开放科学背景下的科学数据开放共享：国家青藏高原科学数据中心的实践

潘小多[1,2]，李新[1,2]，冉有华[3]，郭学军[2]
1. 中国科学院青藏高原研究所国家青藏高原科学数据中心，北京 100101；
2. 中国科学院青藏高原研究所青藏高原地球系统与资源环境国家重点实验室，北京 100101；
3. 中国科学院西北生态环境资源研究院，甘肃 兰州 730000

摘要
介绍了开放科学和开放数据实践活动的概念、内涵和对科学研究的重要性；详细阐述了现阶段开放数据面临的挑战，如数据引用、数据计量、数据互操作和大数据分析等；并以国家青藏高原科学数据中心为例，阐述其在数据引用、数据互操作和大数据分析等开放数据方面的举措和数据共享成效；最后展望了数据中心对开放数据的促进作用。

关键词
开放数据；数据引用；数据计量；数据互操作；大数据分析；地球科学

中图分类号：N37 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022010

Open access of scientific data in the context of open science: the practice of the National Tibetan Plateau Data Center

PAN Xiaoduo[1,2], LI Xin[1,2], RAN Youhua[3], GUO Xuejun[2]
1. National Tibetan Plateau Data Center, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China
2. State Key Laboratory of Tibetan Plateau Earth System and Resources Environment, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China
3. Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China

Abstract
The concept, connotation and importance of open science and open data practice to scientific research were introduced. The challenges faced by open data currently were described in detail, that included data citation, data metrics, data interoperability, and big data analysis. Taking the National Tibetan Plateau Data Center as an example, its measurements and results in data citation, data interoperability and big data analysis were expounded. Finally, the role of data center in promoting open data was prospected.

Key words
open data, data citation, data metrics, data interoperability, big data analysis, earth science

0 引言

开放科学指科学家通过互联网免费分享他们的研究数据、方法、代码、实验室笔记和其他研究过程，以便能够被重复使用和复现，实现协作研究的科学实践，其旨在消除科学研究过程中的访问障碍，使得研究者可共享任何类型的研究成果、资源、方法或工具，促进科学的自由传播，加强科学合作和信息共享，推动科学进程[1-2]。开放科学运动最早可以追溯到17世纪的启蒙运动时期，但是“开放科学”这个术语直到1998年才被史蒂夫·曼提出，当时他注册了域名openscience.com和openscience.org，这标志着开放科学开始步入人们的视野。开放科学代表了一种全新的科学研究范式，它依赖于建立在数字化技术和新型协作工具上的协作型研究和新的知识传播方式。开放科学的理念反映了50年来科学研究的范式转变：以前的标准做法是将研究成果发表在科学出版物上，而现在的趋势是在研究过程的早期阶段就共享和使用所有可用的知识[3]。

开放科学的实践得到越来越多人的认可和采用，比如开放获取的学术论文[4-5]、共享数据[6-8]和开放代码[6,9]等。McKiernan E C等人[1]通过回顾大量文献发现，开放获取学术论文有利于增加科研工作者的学术论文引用、媒体关注、潜在合作者、就业机会和资助机会等。虽然开放数据对于用户和促进科学发展等方面的益处不言而喻，但是科学数据没有像学术论文那样广泛采用开放获取的共享方式，一个重要的原因是研究人员不确定分享科学数据会对他们的职业生涯造成什么影响。开放数据也是欧盟开放科学战略八大目标的首要目标[10]。为抗击新型冠状病毒肺炎（COVID-19）疫情，全球科学出版商取消了访问COVID-19相关研究的限制，研究人员充分认识到开放获取的数据共享对于科学研究的意义，进一步促成了开放科学实践活动。

科学数据是国家科技创新和发展的基础性战略资源，随着大数据时代的到来，科学数据日益呈现出4V特征：体量（volume）、类型（variety）、速度（velocity）和最重要的价值（value），并具有巨大的潜在价值和可开发价值。开放科学重视数据的开放，它不仅倡导论文的开放获取，而且认为论文中的数据开放也是不可或缺的。开放科学数据是开放科学的重要物质基础，强调数据的活用和重用，并把研究数据与论文或论著置于同等甚至更为重要的位置，随着开放数据的持续深入推进，科研期刊要求著作者提交数据可获取声明已成为常态[11]。

科学数据的开放共享已经从全面开放（full and open）原则过渡到目前普遍遵循的FAIR（可发现性（findability）、可获取性（accessibility）、可互操作性（interoperability）、可重用性（reusability））原则[12-13]，并进一步提出了CARE（集体收益（collective benefit）、质量保证（authority to control）、责任（responsibility）、伦理（ethics））原则[14]。FAIR原则强调技术进步，而CARE原则更侧重政策变革，两者相辅相成，体现了大数据时代科学数据共享技术和政策双轮驱动的特征。在我国，为了完善科技资源共享服务体系，推动科技资源向社会开放共享，国务院办公厅在2018年印发了《科学数据管理办法》，明确了数据开放是受政府预算资金资助的研究项目的基本原则。2019年6月，国家青藏高原科学数据中心等20个国家科学数据中心成立，开启了我国科学数据

开放共享的新阶段。目前，中国在科学数据开放共享方面取得了巨大进展，在地学数据共享方面，国家自然科学基金委员会地学领域的重大研究计划、中国科学院的地球大数据科学工程都已成为地学数据开放共享的标杆[15-16]。

上述科学数据开放共享的原则或政策对于促进开放数据是非常有价值的，但它们并没有消除研究人员对于“开放数据可能会给自己的科研工作带来风险”的顾虑，一定程度上影响了科研工作者自下而上自发地开放科学数据的意愿。要实现范式转变，仍需要政府、研究人员和数据中心的积极努力。我国还要在政策、管理、技术和国际化等方面采取更具体的行动，以更大的力度和更多的措施促进科学家共享数据的意愿，提高我国科学数据中心的影响力，推动更加广泛的数据共享[15]。科学数据中心作为数据存储、管理和运营的核心，连接着数据贡献者和数据用户，促进数据贡献者自下而上地开放共享意愿，从而在推动开放数据的实践方面发挥关键作用，但面临的挑战不容小觑。

