

数据论文的写作与发表

孔丽华、习妍
中国科学院计算机网络信息中心
《中国科学数据》编辑部
2018.5.10 · 北京



中国医学科学院 医学信息研究所
May.10, 2018 Beijing

• 提 纲

- 一．为什么要开放数据、出版数据
- 二．科学数据与医学大数据
- 三．数据论文与数据期刊
- 四．如何撰写和发表数据论文——以《中国科学数据》为例

- **数据的可重复验证性 是科学的一个基本原则**
- 一项新发现被接受的基本前提是基于数据的可重复验证性。
- 2014年，学术界最轰动STAP细胞“学术造假”事件
- 2017年8月3日，《自然·生物技术》杂志在其网站上刊登出此前饱受争议的韩春雨团队的撤稿声明。围绕在韩春雨身上持续一年多有关NgAgo-gDNA基因编辑问题的争议，终于，以他的主动撤稿而告一段落了。



小保方晴子



河北科技大学
副教授韩春雨

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

Journal	Impact Factor	Policy of Required Depositions for Types of Data			Policy of Provision of Materials and Methods			Full data available
		Micronutrient	Nutrient	Protein	Macromolecular	Materials upon request	Proteins upon request	
New England Journal of Medicine	51.589							0
Journal of the American Medical Association	21.881							0
Lancet	28.638							0
Nature Medicine	15.702							0
Science	26.277							0
Nature Immunology	20.218							0
Nature Biotechnology	25.547							1
JAMA	21.648							0
Nature Biotechnology	19.782							0
Immunity	19.266							0
Journal of Clinical Investigation	16.915							0
Archives of General Psychiatry	13.976							0
Journal of the National Cancer Institute	16.678							0
Nature Neuroscience	16.864							0
Journal of Neurological Medicine	15.516							0
Annals of Internal Medicine	15.468							0
Journal of Clinical Oncology	16.478							0
Nature Methods	14.795							0
Genes and Development	14.785							0
Cell	13.661							0
PNAS	13.661							0
PNAS	13.661							0
Molecular Cell	13.156							0
Circulation	12.765							0
Developmental Cell	12.638							0
Immunology	12.617							0
Cancer Research	11.224							0
American Journal of Human Genetics	11.002							0
Human Molecular Genetics	11.002							0
Journal of the American College of Cardiology	11.004							0
Journal of the American College of Cardiology	11.004							0
Hepatology	10.374							0
Cancer Biology	10.370							0
PNAS	9.725							0
British Medical Journal	9.721							0
Genetics Research	8.671							0
Plant Cell	8.651							0
Nature Letters	8.627							0
Journal of Cell Biology	8.609							0
Genetics and Cellular Physiology	8.609							0
PNAS	8.316							0
American Journal of Psychiatry	8.177							0
American Journal of Respiratory and Critical Care Medicine	8.130							0
Annals of Neurology	8.130							0

C. Glenn Begley and Lee M. Ellis. Raise standards for preclinical cancer research. 29 MARCH 2012. V4 8 3. NATURE

Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.

*Source of citations: Google Scholar, May 2011.

研究结果应该建立在确凿的实验、试验、观察或调查数据的基础上，因此论文中的**数据必须是真实可靠的**，不能有丝毫的虚假。研究人员**应该忠实地记录和保存原始数据，不能捏造和篡改**。虽然在论文中由于篇幅限制、写作格式等原因，而无法全面展示原始数据，但是一旦有其他研究人员对论文中的数据提出疑问，或希望做进一步了解，论文作者应该能够向质疑者、询问者**提供原始数据**。因此，**在论文发表之后，有关的实验记录、原始数据仍然必须继续保留一段时间，一般至少要保存5年，而如果论文结果受到了质疑，就应该无限期地保存原始数据以便接受审核。**

如果研究人员没有做过某个实验、试验、观察或调查，却谎称做过，无中生有地编造数据，这就构成了最严重的学术不端行为之一——捏造数据。如果确实做过某个实验、试验、观察或调查，也获得了一些数据，但是对数据进行了篡改或故意误报，这虽然不像捏造数据那么严重，但是同样是一种不可接受的不端行为。常见的篡改数据行为包括：去掉不利的数据，只保留有利的数据；添加有利的数据；夸大实验重复次数（例如只做过一次实验，却声称是**3次重复实验的结果**）；夸大实验动物或试验患者的数量；对照片记录进行修饰。

