

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



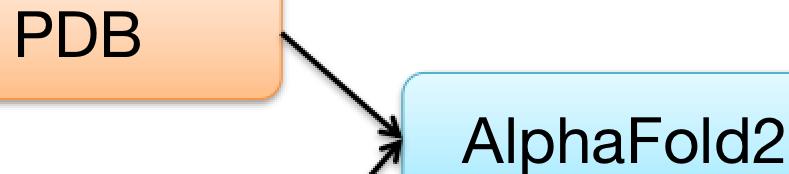
“中国天眼”
500米口径球面射电望远镜 (FAST)

每秒采集数量最高可达38GB

每天新增数据2000TB

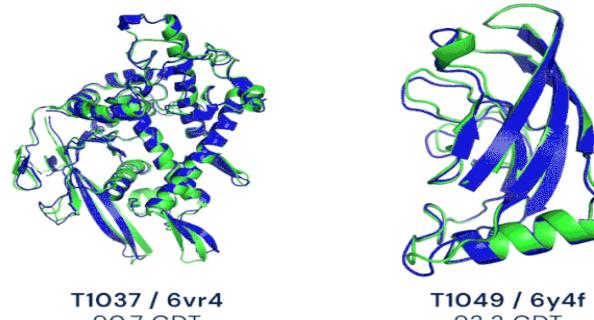
<https://news.sciencenet.cn/sbhtmlnews/2021/12/366880.shtml>

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

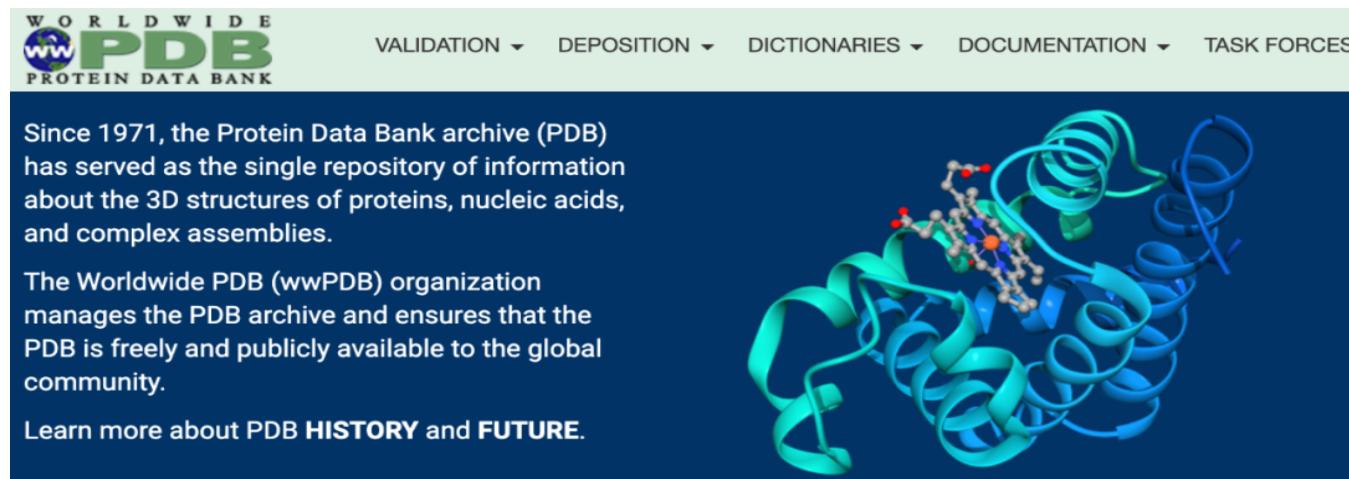


AlphaFold 用于训练的数据来自蛋白质结构数据集 PDB 和包含未知结构蛋白质序列的大型数据库 UniProt，共包括约 170,000 个蛋白质结构。

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。



- Experimental result
- Computational prediction



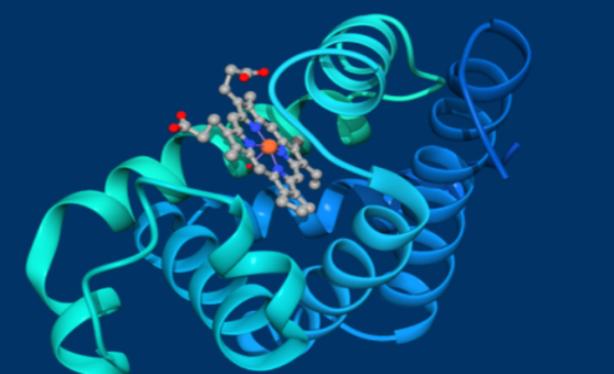
WORLDWIDE
PDB
PROTEIN DATA BANK

VALIDATION ▾ DEPOSITION ▾ DICTIONARIES ▾ DOCUMENTATION ▾ TASK FORCES

Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

Learn more about PDB **HISTORY** and **FUTURE**.