1 开放数据面临的挑战

开放数据面临的首要挑战是数据引用和数据计量。科学的数据计量和规范化的数据引用能够解决再现性、可靠性和可重用性方面的问题，能够定量化开放数据的贡献，能够提高公开数据所关联文献的引用量[17]，能够为相关机构提供考核依据，从而激发数据贡献者开放共享数据的意愿，进一步促进开放科学和开放数据的实践。数据作为科学发现的重要支持，应被视为合法的和可引用的研究成果[18]，并像学术文献一样被直接引用；如后续有增值数据，原始数据也应被引用，明确原始数据的价值，确保增值数据的可靠性追溯[19]。然而目前大部分数据中心缺乏数据引用信息或者不同数据中心之间缺乏统一的数据引用标准，很难进行追踪计量；对共享数据的计量大部分等同于其关联文章被引用的情况，这不利于对那些没有关联文章的共享数据的评价。因此，数据中心作为数据的重要载体和管理方，应尽量遵循由全球大量数据相关机构共同制定的数据引用原则[18]，开发相应的工具，为共享数据提供数据引用信息（包含数据贡献者、数据集名称、数据制备年份、数据的数字唯一标识符和数据分发机构等），并能根据不同引用方式灵活提供数据引用信息。

同时，传统的期刊影响因子及论文引用量并不能充分反映科研成果的科学、社会、政治和经济效应，开放科学为开发新的科研成果计量方式创造了机遇，有助于激励科研人员自发共享除学术论文外的科研成果，比如科学数据和软件代码等。数据中心需要抓住机遇，加强开放数据的科学计量，开发新一代开放科学计量工具，综合反映科学数据的科研、社会、政治和经济效益。实现这些基本的计量只是第一步，如何实现更科学的数据计量，并合理设计相应的激励机制，还需要更多的研究与探索。

第二个挑战是数据的互操作性。在FAIR原则中，互操作性是体现数据信息增值最大化的核心属性，相较于其他属性，该属性最能激发数据贡献者的数据共享意愿。从宏观上来讲，互操作性是要建立一个被广泛认可的关于数据交换、数据安全和信息传递的规范、标准、方法、过程或实践等准则[20]，从技术、结构、语义和组织等不同层次实现数据互操作的标准化。从数据实体来讲，互操作是能够实现多源异构数据的集成、分析和处理，进而实现大数

据分析和决策的技术和方法，具体而言，即保证：①数据/元数据使用正式、可访问、共享和广泛适用的语言来表示知识；②数据/元数据使用遵循公平原则的词汇表；③数据/元数据包括对其他（元）数据的限定引用[21]。

第三个挑战是数据共享模式从数据仓库到大数据平台的转换。实现数据共享模式从数据仓库到大数据平台的转换是从数据角度支持开放科学的关键，应对这一挑战的核心是建设集数据存储、管理、建模、分析、可视化、决策支持于一体的大数据平台，并将其作为开放科学时代的信息基础设施，实现从地球系统的观测、数据综汇、开放获取、信息提取、知识挖掘到智慧决策的技术贯通。而模型驱动与数据驱动方法的深度结合可能是最大的技术瓶颈，也是最有前景的研究方向[22-23]。此外，兼容传统数据共享模式，提供更加强大的数据搜索引擎、智能数据处理工具，更有效地为用户和机器提供更加友好、智能的服务，也是实现上述技术升级转换的桥梁。

2 开放数据实践

为了应对数据开放共享存在的问题和面临的挑战，国家青藏高原科学数据中心初步开展了一些尝试，包括采用国际标准提供数据引用方式和数据关联文献引用方式，支持数据出版，开发在线大数据分析、模型应用等功能，促进第三极地区科学数据开放共享[24-25]。

具体来讲，国家青藏高原科学数据中心开发了中英文双语数据管理与共享平台，大部分数据采用开放获取方式（其中大部分开放数据实现免登录下载），目的是降低数据下载门槛。但是需要有知识产权保护作为开放获取的前提，国家青藏高原科学数据中心采用以下方式来保障数据作者的知识产权（**图1**）：①为每个自有产权的数据赋予唯一的数字对象标识符（digital object identifier，DOI）和中国科技资源（China science and technology resource，CSTR）标识，体现数据的跟

1. 数据DOI：支持CSTR
   - √ 赋预新的数据DOI/CSTR：10.11888/category.tpdc.ID
   - √ 沿用原数据平台的数据DOI/CSTR
2. 数据协议：默认采用CC BY 4.0协议
3. 数据引用：必引，通过数据引用体现数据的知识产权
4. 文献引用：原创数据建议引用数据作者提供的文献
5. 申请获取与数据保护期：尊重数据作者对数据共享方式的要求
6. 开放获取数据：推动开放获取方式的数据集共享，通过DOI、CC协议、数据引用、文献引用等保障数据的知识产权
7. 论文关联数据：通过DOI、CC协议、数据引用、文献引用等保障数据的知识产权；同时采用免登录预览数据设置，保障数据作者优先发表数据关联论文的权益
8. 新媒体和内容平台：区分原创和转载，原创注明作者，转载注明出处

图1 数据知识产权保护措施

踪价值、引用价值、集成价值和互联价值；②采用知识共享（creative commons，CC）4.0协议，保留作者版权，同时授权他人在协议限定范围内的转载、使用和二次演绎等行为；③建议和鼓励用户进行数据引用和数据关联文献引用，并在数据详情页提供数据引用和数据关联文献引用信息；④秉承数据开放获取的原则，同时兼顾数据作者对特殊数据保护的诉求，可设置不超过两年的数据保护期，或根据数据作者对数据共享需要附加额外条件的要求，设置数据申请审批流程。

国家青藏高原科学数据中心积极申请成为国际重要期刊和组织认证的数据仓储，不断完善数据中心的功能，提升数据中心的共享和服务能力。2020年7月国家青藏高原科学数据中心成为国内首个通过*Nature*旗下*Scientific Data*认证的数据仓储中心。2020年7月国家青藏高原科学数据中心成为美国地球物理学会（American Geophysical Union，AGU）推荐的数据仓储，并成功注册综合性的全球研究数据存储库系统（re3data.org和FAIRsharing）和项目（Enabling FAIR Data），促进了数据中心、国际地球科学领域其他数据中心和研究人员的合作与交流。

在数据互操作方面，国家青藏高原科学数据中心尽量采用地学数据领域广泛认可的标准和规范来减少互操作性障碍，如数据交换服务协议选用开放源代码的网络数据访问协议（open-source project for a network data access protocol，OPeNDAP）和开放地理空间信息联盟（open geospatial consortium，OGC）标准。关于数据层面的互操作性，虽然没有要求数据作者使用特定的格式，但建议数据作者尽可能按照气候和预测（climate and forecast）公约，采纳网络通用数据格式（network common data format，NetCDF）对数据进行编码。国家青藏高原科学数据中心按照谷歌数据搜索引擎的要求，在数据集描述页面添加符合Schema.org标准的元数据信息，使得数据中心的数据能够在谷歌数据搜索引擎中被查询到。