人们已习惯用图像软件对图像数据进行处理绘制论文插图，因此又出现了篡改数据的新形式。例如，由于原图的阳性结果不清晰，就用图像软件添加结果。如果没有篡改原始数据，只是通过调节对比度等方式让图像更清晰，这是可以的，但是如果添加或删减像素，则是不可以的。

• 论文的数据去哪里了？



● 国务院发布《科学数据管理办法》

- 2018年1月23日，国务院办公厅通过《科学数据管理办法》，2018年3月17日起正式实施。
- 明确了我国科学数据管理的总体原则、主要职责、数据采集汇交与保存、共享利用、保密与安全等方面内容，着重从五个方面提出了具体管理措施。
- **一是明确各方职责分工，强化法人单位主体责任，明确主管部门职责，体现“谁拥有、谁负责”，“谁开放、谁受益”。**
- **二是按照“分级分类管理，确保安全可控”的原则，主管部门和法人单位依法确定科学数据的密级及开放条件，加强科学数据共享和利用的监管。**
- **三是加强知识产权保护，对科学数据使用者和生产者的行为进行规范，体现对科学数据知识产权的尊重。**
- **四是要求科技计划项目产生的科学数据进行强制性汇交，并通过科学数据中心进行规范管理和长期保存，加强数据积累和开放共享。**
- **五是提出法人单位要在岗位设置、绩效收入、职称评定等方面建立激励机制，加强科学数据管理能力建设。**

国内外数据共享形势与政策

- **美国政府**在其一系列信息公开法律, 如《信息自由法》、《咨询委员会法》、《隐私法》、《阳光下的政府法》、《美国联邦信息资源管理政策》、《电子信息自由法》等框架下制订了各部门的信息公开政策, 如NASA制定《数据与信息政策》、教育部颁布《数据共享实施指南》。
- **英国政府**自1997 年开始制订并分步实施新的《信息自由法案》, 至2005 年1 月该法案全面生效。该法案赋予公众获取公共部门相关信息的权利, 英国各级政府部门均按照该法案颁布实施了本部门的信息公开目录和指南。如卫生部、教育部、能源与气候变化部、边境署、工作与养老金部、区域发展部、农业和农村发展部等。
- **澳大利亚**有关政府信息公开的法律主要是1982年颁布、2010 年进行重要修订的《信息自由法》(Freedom of Information Amendment (Reform) Act ,2010), 赋予每个公民和社团获取政府信息的权利, 联邦政府部门须公布涉及公众的信息、部门工作手册、涉及公众的决策和建议的文件等, 并为公众提供获取这些信息的渠道。如民政部。



DATA.GOV



Australian Government



data.gov.au



DATA.GOV.UK ^{Beta}
Opening up Government

● 科技计划项目管理机构在数据管理方面政策

- **美国**国家科学基金会 (NSF) 发布的项目管理指南(NSF Award and Administration Guide)规定：所有提交到NSF的项目申请书必须包含一份不超过两页的“数据管理计划” (Data Management Plan)。没有数据管理计划的项目申请将不予接受；
- **美国**卫生与人类服务部 (HHS) 下属的国立卫生研究院(NIH)制定了《数据共享计划》，要求从2003年10月1日开始，所有向NIH 申请项目经费在50万以上的科研人员都须包含一个数据共享计划或者数据不共享的说明；
- **美国**航空航天局(NASA)为目前正在实施的每个项目都制定了项目数据管理计划；国农业部 (USDA) 的州际研究、教育和推广局(CSREES)要求所有受资助的研究数据必须没有限制地向公众开放。
- **英国**研究理事会(RCUK) 颁布了相应的数据管理政策。如BBSRC要求研究者提交数据共享计划, 研究者须将出版物存放在英国公共医学中心(UK PubMed Central)。
- **澳大利亚**研究理事会(ARC)和国家健康与医学研究理事会(NHMRC)两个机构有关于科学数据共享的相关规定。鼓励研究者将项目研究中产生的数据和出版物存放在合适的学科库或机构库中, 并在项目最终报告中说明存放的地方。如果研究者不愿意在规定时间内对数据进行存放, 必须在项目最终报告里说明原因。