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 tasks, including passing a simulated bar exam with scores among 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 fluency and adherence to instructions. We also describe the hardware design, deployment 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

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

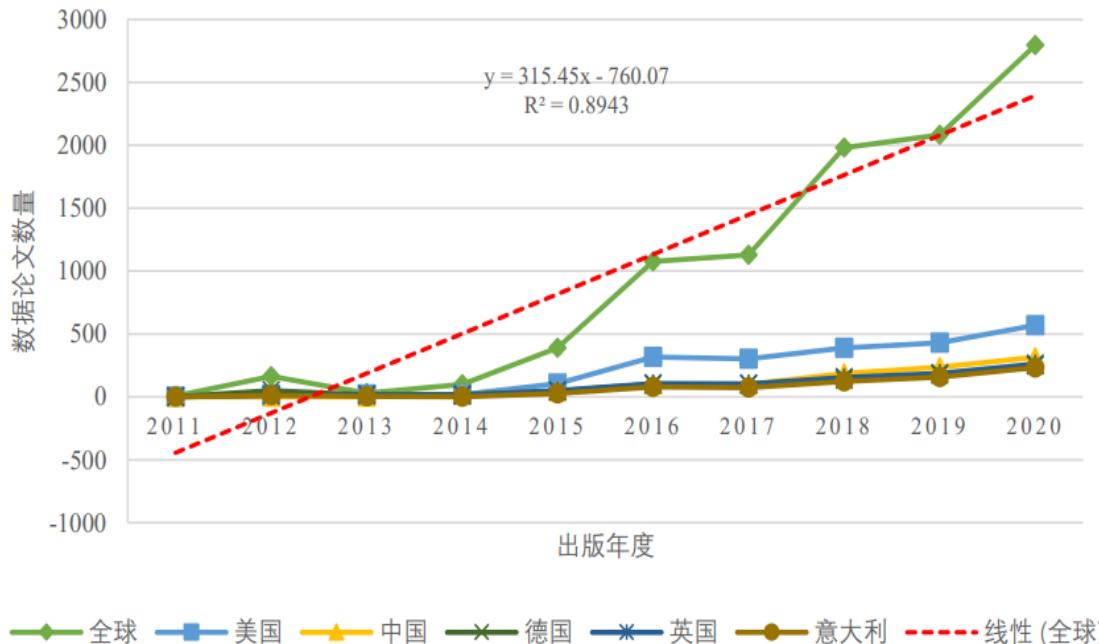
自2010年以来，每年ImageNet大规模视觉识别挑战赛，研究团队在给定的数据集上评估其算法，并在几项视觉识别任务中争夺更高的准确性。