在大数据分析方面，国家青藏高原科学数据中心通过增量集成和自主研发，构建大数据质量控制、自动建模与分析、数据挖掘及交互式可视化的方法库，形成具有高可靠性、高可扩展性、高效性和高容错性的工具箱，实现青藏高原及周边多源异构、多粒度、多时相、长时间序列大数据的协同分析方法的集成和共享，以及高效和在线的大数据分析处理，并通过青藏高原关键地表过程的大数据分析应用示范，打通数据深度挖掘的整体技术链路[26]。国家青藏高原科学数据中心目前包含机器学习、数据同化、参数估计、时间序列分析、高级地统计、数据后处理和因果分析七大类大数据分析方法库，通过方法库的元信息对方法进行管理和智能搜索/推荐，建立代码共享机制，并在GitHub上托管。

目前，国家青藏高原科学数据中心集成了青藏高原及周边科学数据集4 350个（数据量接近172 TB），其中开放获取的科学数据集有2 797个，占比超过64%。自2021年3月以来，国家青藏高原科学数据中心对开放获取的数据实行免登录设置，平均每月数据下载量达1.6万多次，较之前增长了两倍多，大大提升了数据共享服务量。境外用户的数据下载量占比超过35%，随着国家青藏高原科学数据中心国际化建设的进一步推进，国际数据贡献者和数据用户有望进一步增多，从而进一步提升数据中心的国际影响力。截至2021年9月，已有2 800多篇论文使用和引用了国家青藏高原科学数据中心的数据集，用于冰冻圈变化、亚洲水塔变化、生态系统脆弱性评

估、重大工程风险评估和遥感反演评估等研究，为青藏高原地球系统科学研究提供了数据支撑，有效地提高了第三极地区科学数据的共享水平与利用效率，推动了青藏高原及周边地区地球系统的科学研究和前沿创新。

3 结束语

随着以地球观测系统和众源地理信息为典型代表的各类科学数据以指数级速度的持续增长，地球系统科学研究进入了“大数据”时代，科学研究的第四范式——数据密集型科学发现不约而至，开放科学和开放数据是科学发现新范式的一种适应和必然过程，每一位科研工作者都将是开放科学和开放数据的贡献者和受益者。数据中心是开放数据系统中连接决策者、数据贡献者、数据和数据用户的中介机构，可从政策、管理、技术和国际化等方面加强开放数据措施，并让数据贡献者和数据用户受益，形成科学和社会收益的强化反馈。

参考文献:

[1] MCKIERNAN E C, BOURNE P E, BROWN C T, et al. How open science helps researchers succeed[J]. eLife, 2016, 5: e16800.

[2] European Commission. Open innovation, open science, open to the world: a vision for Europe[R]. 2016.

[3] WOELFLE M, OLLIARO P, TODD M H. Open science is a research accelerator[J]. Nature Chemistry, 2011, 3(10): 745–748.

[4] SWAN A. The open access citation advantage: studies and results to date[R]. 2010.

[5] BJÖRK B C. The hybrid model for open access publication of scholarly articles: a failed experiment?[J]. Journal of the American Society for Information Science and Technology, 2012, 63(8): 1496–1504.

[6] STODDEN V, GUO P, MA Z. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals[J]. PLoS One, 2013, 8(6): e67111.

[7] HEIMSTÄDT M, SAUNDERSON F, HEATH T. From toddler to teen: growth of an open data ecosystem[J]. eJournal of eDemocracy and Open Government, 2014, 6(2): 123–135.

[8] MICHENER W K. Ecological data sharing[J]. Ecological Informatics, 2015, 29: 33–44.

[9] SHAMIR L, WALLIN J F, ALLEN A, et al. Practices in source code sharing in astrophysics[J]. Astronomy and Computing, 2013, 1: 54–58.

[10] RAMJOUÉ C. Towards open science: the vision of the European commission[J]. Information Services & Use, 2015, 35(3): 167–170.

[11] 吴建中. 推进开放数据 助力开放科学[J]. 图书馆杂志, 2018, 37(2): 4–10.
WU J Z. Promoting open data for open science[J]. Library Journal, 2018, 37(2): 4–10.

[12] WILKINSON M D, DUMONTIER M, AALBERSBERG I J, et al. The FAIR guiding principles for scientific data management and stewardship[J]. Scientific Data, 2016, 3: 160018.

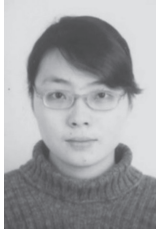
[13] STALL S, YARMEY L, CUTCHER-GERSHENFELD J, et al. Make scientific data FAIR[J]. Nature, 2019, 570(7759): 27–29.

[14] CARROLL S R, GARBA I, FIGUEROA-RODRÍGUEZ O L, et al. The CARE principles for indigenous data governance[J]. Data Science Journal, 2020, 19(1): 43.

[15] LI X, CHENG G D, WANG L X, et al. Boosting geoscience data sharing in China[J]. Nature Geoscience, 2021, 14(8):