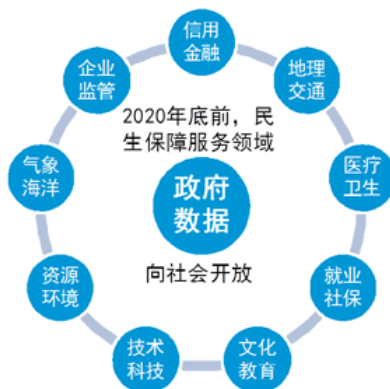
Australian Government
Australian Research Council

机构名称	政策名称	发布时间	相关内容
英国生物技术与生物科学研究理事会 (BBSRC)	BBSRC数据共享政策	2007年4月发布，2016年3月更新	1 需提交数据管理计划或不能共享的原因说明，作为项目资助计划的一部分，强制性的附件形式上交 2 评估数据管理计划，通过同行评议提高可行性 3 数据共享活动费用作为资助项目的一部分 4 数据管理活动贯穿整个项目周期 5 研究人员需要遵守行业道德，注意相关数据的道德性使用 6 保护研究人员的知识产权 7 科学数据被安全保存至少10年
英国医学研究理事会 (MRC)	MRC数据共享政策	2006年1月发布，2011年9月更新	1 数据管理计划作为项目申请的组成部分 2 要求研究者将文章发表在允许作者(或者作者所在研究机构)能保留版权、能够对外开放数据的期刊上 3 资助的个人或机构须遵守数据创建、收集或管理的相应责任 4 监督和评估数据共享活动，汇报资助款项中的数据共享活动的实施情况 5 科学数据被安全保存至少10年
英国癌症研究慈善机构 (CRUK)	CRUK科学数据共享和保护政策、CURK数据共享指南	2009年1月	1 对涉及被试者个人信息的科学数据共享时要保护个人隐私数据，如数据匿名化、删除标识符 2 共享科学数据的时间最迟不晚于其主要陈国在出版物上发表的时间 3 要求科学数据在机构库中至少保存5年

机构名称	政策名称	发布时间
美国国家卫生研究院(NIH)	数据共享政策与实施指南	
美国环境卫生科学研究院 (NIEHS)	环境卫生科学数据共享战略	
美国疾病控制与预防中心 (CDC)	数据发布和共享政策、增加CDC自主的数字科学数据和科学出版物的获取计划	
澳大利亚国家健康与医学研究理事会 (NHMRC)	数据共享声明	2012年发布， 2018年1月更新
加拿大基因组数据共享政策咨询委员会	加拿大基因组数据发布与共享政策	2016年9月
加拿大卫生研究院 (CIHR)、自然科学与工程研究理事会 (NSERC)、社会科学人文研究理事会 (SSHRC)	三机构出版物开放获取政策	2016年12月
爱尔兰健康研究委员会(HRB)	HRB开放获取和无约束访问已出版研究成果政策	2014年5月

● 数据共享在国内也是大势所趋

- 长期以来，科技部积极推动科学数据管理与共享工程，从科技发展的战略高度出发，启动了科学数据共享工程。2001年启动了第一个试点以来，在资源环境、农业、人口与健康、基础与前沿等领域共24个部门开展了科学数据共享工作，已经初具规模。
- 数据共享理念已经在科技界得到了广泛认可，形成共享氛围和服务意识，逐渐改变了我国科学数据封闭独享的局面，带动了跨行业的数据交换。
- 2015年国务院印发《促进大数据发展行动纲要》，明确指出要大力发展科学数据，推动由国家公共财政支持的公益性科研活动获取和产生的科学数据逐步开放共享。