2012年，深度神经网络方法达到前所未有的精度，被认为是深度学习新一轮研究热潮开始的标志事件之一

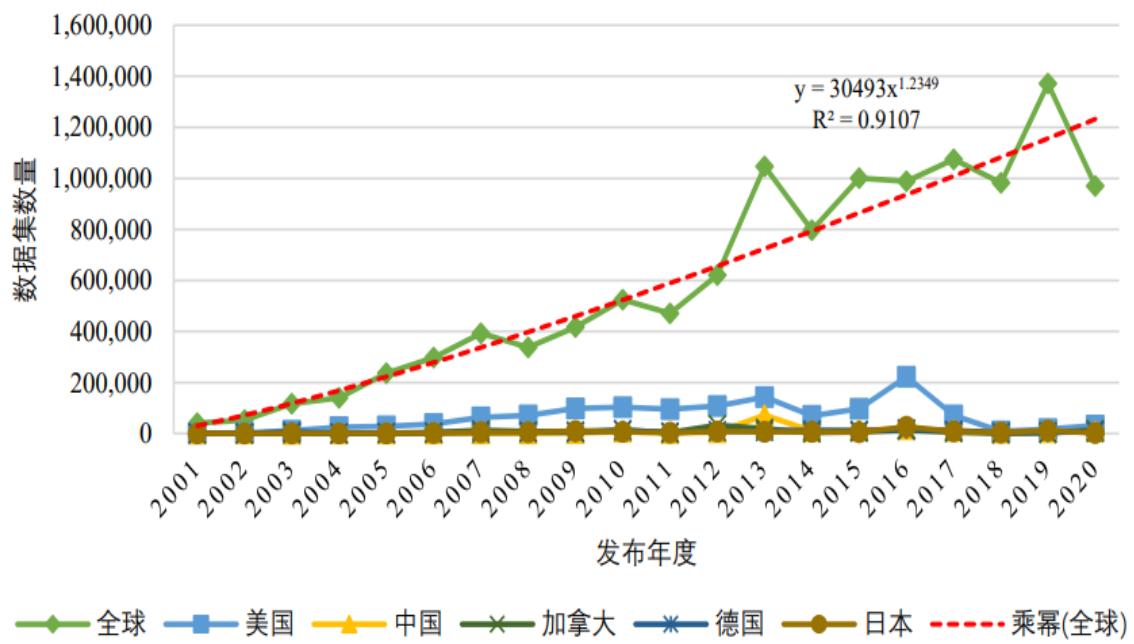
CVPR,2009,被
引起超55000

科学数据开放共享呈现快速上涨态势

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。



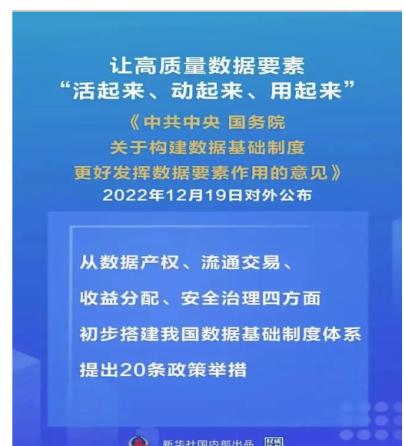
数据论文年度分布情况



数据集年度分布情况

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

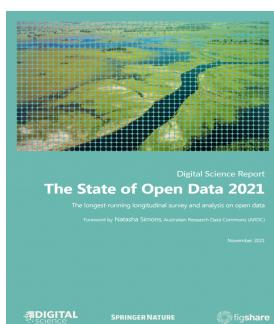
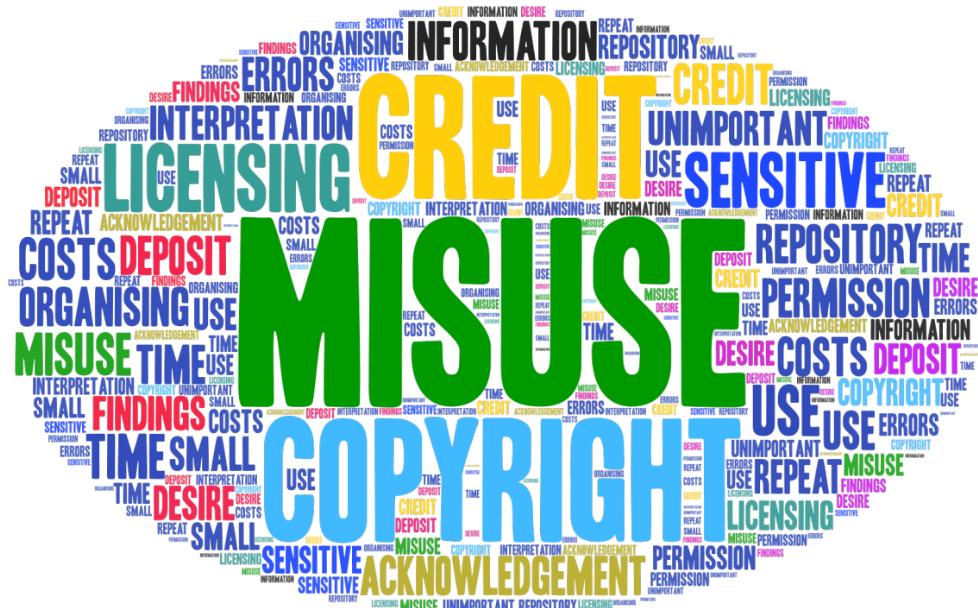
一、科学数据开放共享的发展态势

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

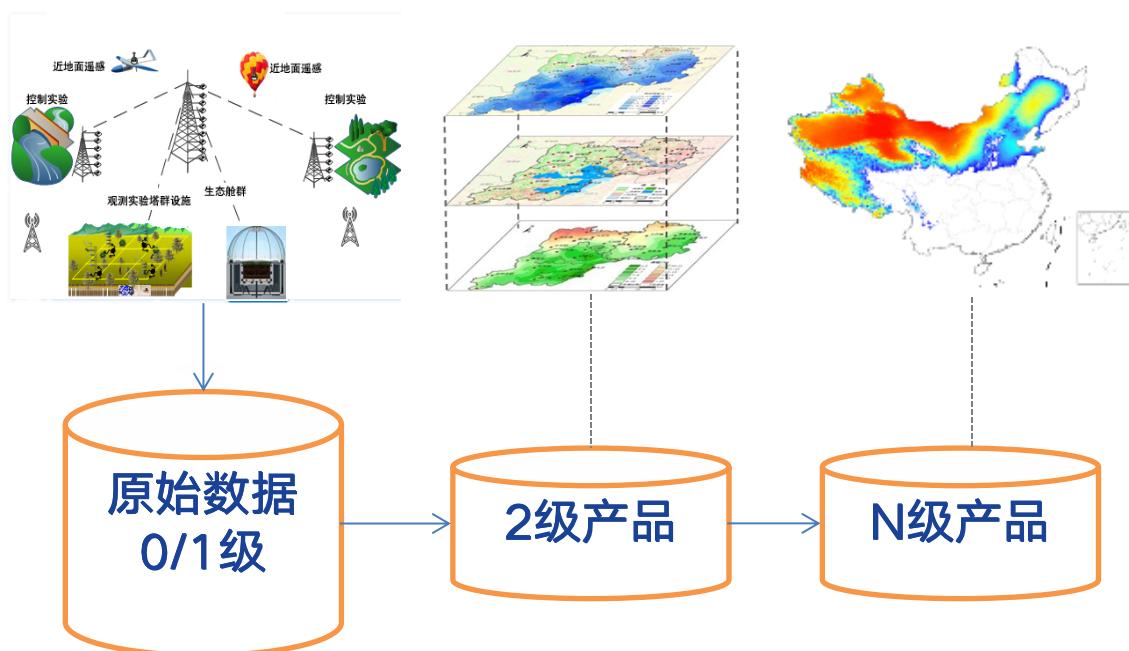
科学数据价值

确保

科学数据安全

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
- 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



科学数据的采集、加工、再加工等衍生过程涉及多主体贡献

Supplementary Material

Summary

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

Fig. S2. Characterization of iPSCs generated from iCD1 or Δ 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- κ B 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