541–542.
[16] LI X, NAN Z T, CHENG G D, et al. Toward an improved data stewardship and service for environmental and ecological science data in West China[J]. International Journal of Digital Earth, 2011, 4(4): 347–359.
[17] PIWOWAR H A, VISION T J. Data reuse and the open data citation advantage[J]. PeerJ, 2013, 1: e175.
[18] Data Citation Synthesis Group. Joint declaration of data citation principles[S]. 2014.
[19] COUSIJN H, KENALL A, GANLEY E, et al. A data citation roadmap for scientific publishers[J]. Scientific Data, 2018, 5: 180259.
[20] PONS M. Arrangements for a successful interoperability workshop[C]// Proceedings of 2008 AIChE Annual Meeting. [S.l.:s.n.], 2008.
[21] GUIZZARDI G. Ontology, ontologies and the "I" of FAIR[J]. Data Intelligence, 2020, 2(1–2): 181–191.
[22] BAUER P, DUEBEN P D, HOEFLER T, et al. The digital revolution of earth-system science[J]. Nature Computational Science, 2021, 1(2): 104–113.
[23] REICHSTEIN M, CAMPS-VALLS G, STEVENS B, et al. Deep learning and process understanding for data-driven earth system science[J]. Nature, 2019, 566(7743): 195–204.
[24] PAN X D, GUO X J, LI X, et al. National Tibetan Plateau Data Center: promoting earth system science on the third pole[J]. Bulletin of the American Meteorological Society, 2021, 102(11): 2062–2078.
[25] CHEN F H, DING L, PIAO S L, et al. The Tibetan Plateau as the engine for Asian environmental change: the Tibetan Plateau earth system research into a new era[J]. Science Bulletin, 2021, 66(13): 1263–1266.
[26] 李新, 潘小多, 郭学军, 等. 大数据系统助力青藏高原和泛第三极地球系统科学研究[M]// 中国科研信息化蓝皮书2020. 北京: 电子工业出版社, 2020.
LI X, PAN X D, GUO X J, et al. Big data promotes the Tibetan Plateau and Pan-Third Pole earth system science[M]// China's e-science blue book 2020. Beijing: Publishing House of Electronics Industry, 2020.

作者简介



潘小多（1978– ），女，博士，中国科学院青藏高原研究所研究员、博士生导师，主要从事区域气候变化、数据同化、数据集成和大数据分析等研究，在*Bulletin of the American Meteorological Society*、*Journal of Geophysical Research*和《高原气象》等期刊上发表学术论文60多篇。自2018年以来，在国家青藏高原科学数据中心负责科学数据集成与服务方面的工作，已申请3项国家发明专利和1项计算机软件著作权。

李新（1969– ），男，博士，中国科学院青藏高原研究所研究员、博士生导师，国家杰出青年科学基金获得者。发展了我国大尺度陆面数据同化系统及高分辨率的流域尺度陆面水文数据同化系统，组织实施了"黑河综合遥感联合试验"和"黑河生态水文遥感试验"。获甘肃省自然科学奖一等奖及中国科学院杰出科技成就奖。已发表学术论文400多篇（其中SCI收录260多篇），论文总引用17 000多次。

冉有华（1980– ），男，博士，中国科学院西北生态环境资源研究院副研究员，主要从事冰冻圈生态水文遥感与模型、遥感产品真实性检验等相关研究工作，已发表学术论文70多篇（其中SCI收录30多篇），论文总引用3 000多次。

郭学军（1977– ），男，博士，中国科学院青藏高原研究所研究员级高级工程师，主要从事科研信息化和科学大数据等方面的工作，主持中国科学院战略性先导科技专项（A类）“泛第三极环境变化与绿色丝绸之路建设”子课题、中国科学院信息化专项等多个项目。

收稿日期： 2021-10-09

通信作者： 李新，xinli@itpcas.ac.cn

基金项目： 青藏高原地球系统基础科学中心资助项目（No.41988101）；中国科学院战略性先导科技专项（No.XDA20060600）

Foundation Items: Basic Science Center for Tibetan Plateau Earth System (No.41988101), The Strategic Priority Research Program of the Chinese Academy of Sciences (No.XDA20060600)

“国家科学数据中心”联合专刊总目录

方向一：科学数据治理

作者	题目	期刊	卷期	DOI
胡皓，齐法制，孙晓康，罗齐	高能同步辐射光源科学数据管理策略研究与应用	《大数据》	第8卷第1期	10.11959/j.issn.2096-0271.2022001
许琦，邹自明，袁雅琴，胡晓彦，佟继周，马文臻	科技计划项目数据管理过程模型	《大数据》	第8卷第1期	10.11959/j.issn.2096-0271.2022002
高飞，周国民，满芮	基于生命周期理论的农业科学数据中心化管理模式	《大数据》	第8卷第1期	10.11959/j.issn.2096-0271.2022003
李茹姣，张欣，宋述慧，王彦青，邹东，肖景发，赵文明，章张，鲍一明	基因组科学数据的安全管理与应用	《大数据》	第8卷第1期	10.11959/j.issn.2096-0271.2022004
陈异凡，闫桑，杨亚超，胡林，樊景超，张翔鹤，周国民	我国农业科学数据共享协议	《大数据》	第8卷第1期	10.11959/j.issn.2096-0271.2022005
智峰，田锋，赵若凡	计量科学大数据分级分类	《大数据》	第8卷第1期	10.11959/j.issn.2096-0271.2022006
孙苗，王子珂，童心，符昱，王漪，康林冲，姜晓轶	典型海洋环境观测数据产品应用现状及对我国的启示	《大数据》	第8卷第1期	10.11959/j.issn.2096-0271.2022007
苏文，张黎，郭学兵，何洪林，唐新斋，任小丽	生态系统长期观测数据产品体系	《大数据》	第8卷第1期	10.11959/j.issn.2096-0271.2022008
曹乔卓然，陈祖刚，李国庆，李静	科学数据中心资源和用户访问控制体系	《大数据》	第8卷第1期	10.11959/j.issn.2096-0271.2022009
潘小多，李新，冉有华，郭学军	开放科学背景下的科学数据开放共享：国家青藏高原科学数据中心的实践	《大数据》	第8卷第1期	10.11959/j.issn.2096-0271.2022010

方向二：科学大数据在采集、汇交、保存、安全、分析、挖掘、呈现等方面的理论与前沿技术

作者	题目	期刊	卷期	DOI
张耀南	数据工程学建设思考与实践	《数据与计算发展前沿》	第4卷第1期	10.11871/jfdc.issn.2096-742X.2022.01.001
韩军，樊东卫，陶一寒，许允飞，李珊珊，米琳莹，李长华，崔辰州	FAST科学观测项目管理信息系统	《数据与计算发展前沿》	第4卷第1期	10.11871/jfdc.issn.2096-742X.2022.01.002
胡庆宝，郑伟，王佳荣，汪璐，颜田	高能物理科学数据中心智能运维系统	《数据与计算发展前沿》	第4卷第1期	10.11871/jfdc.issn.2096-742X.2022.01.003
张翔鹤，闫桑，樊景超	多源异构作物组学数据融合方法研究——以高粱为例	《数据与计算发展前沿》	第4卷第1期	10.11871/jfdc.issn.2096-742X.2022.01.004
唐新斋，陈昕，何洪林，郭学兵，苏文，谢传节，沈志宏，张黎，任小丽，侯艳飞，刘峰	新一代“生态网络云”大数据平台的设计与实现	《数据与计算发展前沿》	第4卷第1期	10.11871/jfdc.issn.2096-742X.2022.01.005
卢逸航，李国庆，陈祖刚	科学数据中心间互操作模式研究	《数据与计算发展前沿》	第4卷第1期	10.11871/jfdc.issn.2096-742X.2022.01.006
李进，陈祖刚，李国庆	对地观测知识枢纽研究进展	《数据与计算发展前沿》	第4卷第1期	10.11871/jfdc.issn.2096-742X.2022.01.007
石京燕，黄秋兰，汪璐，李海波，杜然，姜晓巍，胡庆宝，郑伟，闫晓飞，张玄同	国家高能物理科学数据中心分布式数据处理平台	《数据与计算发展前沿》	第4卷第1期	10.11871/jfdc.issn.2096-742X.2022.01.008
张喆，杨海彦，王海雪，王格，何战科，徐永亮，孙保琪，杨旭海	国际GNSS监测评估系统数据采集与服务的研究及应用	《数据与计算发展前沿》	第4卷第1期	10.11871/jfdc.issn.2096-742X.2022.01.009