- 一些行业部门、政府部门、高校和自“十一五”以来建设的基础条件平台等都陆续确立了自己的数据管理与共享办法、意见等。

机构类型	发布机构	政策名称
行业部门	中国气象局、中国水科院、中国地震局、中华人民共和国国土资源部、国家海洋局、国家国防科技工业局高分观测专项办公室、江西省国土资源厅、浙江省教育厅、国家海洋局、国家海洋局、国家海洋信息中心、中华人民共和国交通运输部	气象资料共享管理办法、水利科学数据共享管理办法（试行）、地震科学数据共享管理办法、国土资源数据管理暂行办法、中国极地科学考察样品和数据管理办法（试行）、高分辨率对地观测系统重大专项卫星遥感数据管理暂行办法、江西省国土资源数据管理暂行办法、浙江省教育数据暂行管理办法、国家海洋局关于规范海洋生态环境监测数据管理工作的意见、海洋生态环境监测数据共享服务程序(试行)、交通运输部办公厅关于推进交通运输行业数据资源开放共享的实施意见
政府部门	北京经信委、浙江省人民政府、湖北省人民政府办公厅、浙江省人民政府办公厅、山东省人民政府办公厅、青海省政府办公厅、河北省政府办公厅、上海市人民政府、贵州省政府办公厅、江西省人民政府、安徽省人民政府办公厅、湖南省人民政府，等等	北京市政务信息资源共享交换平台管理办法（试行）、浙江政务服务网信息资源共享管理暂行办法、山东省政务信息资源共享管理办法、青海省地理空间数据交换和共享管理办法、河北省政务信息资源共享管理规定、上海市政务数据资源共享管理办法、贵州省政务数据资源管理暂行办法、江西省政务信息资源共享管理实施细则安徽省政务信息资源共享管理暂行办法、湖南省地理空间数据管理办法
基础条件平台	国家人口与健康科学数据共享平台、	

日前，国务院办公厅发布《科学数据管理办法》，不少科学家表示，合理、完善的科学数据管理办法，将有助于我国的科技投入发挥出更巨大的能量。

2018年1月经中央深改组审议通过《科学数据管理办法》，近日由国务院办公厅正式印发。《科学数据管理办法》旨在进一步加强和规范科学数据管理，保障科学数据安全，提高开放共享水平，更好地为国家科技创新、经济社会发展和国家安全提供支撑。这是我国第一次在国家层面出台科学数据管理办法。