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)*

Baseline

Accuracy

76.2%

Model-Centric

+0%

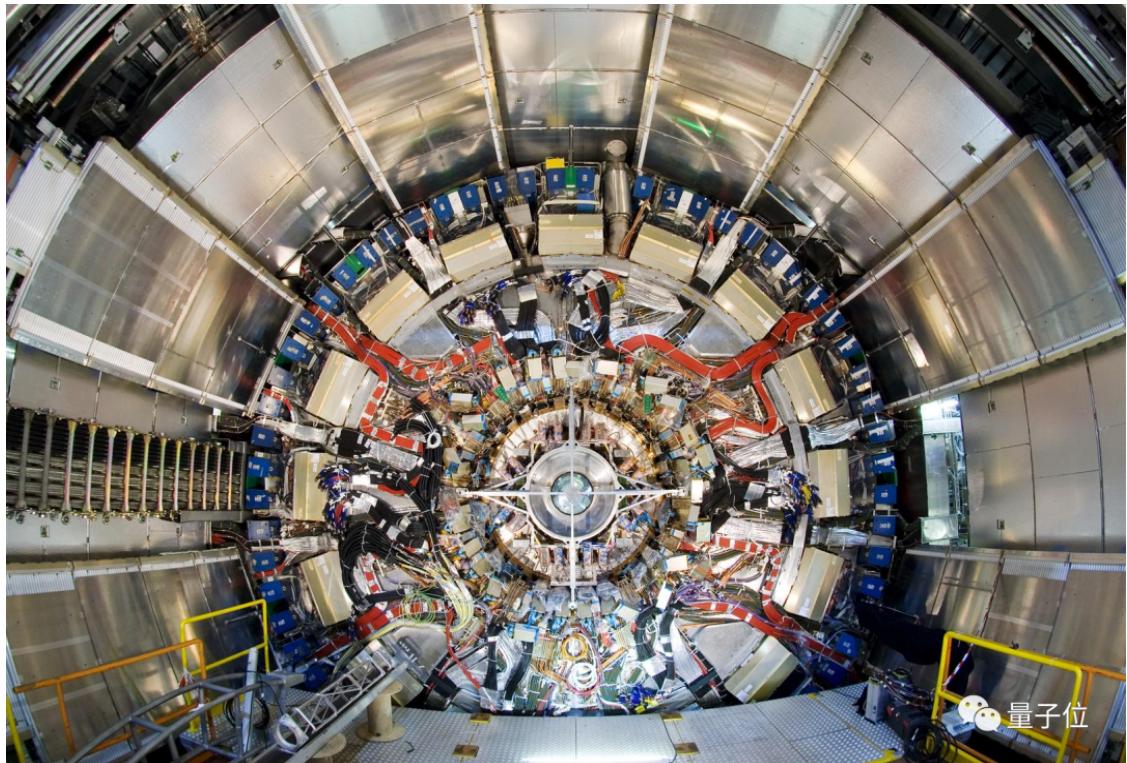
Data-Centric

+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

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



Jason Priem
@jasonpriem

A physics paper with 8,778 authors. I want them all to built a little city to live in together, with schools and parks and a nice cozy supercollider. hal.archives-ouvertes.fr/hal-03583443

不分第一作者，一律按姓氏字母顺序排列



264 转推 102 引用推文 1,386 喜欢次数

量子位

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。

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



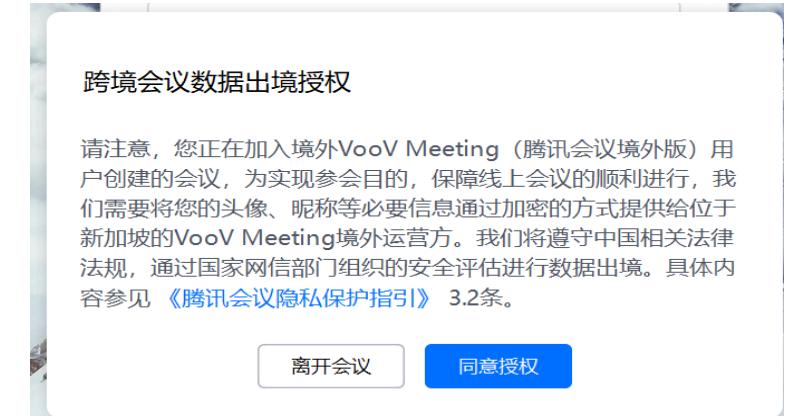
随着近年来大量科研工作的线上开展，黑客组织瞄准科研机构，2020年密歇根州立大学物理和天文学系遭受勒索软件攻击使研究人员无法访问他们的科学数据，导致50%-70%的研究被迫停止，2023年以色列理工学院被黑客入侵，大量内部文件遭到泄露。