方向三：数据标准规范、优质数据集出版

作者	题目	期刊	卷期	DOI
孟晓阳，王佳权，苑尚博，宋佳军，马启明	2020年基于VLF/LF三维闪电定位系统的全国闪电数据集	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0059.zh
田锋，智峰，赵若凡	社会公用计量标准数据集	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0062.zh
李英勇	中药材化学成分的晶体结构数据集	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0052.zh
李文杰，王江宁，卜翠萍，葛斯琴，林聪田，韩艳，纪力强	基于动物志的粉蝶形态特征数据集	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0063.zh
陈中，郑为民，陈肖，薛岩松	2007—2020中国探月工程VLBI测量数据集及其应用	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0068.zh
肖翠，金冬梅，李颖超，李晓京，张路杨，郑柏岩，吉小冬，林秦文	云南漾濞石门关景区动植物资源数据集	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0064.zh
郭学兵，唐新斋，苏文，何洪林	生态系统要素长期观测（EcoLTO）数据产品规范研制	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0058.zh
张乃静，肖云丹，侯瑞霞，魏胜蓉，纪平	三北工程区生态系统土壤保持能力评估数据集（2000—2020年）	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0073.zh
肖云丹，张乃静，王俊荣，侯瑞霞，魏胜蓉，纪平	基于水资源承载力的华北地区降水与地下水要素数据集（2005—2016年）	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0076.zh
白燕，杨雅萍，孙九林	黄河流域250 m分辨率植被生长季时空演变数据集（2000—2020年）	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0045.zh
徐洋，杨雅萍	1982—2020年中国5 km分辨率逐月NDVI数据集	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0041.zh
严涛，金佳鑫，朱青松，刘颖	1992—2018年中国及其毗邻地区基于植被功能类型的土地覆盖与香农多样性指数数据集	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0043.zh
陈逸聪，邵华，李杨，戴玲	2015年长三角地区30 m土地覆被融合数据	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0040.zh
刘颖，周士杰，金佳鑫，严涛	基于“两叶”模型的2001—2016年贵州省LAI与APAR数据集	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0047.zh
赵秀宽，宁百齐，胡连欢，刘立波，李国主，解海永，李凤琴，杨敏	1960年武汉站电离层测高仪数据集	《中国科学数据（中英文网络版）》	第7卷第1期	10.11922/11-6035.csd.2021.0065.zh

续表

作者	题目	期刊	卷期	DOI
孙定中, 马俊才	生物数据的标准化与微生物数据标准的发展	《中国科学数据(中英文网络版)》	第7卷第1期	10.11922/11-6035.csd.2021.0070.zh
袁媛, 陈雷	IDADP-葡萄病害识别研究图像数据集	《中国科学数据(中英文网络版)》	第7卷第1期	10.11922/11-6035.csd.2021.0077.zh

方向四：科学数据资源管理研究、精选数据集案例、优秀服务案例、开放共享最佳实践

作者	题目	期刊	卷期	DOI
苗晨, 张连翀, 李国庆, 曾庆双, 李静, 夏俊士	基于开放科学的全球重大自然灾害数据应急援助机制研究和实践	《中国科技资源导刊》	第54卷第1期	10.3772/j.issn.1674-1544.2022.01.001
胡良霖, 朱艳华, 李坤, 胡泊, 王璐, 高瑜蔚, 李国庆	科学数据伦理关键问题研究	《中国科技资源导刊》	第54卷第1期	10.3772/j.issn.1674-1544.2022.01.002
曾庆双, 张连翀, 李国庆, 郭志斌	基于区块链的灾害应急遥感用户信息共享	《中国科技资源导刊》	第54卷第1期	10.3772/j.issn.1674-1544.2022.01.003
曾珊, 陈刚, 齐法制, 张红梅, 李海波, 李亚康, 田浩来	高能物理科学数据服务与应用	《中国科技资源导刊》	第54卷第1期	10.3772/j.issn.1674-1544.2022.01.004
王寅初, 李文军, 任承钢, 刘正一, 秦松	海洋生物产品服务平台的构建与应用	《中国科技资源导刊》	第54卷第1期	10.3772/j.issn.1674-1544.2022.01.005
李珊珊, 樊东卫, 崔辰州, 何勃亮, 陶一寒, 霍志英, 米琳莹, 罗阿理, 陈建军, 侯文, 孔啸, 李荫碧, 郭炎鑫, 李双, 李长华, 许允飞, 韩军, 杨丝丝, 杨涵溪, 赵永恒	LAMOST天体光谱数据开放共享的回顾与展望	《中国科技资源导刊》	第54卷第1期	10.3772/j.issn.1674-1544.2022.01.006
康建芳, 张耀南, 王家耀, 贾泽祥, 韩立钦, 刘春, 敏玉芳, 李红星, 吴亚敏, 张彩荷	黄河流域生态保护与高质量发展体系化科学数据建设与实践	《中国科技资源导刊》	第54卷第1期	10.3772/j.issn.1674-1544.2022.01.007
娄洪伟, 周影, 吴昊轩, 盛磊	光学技术数据库对光学设计软件的支持	《中国科技资源导刊》	第54卷第1期	10.3772/j.issn.1674-1544.2022.01.008
安波	文字知识图谱构建及应用	《中国科技资源导刊》	第54卷第1期	10.3772/j.issn.1674-1544.2022.01.009
纪珍, 佟继周, 胡晓彦, 邹自明, 马福利, 熊森林	空间科学数据产品组织模型的应用研究	《中国科技资源导刊》	第54卷第1期	10.3772/j.issn.1674-1544.2022.01.010
袁雅琴, 胡晓彦, 佟继周, 邹自明	大数据开放背景下的我国空间科学数据出版实践	《中国科技资源导刊》	第54卷第1期	10.3772/j.issn.1674-1544.2022.01.011
何洪林, 陈智, 张黎, 任小丽, 何念鹏, 贾彦龙, 王秋凤, 郭学兵, 苏文, 唐新斋, 葛蓉, 牛忠恩, 朱先进, 张心昱, 高扬, 朱剑兴, 常清青, 于贵瑞	基于模型数据融合的我国陆地生态系统碳氮水循环研究应用	《中国科技资源导刊》	第54卷第1期	10.3772/j.issn.1674-1544.2022.01.012