《科学数据管理办法》要求科技计划项目产生的科学数据进行强制性汇交，并通过科学数据中心进行规范管理和长期保存。

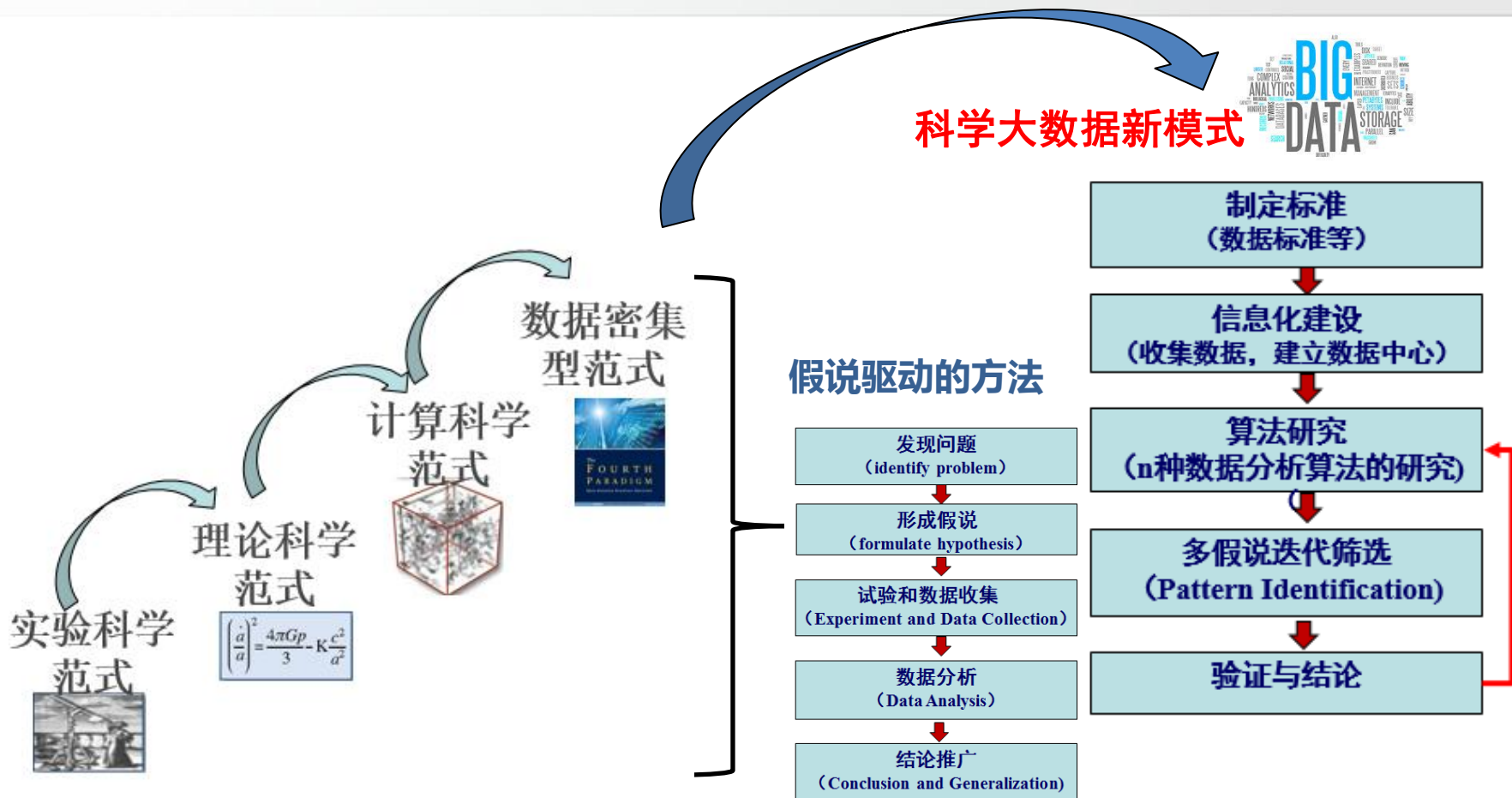
● 《科学数据引用》国家标准正式发布

- 国家标准化管理委员会通过官方网站发布《中华人民共和国国家标准公告（2017年第32号）》（<http://www.sac.gov.cn/gzfw/ggcx/gjbzgg/201732/>），中国科学院计算机网络信息中心主持研制的《信息技术 科学数据引用》（GB/T 35294-2017）正式发布，自2018年7月1日起正式实施。
- 《科学数据引用》规定了科学数据引用元素描述方法、引用元素详细说明、引用格式等方面的内容，适用于科学数据传播机构、数据使用者等。其中，科学数据传播机构可根据本标准设计数据引用系统，并声明数据引用规则；数据使用者可根据本标准著录科学数据引用信息。
- 《科学数据引用》国家标准的正式发布，标志着**科学数据可以像学术论文一样被学术同行标准化引用**，将在一定程度上促进数据拥有者开放共享其数据。



科研范式的转变与期刊届的数据政策 (Data Policy)

● 科研模式的转变



- 科学研究需要一种开放的模式，包括开放**数据**

- 科学研究需要一种开放的模式，包括**开放数据**、**科学出版物的开放存取**以及对其内容的有效交流——英国皇家学会《科学是开放事业》（Science as an Open Enterprise）其中，提出**数据质量**应该能够达到如下标准
 - 可获取的
 - 可理解的
 - 可评估的
 - 可利用的

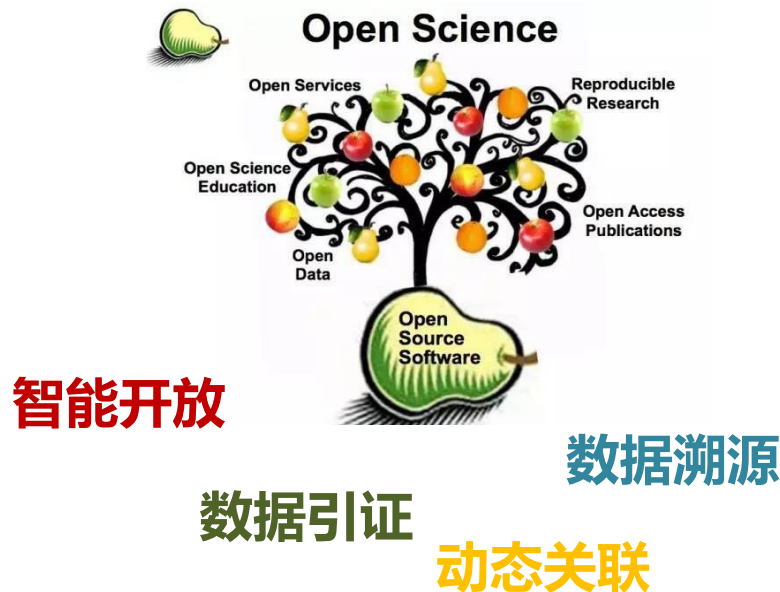


FAIR Principles

FAIR - a set of guiding principles to make data **F**indable, **A**ccessible, **I**nteroperable, and **R**e-usable