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。

汇报提纲

一、科学数据开放共享的发展态势

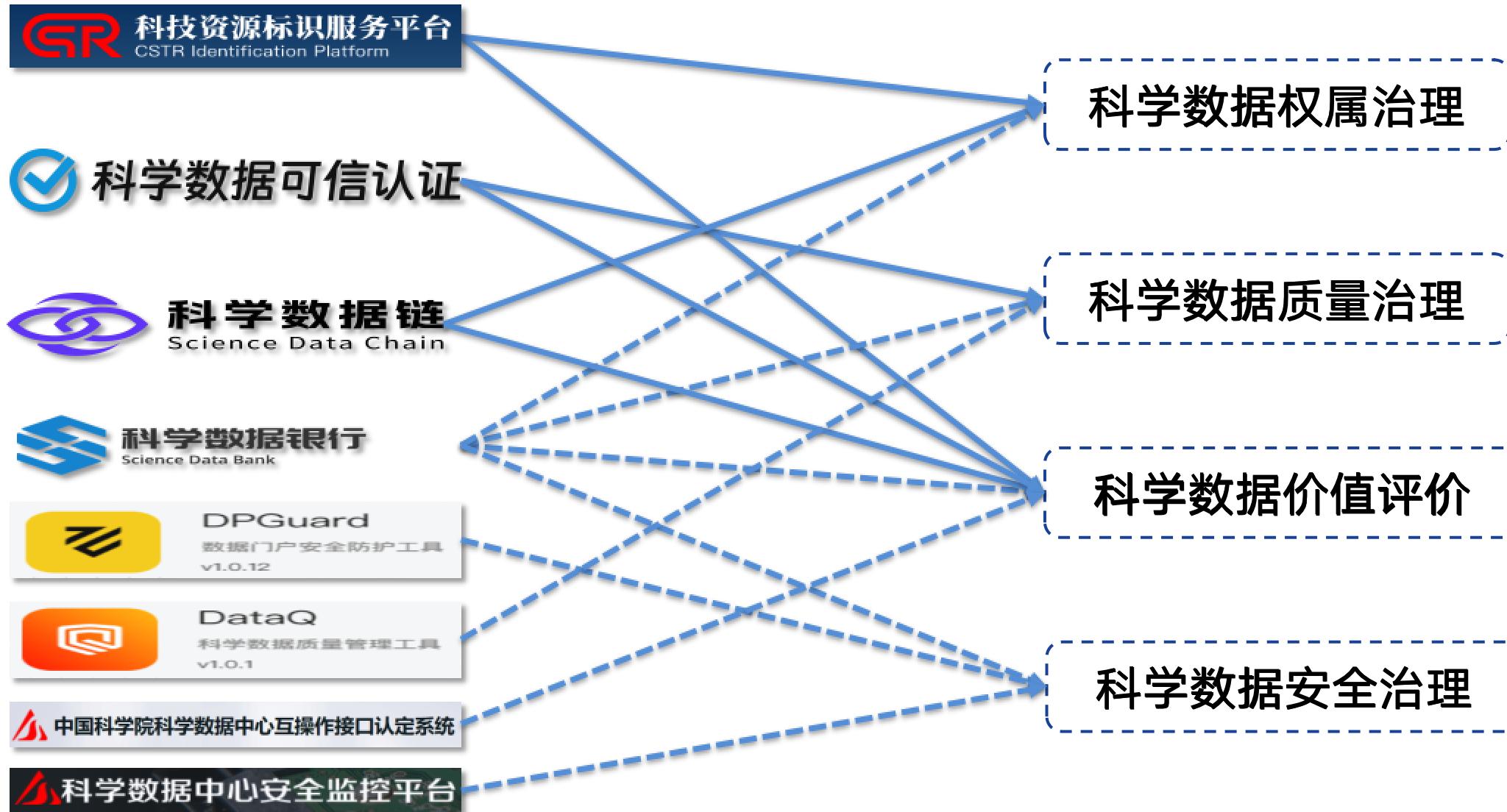
二、科学数据开放共享面临的挑战

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



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



CSTR服务平台，促进科技资源信息互通



GRID机构标识
107,102



ROR机构标识
100,467



National Science Foundation
WHERE DISCOVERIES BEGIN



ORCID人员标识
12,336,757



欧盟资助项目标识
115,642

NSF资助项目标识 491,339

The screenshot shows the homepage of the CSTR Identifier service. The header includes the logo and text '科技资源标识服务平台 CSTR Identifier'. Navigation links include '成为会员', '标准注册', '关于我们', '登录', 'EN 中文'. The main content features the 'CSTR' logo and the tagline '科技资源唯一标识, 推动科技资源开放共享'. Below this are search filters for '科学数据', '科研人员', '科研机构', and '科研项目'. A search bar contains the text '关键字、标题、CSTR' and a magnifying glass icon. At the bottom, there is a small note: '版权所有©2019-2022 中国科学院计算机网络信息中心 | 京ICP备 09112257号-127'.

- 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**

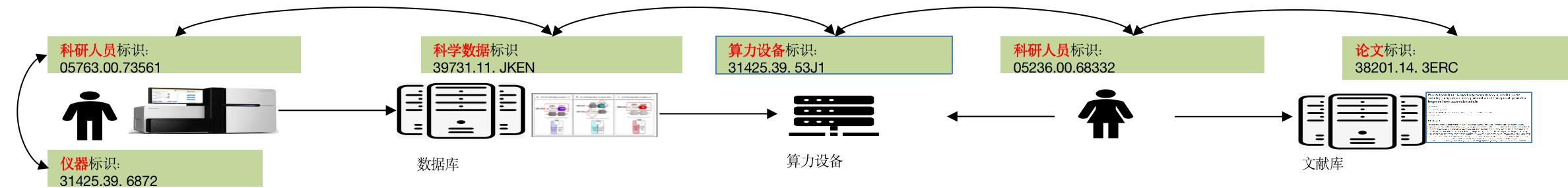
汇聚量**256,613,592条**

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



中国科学院冰川冻土沙漠科学数据中心
CSTR标识码: 11738

引用量TOP(3)