2020 年《大数据》高被引论文 Top10

排名	论文信息
1	邓建国, 张素兰, 张继福, 等. 监督学习中的损失函数及应用研究[J]. 大数据, 2020, 6(1): 60–80. DENG J G, ZHANG S L, ZHANG J F, et al. Loss function and application research in supervised learning[J]. Big Data Research, 2020, 6(1): 60–80.
2	叶雅珍, 刘国华, 朱扬勇. 数据资产化框架初探[J]. 大数据, 2020, 6(3): 3–12. YE Y Z, LIU G H, ZHU Y Y. An initial exploration on framework of data assetization[J]. Big Data Research, 2020, 6(3): 3–12.
3	李雨霏, 刘海燕, 闫树. 面向价值实现的数据资产管理体系构建[J]. 大数据, 2020, 6(3): 45–56. LI Y F, LIU H Y, YAN S. Construction of a value-oriented realization of data asset management system[J]. Big Data Research, 2020, 6(3): 45–56.
4	王健宗, 孔令炜, 黄章成, 等. 联邦学习算法综述[J]. 大数据, 2020, 6(6): 64–82. WANG J Z, KONG L W, HUANG Z C, et al. Research review of federated learning algorithms[J]. Big Data Research, 2020, 6(6): 64–82.
5	戴炳荣, 闭珊珊, 杨琳, 等. 数据资产标准研究进展与建议[J]. 大数据, 2020, 6(3): 36–44. DAI B R, BI S S, YANG L, et al. Research status quo and suggestions on data assets standardization[J]. Big Data Research, 2020, 6(3): 36–44.
6	杨孟辉, 杜小勇. 政府大数据治理: 政府管理的新形态[J]. 大数据, 2020, 6(2): 3–18. YANG M H, DU X Y. Big data governance in governments: a new form of the government administration[J]. Big Data Research, 2020, 6(2): 3–18.
7	夏大文, 王林, 张乾, 等. 大数据应用技术课程教学改革与实践[J]. 大数据, 2020, 6(4): 115–124. XIA D W, WANG L, ZHANG Q, et al. Teaching reform and practice of big data application technology course[J]. Big Data Research, 2020, 6(4): 115–124.
8	臧根林, 王亚强, 吴庆蓉, 等. 智慧城市知识图谱模型与本体构建方法[J]. 大数据, 2020, 6(2): 96–106. ZANG G L, WANG Y Q, WU Q R, et al. Model and construction method of the ontology of knowledge graph of smart city[J]. Big Data Research, 2020, 6(2): 96–106.
9	刘彦松, 夏琦, 李柱, 等. 基于区块链的链上数据安全共享体系研究[J]. 大数据, 2020, 6(5): 92–105. LIU Y S, XIA Q, LI Z, et al. Research on secure data sharing system based on blockchain[J]. Big Data Research, 2020, 6(5): 92–105.
10	董祥千, 郭兵, 沈艳, 等. 基于利润最大化的数据资产价值评估模型[J]. 大数据, 2020, 6(3): 13–20. DONG X Q, GUO B, SHEN Y, et al. Data assets value evaluation model based on profit maximization[J]. Big Data Research, 2020, 6(3): 13–20.

数据来源：中国知网；统计日期：2022-01-04

·学术科普·

漫威电影中的智能大脑

王元卓　中国科学院计算技术研究所
陆　源　北京科技大学
崔原豪　北京邮电大学

科幻电影通常以科学为基础展开叙事，以合理的科学推理推动精彩的故事场景。因此科幻电影大多发生在未来，而常见的主题往往是太空飞船、机器人、时空穿越，以及人工智能等。

从2008年的《钢铁侠》开始，漫威宇宙给我们带来了20多部经典的漫威电影，介绍了美国队长、绿巨人、蜘蛛侠、雷神等几十位超级英雄。相信许多观众跟我一样喜欢钢铁侠，钢铁侠的战甲在漫威电影中是科技感的体现，拥有不可动摇的地位。从第一部《钢铁侠》开始出现的Mark1战甲，到《复仇者联盟4》中的纳米战甲2.0，每一代新战甲都惊艳了观众，当然，还有它的前后4代智能大脑系统，如图1所示。

智能管家“贾维斯”最早出现在2008年第一部《钢铁侠》电影里，他的主要作用是为钢铁侠托尼提供生活上的帮助、收集信息和反馈信息。贾维斯并不会直接为托尼在战争中提供实质性的帮助，它更像一个全能的情报和信息处理中心。虽然贾维斯没有固定形象，但这不妨碍它成为钢铁侠最经典的一款智能大脑系统。Meta（原Facebook）CEO马克·扎克伯格就曾经用150个小时制作出一个能进行简单生活管理的智能大脑贾维斯，它可以根据