FAIR原则

- 可发现性 (**F**indability)
- 可获得性 (**A**ccessibility)
- 互操作性 (**I**nteroperability)
- 可重用性 (**R**eusability)

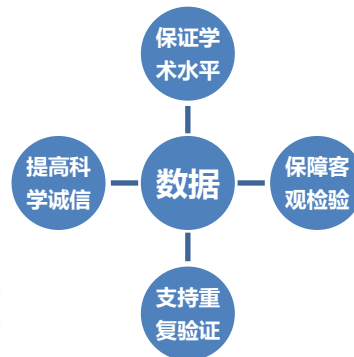


-
- ```
graph TD; A((Identify problem / observe phenomenon)) --> B((Formulate hypothesis / create experiment)); B --> C((Run experiment / create data)); C --> D((Analyse data)); D --> E((Disseminate results)); E --> F((Quality evaluation / peer review)); F --> A; D --> B;
```

# ● 公布支撑数据成为学术期刊的基本要求

论文投稿时提交和发表时公布支撑数据,已成为越来越多学术期刊的基本要求,例如Science、Nature、Cell、PLOS Biology等。

- 2014.03 **PLOS**发布了[PLOS' New Data Policy](#) (Data 可以开放共享并容易获取)
- 2016.06.03 **Science**杂志的社论,提到实行、推广**TOP** (**Transparency and Openness Promotion**)。TOP最早由美国开放科学中心 (COS) 开展心理学研究可重复性项目时提出,包括对科学期刊和科学家有关研究透明性、数据共享等多方面的建议。
- 2017.03.15 **Nature**旗下期刊将推进采用更严格的期刊政策,以增加研究成果报道的透明性。简单说,就是以后这些期刊将更严格的要求作者公开原始数据、研究材料和方法等。



- 数据提交强制级别：强制提交、鼓励提交、无明确要求
- 期刊示例：
  - （来源：2017 Journal Citation Reports 中的TOP 30期刊，TOP 30中有20种可投生物医药类论文）
- ◆ NEW ENGLAND JOURNAL OF MEDICINE ( Rank 2 )
  - 没有数据论文相关栏目，但是在网站精选栏目中有数据共享专题。在2017年4月3-4日，通过the Aligning Incentives for Sharing Clinical Trial Data web event召集了来自世界各地的学者、分析人士和患者社区，举办了一场关于临床试验数据共享的公开讨论。
- ◆ NATURE REVIEWS DRUG DISCOVERY ( Rank 3 )
  - nature旗下的期刊，遵循Availability of data, material and methods政策。要求作者的数据可获取：评审阶段数据对编辑和同评专家可见；出版后应读者合理要求开放数据。规定某些类型的数据必须强制存储，并指定了对应的存储库。
  - 执行该数据共享政策TOP30期刊还有NATURE REVIEWS MOLECULAR CELL BIOLOGY ( 6 )、NATURE BIOTECHNOLOGY ( 8 )、NATURE REVIEWS GENETICS ( 9 )、NATURE ( 10 )、NATURE REVIEWS IMMUNOLOGY ( 12 )、NATURE MEDICINE ( 23 )、NATURE REVIEWS NEUROSCIENCE ( 27 )

### ◆ LANCET ( Rank 5 )

- 研究数据是研究论文的有力支撑。建议作者将数据提交到存储库，如Mendeley Data，以便于数据引用、与文章关联，并提供了数据库关联工具Database Linking Tool，列出了该工具支持的存储库。