序号 标识码 期刊/会议 刊期/卷期 变更时间 变更原因

1 CSTR:11738.11.05.0001 地球化学 17/09 2020-01-01 Non-monotonic changes in Asian Water Towers' streamflow at increasing warming levels

2 CSTR:11738.11.05.0002 地球化学 11/05 2020-01-02 A global 3-km spatially explicit hydrographer dataset (1962–2015) estimated by the Shallowest-Wave model

3 CSTR:11738.11.05.0003 地球化学 11/05 2020-12-15 A daily 30 km spatially explicit hydrographer and gross production dataset across China during 2000–2020

4 CSTR:11738.11.05.0004 地球化学 6/947 2020-02-07 A greater loss of species in the future moderate warming will probably predict a distribution of woody species

5 CSTR:11738.11.05.0005 地球化学 5/965 2020-01-10 Inter-comparison and validation against *in situ* measurement of soil moisture estimation of increasing soil salinity for Central Africa from the annual mean to the decadal scale

6 CSTR:11738.11.05.0006 地球化学 5/417 2020-12-07 New insights into the role of the Asian Plateau in the global hydrological cycle: the basin impact of mountain hydro-power development

7 CSTR:11738.11.05.0007 地球化学 4/371 2020-12-12 Frozen soil Adheres the Effect of Spring Snow Cover Annex to Subsequent Precipitation Over Tibetan Plateau

8 CSTR:11738.11.05.0008 地球化学 4/377 2020-12-20 Spatio-temporal evolution of arid soil in China—an insight into the relationship between arid soil and national policies to control desertification

总数量 2,443 次 总体排名 12/233 总体失败 0 次 总体排名 1/233 总体失败 0 次

引用量TOP

来源国TOP

中国科学院冰川冻土沙漠科学数据中心
CSTR标识码: 11738

注册量 3,770 次 总体排名 10/233 总体失败 74 次 总体排名 2/233 总体失败 0 次

引用量TOP

来源国TOP

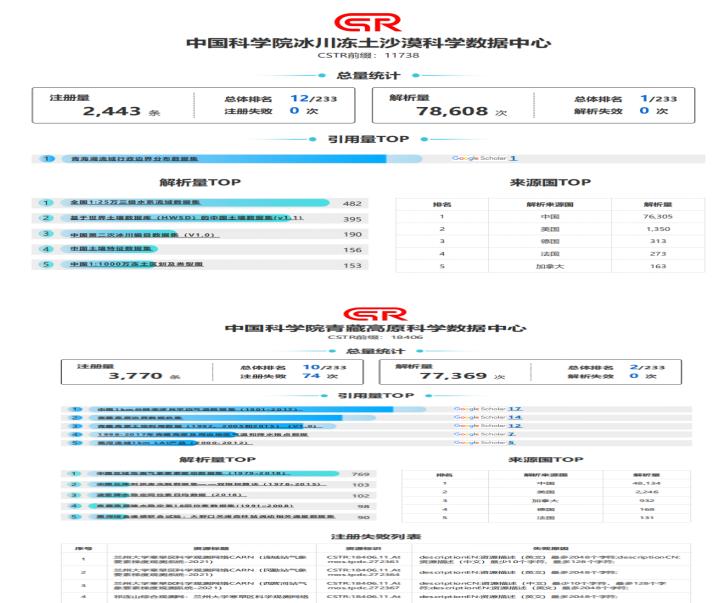
注册失败列表

Article Google Scholar

Non-monotonic changes in Asian Water Towers' streamflow at increasing warming levels

64. Doxey-Whitfield, E. et al. Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4. *Pop. Appl. Geogr.* 1, 226–234 (2015).

65. Zhang, Y., Ren, H. & Pan, X. Integration dataset of Tibet Plateau boundary. <https://cstr.cn/18406.11.Geogra.tpdcc.270099> (2019).



科技资源标识服务平台
CSTR标识码: 11738

基础信息
中文标题: Dual ligand engagement for noncanonical inflammasome activation
CSTR:11738.14.41590-022-0118-0
作者: Zhang-Hua Yang, Jiahuai Han

引文文献
Semantic Scholar (2)

Comparative study of aluminum speciation on brain-type CK-88
作者: Sjija Liu,Yiyan Ding,Qianqian Yu,Xuerui Wang,Dai Cheng
副刊标题: Journal of inorganic biochemistry
摘要: Brain-type Creative kinase (CK-BB), which has a high affinity on the bioactivity of CK-BB has been studied that the inhibition kit was 0.67 mM, while Al(mg) was 3.81 mM. Fluorescence spectra AlCl3 was closer to the active site of CK-BB. C6 cells were used to cause insufficient local ATP supply in cells which affected the function subsequently activated tcytochrome C (cyt C)/Caspase-3 might be related to the process of Al-induced energy metabolism