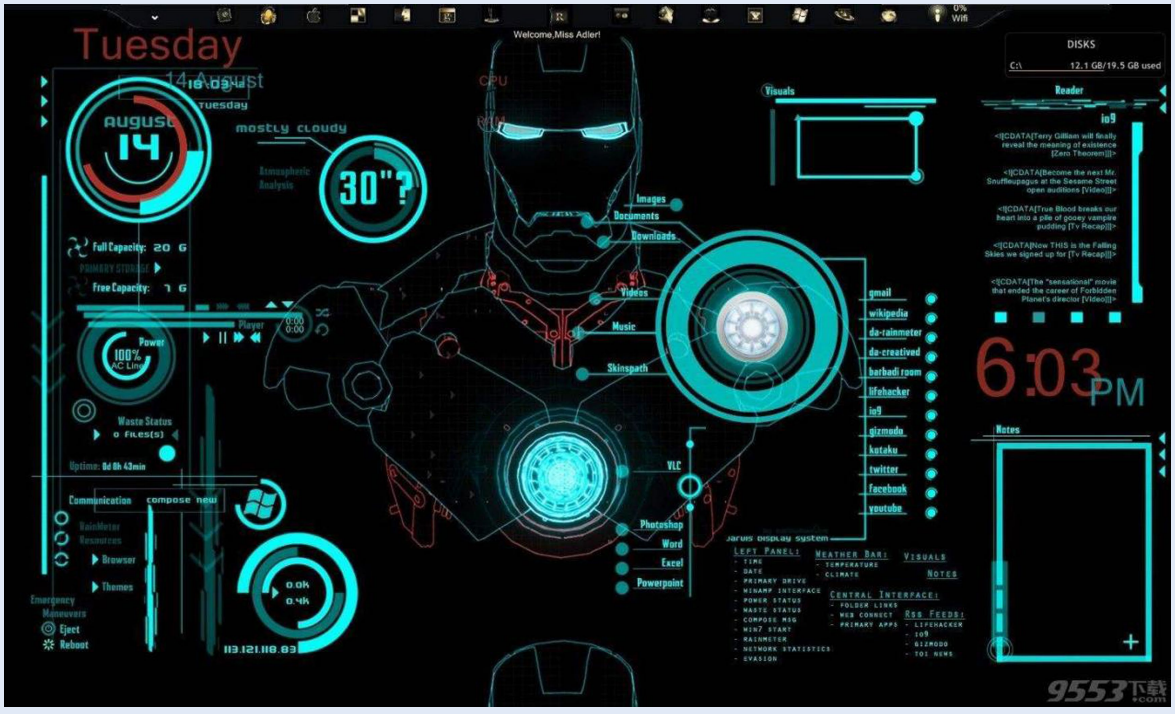


图1　《钢铁侠》中的智能管家

个人喜好播放音乐、为访客开门、烤吐司片，或者只进行工作学习日程提醒等基本的智能互动。

在电影《复仇者联盟2：奥创纪元》中，贾维斯成为幻视之后，钢铁侠启用了第二套人工智能作战系统——星期五。星期五不仅继承了贾维斯的管家模式，在作战上也有了极大的升级。在电影《美国队长3：内战》中，星期五可以自助扫描美国队长的作战模式，模拟对战方案，这使得美国队长处处受制。

第三套智能大脑系统维罗妮卡是专属于反浩克装甲的一套作战系统，只要启动维罗妮卡，反浩克装甲就会从专属地球轨道卫星中脱离。据导演韦登所说，维罗妮卡的灵感来自美国漫画《阿尔奇》。高富帅阿尔奇有两个存在感情纠葛的女友，一个叫维罗妮卡，另一个叫贝蒂。而绿巨人班纳之前的爱人就叫贝蒂，因此为了打败班纳而制造的装甲就被叫作维罗妮卡。

伊迪丝可以说是所有智能大脑系统中最完善的一套了。作为托尼留给蜘蛛侠彼得·帕克的遗产，伊迪丝最早出现在电影《蜘蛛侠：英雄远征》中。她可以操控无人机作战以保护使用者，以及反馈信息情报等，这些功能几乎整合了前面版本所有的优势，在战争中起到了决定性的作用。在这个名字里，托尼也藏了一句话：“Even dead，I’m the hero”，翻译过来就是“尽管我死了，但我仍是英雄”。

讲了很多钢铁侠中的智能大脑系统，很多人觉得这些技术就在我们眼前，其实并不是。一个能力强大的“智能大脑”涉及从大数据的感知到知识图谱的构建，从存储管理到分析挖掘，从各种智能模型再到便捷人机交互的全面支持，我们目前的科技水平离科幻电影还有较大的距离。当然，未来如果真的能实现完善的智能大脑系统，如图2所示，我们可以根据现有技术做一些简单的猜测。

