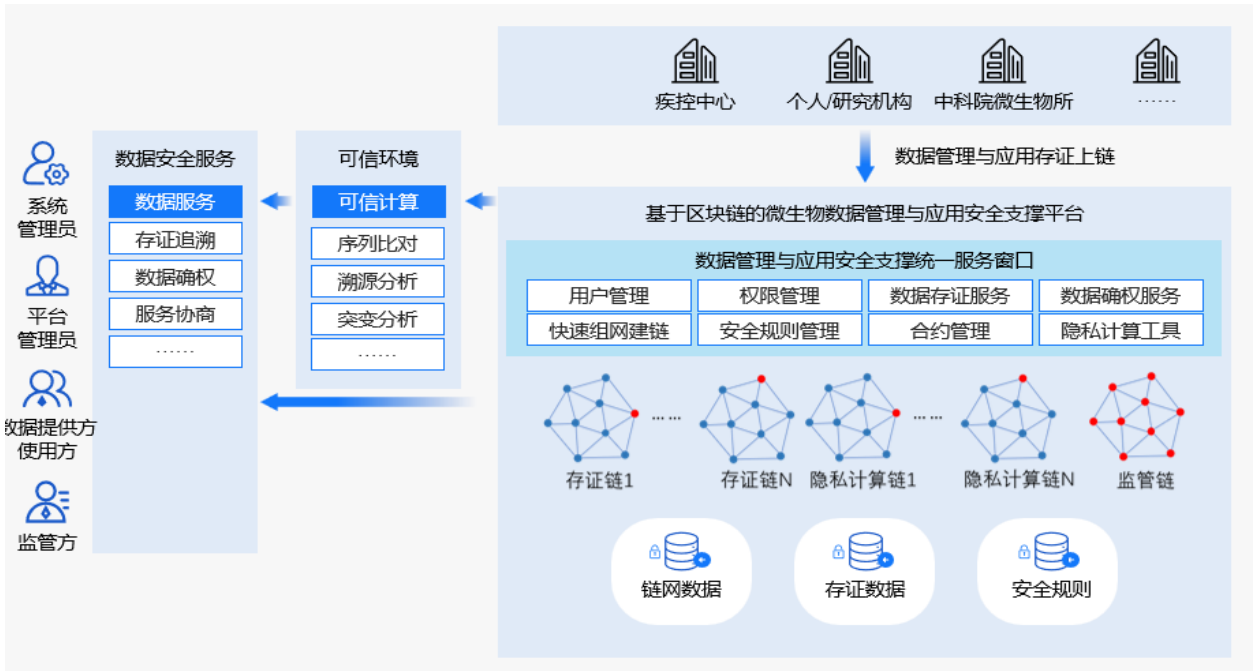

节点信息

| 节点名称(英) | 节点名称(中) | 创建机构名称 | 创建时间 |
|---|---|---|---|
| 国家对地观测科学数据中心节点 | 国家对地观测科学数据中心节点 | 国家对地观测科学数据中心 | 2022-06-28 17:43:49 |
| 国家卫星海洋应用中心节点 | 国家卫星海洋应用中心节点 | 国家卫星海洋应用中心 | 2022-06-28 17:36:29 |
| 天仪研究院节点 | 天仪研究院节点 | 天仪研究院 | 2022-06-28 17:11:56 |
| 长光卫星技术股份有限公司节点 | 长光卫星技术股份有限公司节点 | 长光卫星技术股份有限公司 | 2022-06-28 17:27:18 |
| 国家综合观测数据共享平台节点 | 国家综合观测数据共享平台节点 | 国家综合观测数据共享平台 | 2022-06-28 17:41:25 |
| 国家遥感应用工程技术研究中心节点 | 国家遥感应用工程技术研究中心节点 | 国家遥感应用工程技术研究中心 | 2022-06-28 17:47:26 |
| aircas | 中国科学院空天信息创新研究院节点 | 中国科学院空天信息创新研究院 | 2022-06-28 17:38:41 |
| 国家卫星气象中心节点 | 国家卫星气象中心节点 | 国家卫星气象中心 | 2022-06-28 17:40:00 |
| 珠海欧比特宇航科技股份有限公司节点 | 珠海欧比特宇航科技股份有限公司节点 | 珠海欧比特宇航科技股份有限公司 | 2022-06-28 17:30:58 |
| 中国遥感卫星地面站节点 | 中国遥感卫星地面站节点 | 中国遥感卫星地面站 | 2022-06-28 17:45:52 |

合约信息信息

| 合约名称 | 所属子链 | 数据存储范围 | 数据类型 | 创建时间 |
|---|---|---|---|---|
| 灾害数据合约 | 灾害应急数据链 | 链内共享 | 数据集 | 2022-07-08 12:57:36 |

## 基于区块链的微生物科学数据安全管理与追溯基础支撑平台





- 保障隐私计算任务数据端到端的隐私性
- 保障隐私计算中数据全生命周期的安全性
- 保障隐私计算过程的可追溯性


中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences
敬请批评指正！