Open Access Article

Prevalence, Genetic Homogeneity, and Antibiotic Resistance of Pathogenic *Yersinia enterocolitica* Strains Isolated from Slaughtered Pigs in Bulgaria

by Maya Angelovska ¹, Maya Margaritova Zaharieva ², Lyudmila L. Dimitrova ¹, Tanya Dimova ², Irina Gotova ³, Zoltan Urshev ³, Yana Ilieva ¹, Milla Dobromirova Kaleva ¹, Tanya Chan Kim ¹, Sevda Naydenska ⁴, Zheche Dimirov ³ and Hristo Naydenski ^{1,*}

ORCID人员标识使用效果

¹ The Stephan Angeloff Institute of Microbiology, Bulgarian Academy of Sciences, 26 Akad. G. Bonchev Str., 1113 Sofia, Bulgaria

² Institute of Biology and Immunology of Reproduction, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria

³ LB Bulgarianum Plc., R&D Department, 14 Malashevksa Str., 1000 Sofia, Bulgaria

⁴ University Multiprofile Hospital for Active Treatment, Alexandrovska, Medical University, 1 Georgi Sofiski Str. 1431 Sofia, Bulgaria

* Author to whom correspondence should be addressed.

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、组织架构等方面筛选分类功能

科学数据可信认证

国家基础学科公共科学数据中心

高可信的科学数据服务平台 高质量的科学数据创新服务

了解更多

可信赖科学数据服务平台快速发现

已收录平台 3131 个，覆盖机构 2261 家

Search for Platforms...

平台发现

共查询到 3129 个, 第 1 页, 共 313 页

筛选:

所属国家 (69)

- China (89)
- United States of America (1002)
- Germany (437)
- Canada (354)
- United Kingdom of Great Britain and Northern Ireland (220)

网络安全等级

- 一级 (0)
- 二级 (22)
- 三级 (2)
- 四级 (0)
- 未评测 (3105)

平台类型

- 学科型 (2342)
- 多学科型 (2)
- 机构型 (526)
- 项目型 (0)
- 其他 (259)

View All

查询

为服务平台提供认证、增值服务和影响力提升

国家基础学科公共科学数据中心

National Basic Science Data Center

别名: 中国科学院基础学科公共科学数据中心

完全公开

学科: 物理; 化学; 材料; 植物; 动物; 病毒; 核科学; 信息科学; 交通运输

资源内容类型: 数据集; 期刊论文; 数据论文; 书籍; 标准; 软件

简介: 国家基础学科公共科学数据中心依托中科院计算机网络信息中心, 奉头中科院40年科学数据工作的坚实基础, 联合中科院、教育部、工信部、国防科工局、国家林业局、黑龙江省等下属40余个单位共同建设, 数据资源覆盖17类一级学科, 形成23个主题库, 在线服务数据总量2.29PB, 数据记录近10亿条, 形成完善的标准体系、技术体系和软件栈, 服务体系全面覆盖“数据服务-信息服务-知识服务”...

科学数据可信认证 收录证书

中文名称 (英文名称)

贵平台成为科学数据可信认证收录平台, 特颁发此证书。

CSTR:
上链:

二维码

1 可发现性

- 1.1 使用通用的数据标识体系 ✓
- 1.2 使用自身的数据标识体系 ✓
- 1.3 统一的数据服务平台收录 ✓

2 可访问性

- 2.1 支持用户访问
运行状态: 可访问
- 2.2 服务平台持续访问
历史可访问率: 100.00 %

3 可互操作性

- 3.1 采用通用的数据交换标准 ✓
- 3.2 采用通用的数据操作协议提供数据服务 ✓

4 可用性

- 4.1 采取质量检测措施 ✓
- 4.2 使用作者授权 ✓
- 4.3 使用数据许可 ✓

5 安全性

- 5.1 网络系统符合安全要求 ✓
- 5.2 具有安全与道德审查 ✓

6 持续性

- 6.1 具有稳定的信托机构 ✓
- 6.2 具有专家和技术队伍 ✓
- 6.3 平台具有长期发展方向 ✓

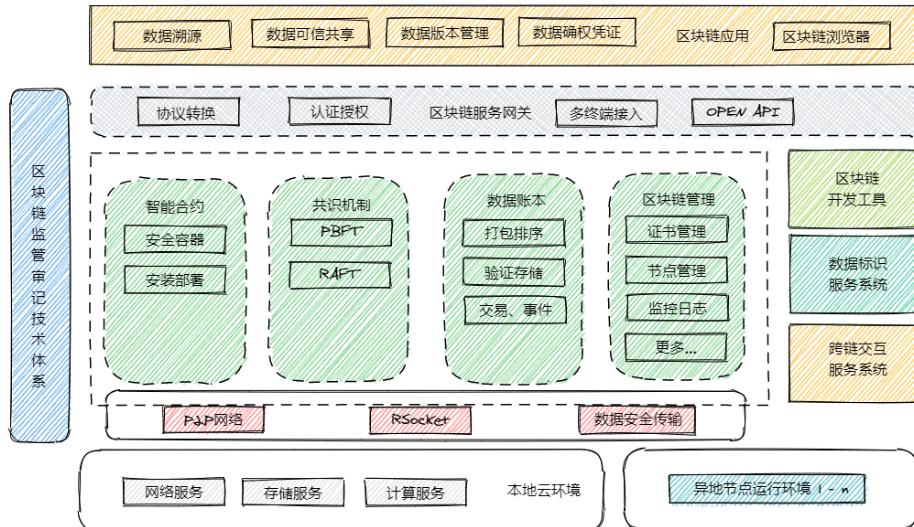
科学数据链-基于区块链的数据安全溯源



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)