首先，比较完善的智能大脑系统要有对外界大数据的感知能力。对于具有智能的计算

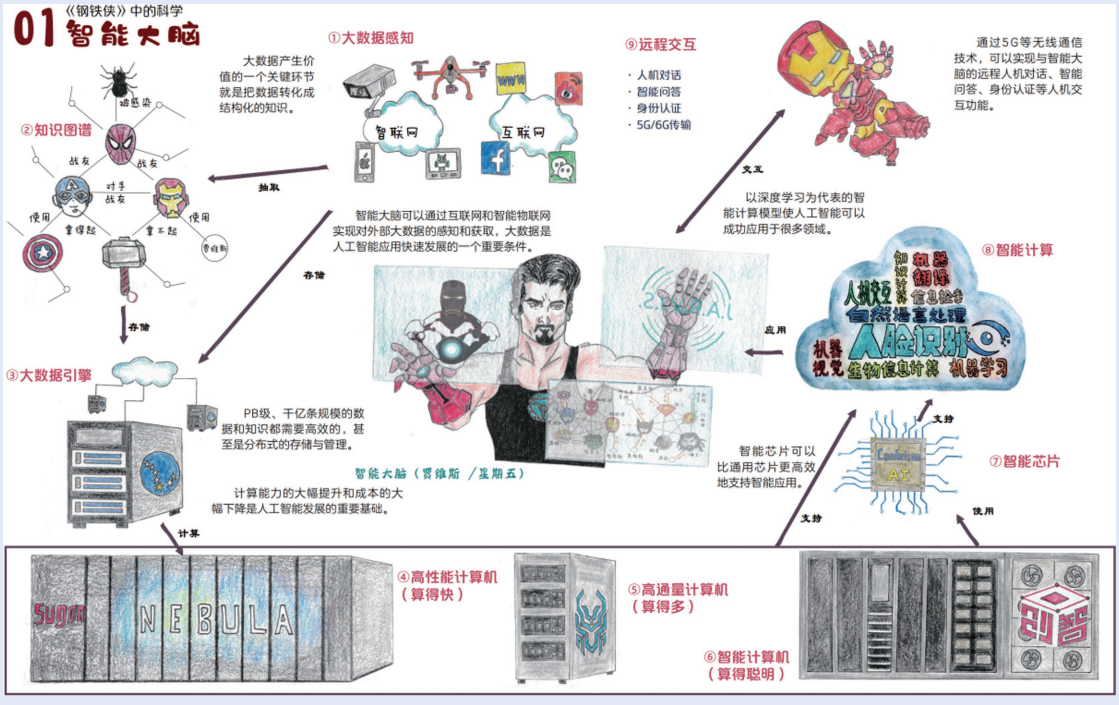


图 2 智能大脑讲解图

（选自《科幻电影中的科学：科学家奶爸的AI手绘》）

设备，它的任务执行和知识学习都离不开大数据处理。而大数据的感知与获取主要通过互联网和智能物联网两个途径来实现。在万物互联的时代，人们预测，到2025年全世界物端连接设备将达到754亿台，大约是2015年的5倍。全世界在互联网上每天有50亿次在线搜索。2025年全世界每年产生的数据将从2018年的33 ZB增长到175 ZB。那么175 ZB的数据到底有多大呢？1 ZB相当于1.1万亿GB，如果把175 ZB全部存在DVD光盘中，那么DVD光盘叠加起来的高度将是地球和月球距离的23倍（月球距离地球最近距离约为39.3万千米），或者绕地球222圈（一圈约为4万千米）。如果按平均网速25 Mbit/s计算，一个人下载完这175 ZB的数据需要18亿年。

这些大数据通常体量大、变化快、模态多样、价值利用率低，让它们产生大价值的一个关键环节就是知识抽取，也就是把大数据转化成结构化的知识。这一步骤，我们通常使用一种叫作“知识图谱”的技术来完成。我们平常传输的数据大部分只是信息，是一种外部的客观事实。而知识，通常是对外部客观规律的归纳和总结。知识图谱本质上是一种揭示知识实体之间关系的语义网络。

为了完成大数据处理，我们需要用到很多类似知识图谱的前沿计算机技术，我们通常将这些用来做数据处理的软件系统和相关算法叫作大数据处理引擎。它可以集数据采集、集成、加工、建模、服务等能力于一体，可以覆盖大数据全生命周期的数据处理能力，从而为各类业务提供全流程、全方位的支撑。现在，先进的大数据引擎通过图结构数据压缩和轻量级并行处理技术，可以实现十亿级规模数据查询的秒级响应。

在拥有数据和领域知识的基础上，就要针对不同的应用类型选择适合的计算机。比如复杂的科学计算最要紧的是计算速度，这就需要选择高性能计算机，实现“算得快”的目标。而高性能计算机主要适用于任务单一、负载变化不频繁、单个任务计算量大，以及计算局部性好的科学与工程计算类应用。

再比如互联网大数据等应用最要紧的问题是大量用户群体带来的高并发量，需要系统能够保证在特定时间内同时处理很多数据请求，这就需要使用高通量计算机来实现“算得多”的目标。高通量计算机的核心特点是保障整体计算过程的并发性、实时性和确定性。

如果是大量智能应用场景，就需要能够实现“算得更聪明”的设备，这类计算机通常被叫作智能计算机，针对不同的应用场景，其可以让人们的决策和反应更快、更准。在智能计算设备上，需要搭载智能芯片，与通用芯片相比，智能芯片可以更高效地支持智能处理任务。比如AI芯片在图像识别、自动翻译、视频分析等智能任务上，比传统芯片速度快100倍，而耗电量却只有传统芯片的1%。

在这些不同特点的计算机上可以运行以深度学习为代表的智能计算模型，使人脸识别、生物信息计算、人机交互、机器翻译、知识计算、信息检索等智能计算技术，以及人工智能从科幻电影走向现实。

在经典的科幻电影中，人工智能可能是最激动人心的体验，这种技术的视觉展现甚至可能超过电影本身，升华电影内核。人类对人工智能的梦想从未停止过。我们也在等待人工智能发展带来的更多便利。虽然还有很长的路要走，但是我们相信，就像电影中无数次描绘的那样，随着人工智能的发展，越来越多的人工智能产品将出现在我们的生活中。或许终有一天，我们能与人工智能培养真实情感，与之和谐共存。

·行业资讯·

《网络数据安全管理条例》内部研讨会成功召开

2021年11月26日上午，360天枢智库、大数据协同安全技术国家工程实验室、《大数据》期刊共同举办了《网络数据安全管理条例（征求意见稿）》（以下简称《条例》）内部研讨会。

会议邀请公安部第三研究所网络安全法律研究中心、中国信息通信研究院互联网法律研究中心、中国现代国际关系研究院科技与网络安全所、北京师范大学网络法治国际中心、汉坤律师事务所、江苏竹辉律师事务所等单位的专家学者，重点讨论了《条例》关于个人信息处理和重要数据安全管理的条款，并对互联网平台企业的合规建设提出了建议。

1 关于个人信息处理的条款

汉坤律师事务所律师解石坡对照《个人信息保护法》，分析了《条例》里特别值得企业关注的内容，尤其是个人信息处理在“告知–同意”基本处理规则方面，需要取得单独同意的6种情形。

江苏竹辉律师事务所律师原浩对比《数据安全管理办法（征求意见稿）》，解读了《条例》中爬虫条款的规定，由此分析了企业可能面临的新问题，包括司法实践的冲突、与竞争法的协调、与政府信息公开的协调、行业自律的价值和作用如何体现。

2 关于重要数据安全管理的条款

来自公安部第三研究所网络安全法律研究中心和中国现代国际关系研究院科技与网络安全所的专家们详细解释了《条例》涉及的基本概念，对《条例》第三部分重要数据处理的条款进行了专业细致的解读。并且，就如何在实践中落实国家的立法和政策指导，提出主动兼顾数据安全的动态性、平衡性和相对性的原则。

3 对互联网平台企业的建议

对于涉及数据处理的互联网平台企业，北京师范大学网络法治国际中心执行主任、中国互联网协会研究中心副主任吴沈括指出，从整体来看，《条例》文本的总体框架已经指出了企业数据大合规的路径。为此，企业需要在技术管理、组织架构和价值生态等多个层面，深度改造发展模式。尔后，专家们就如何加强企业数据合规管理体系建设提出了宝贵建议。

4 交流与讨论

在自由讨论环节，中国网络空间研究院网络安全研究所所长姜伟线上介绍了《条例》的起草背景，中国信息通信研究院互联网法律研究中心主任工程师何波从立法技术的角度，对《条例》涉及的几个核心概念问题做了深入分析。

最后，360天枢智库、大数据协同安全技术国家工程实验室、《大数据》期刊的多位负责人和资深研究员，围绕专家意见和建议进行了交流和讨论。