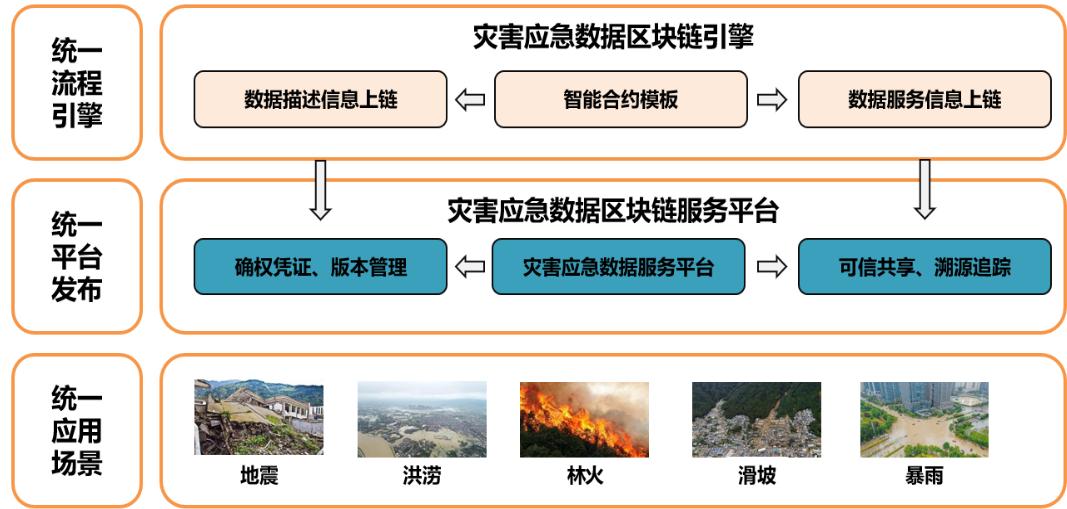
- 以**数据存证、确权、溯源**等应用为需求牵引，为科学数据提供安全、可信的区块链服务
- 基于**自主研发、安全可控的ODC区块链引擎**构建
- 加入科链科学数据链，保证上链数据**安全共享、可信存证、数据追溯、版本管理、确权凭证、使用记录、贡献量统计**等



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

基于区块链的灾害应急数据服务平台



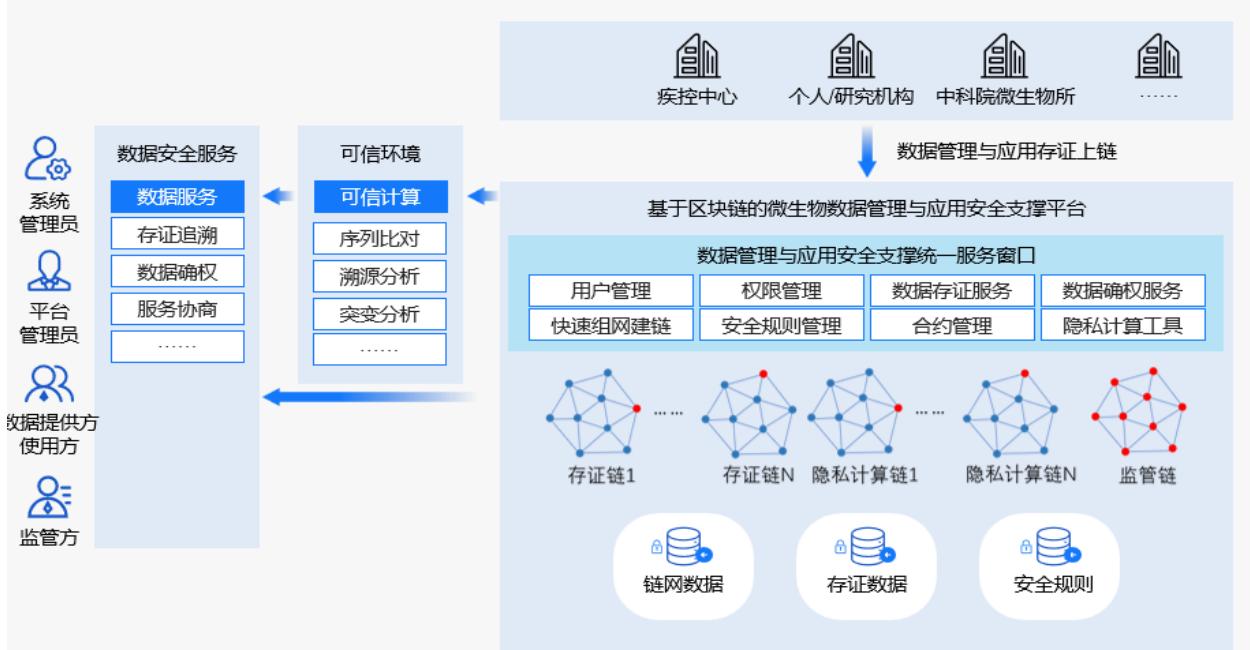
节点信息

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
国家综合观探测数据共享平台节点	国家综合观探测数据共享平台节点	国家综合观探测数据共享平台	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

合約名称	所属子链	数据存储范围	数据类型	创建时间
灾害应急数据合约	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！