### ◆ Science ( Rank 16 )

- 规定cover letter中要指明所有支撑数据的可用性，将被如何存储，数据的使用限制。提交论文的时候也要将数据作为补充材料上传，以便同评专家审阅。

### ◆ Cell ( Rank 22 )

- 其特殊栏目资源 ( Resource )，用于接收介绍重要信息数据库，并基于对该数据库的分析，得出新见解的稿件。

### ◆ Cancer Cell ( Rank 30 )

- 要求作者对数据负责，保留原始数据。鼓励作者将原始数据上传至Mendeley Data、Dryad或者其他图片和数据存储库。规定了强制存储的数据（蛋白序列、DNA和RNA序列、深度测序数据、人类基因组序列、生物大分子结构等）和非强制存储的数据（蛋白质组学数据、蛋白质交互作用数据、化学复合筛选和分析数据、流式细胞仪数据）及其推荐存储库。

## ● 我国部分学术期刊数据政策

• 《**数据分析与知识发现**》（原《现代图书情报技术》）提出《支撑数据提交要求》，从2016年第2期起，要求所有发表论文提交支撑论文结论的数据，并通过适当方式供研究共同体或社会公众共享。

• 《关于投稿人自愿提供稿件支持原始数据的通知》——《**中华外科杂志**》与**国家人口与健康科学数据共享平台肿瘤专题服务网络共享平台**（简称肿瘤专题平台）投稿稿件一经录用，支撑数据将在文章发表的同时，数据将纳入国家人口与健康科学数据共享平台管理和发布。发布内容为去隐私化的全部或部分数据，由肿瘤专题平台根据平台的共享规则确定。

• 《**中华健康管理学杂志**》从2016年起，逐步要求原始研究类稿件提供相应的原始资料，例如：原始数据、原始结果、量表、干预方法、问卷等。可提供的原始研究资料包括但不限于以上内容。

表1 提交数据类型

| 数据类型                             | 具体描述      |                                                        |
|----------------------------------|-----------|--------------------------------------------------------|
| 内在数据（必选）<br>(Integral Content)   | 研究方法详细描述  | 研究方法、实验环境配置、应用环境配置、实验操作过程等。研究方法的要点在正文中提供，详细信息作为数据支持。   |
|                                  | 直接研究工具    | 调查问卷、访谈提纲、测试题目、计算算法、样本数据等。                             |
|                                  | 直接研究结果数据  | 直接产生正文图表和正文所描述结论的样本数据表、测试数据记录、调查或观察数据记录、计算或模拟数据记录等。    |
| 附加数据（可选）<br>(Additional Content) | 研究方法背景信息  | 研究方法、技术、工具（设备）、实验环境等的进一步细节描述。                          |
|                                  | 扩展的研究结果数据 | 为得到研究结论所使用的原始实验数据，它们可能没有直接用于产生正文图表，但有助于理解和支持正文研究结论的数据。 |
|                                  | 扩展的参考文献表  | 对于帮助读者理解研究方法、研究过程和研究结论有帮助的参考文献。                        |

# 《中国内科杂志》在数据出版及增强出版的实践

- 升级了新的阅读体验，在全文页支持视频播放
- 对于手术类，操作类的文章，这是个非常有效的表现形式
- 结构化数据的标准



# 什么是科学数据？ 医学大数据

# • 什么是科学数据

## • 体量大小

- 科学大数据
  - 大科学装置、大型仪器设备和野外台站网等产生的海量数据
  - 行业部门按照统一规范标准长期采集的数据
- 长尾小数据
  - 80%的科研活动都集中于生产长尾小数据
  - 经费低、规模小、参与人员众多

## • 表现形式

- 数值，文本描述，图像、视频、音频等多媒体文件，源代码，模型，研究结果，数据产品、信息、算法等

## • 生产方式

- 观测数据、实验数据、遥感数据、调查数据、统计数据、模拟数据、衍生数据等

- 数据来源：高通量技术的发展、现代医疗设备产生的海量、医院信息化、病因学研究等
- 贯穿从基础研究、药物开发、临床诊疗到健康管理的所有环节
  - 基础研究：高通量基础组、转录组测序产生的数据
  - 药物开发：寻找靶标、建立模型、寻找化合物等过程中产生的数据
  - 临床医学：现代数字化医疗系统产生的海量数据，如X光检查、核磁共振、彩色超声、内窥镜、血管造影等医学影像数据和电子病历数据。

- 大数据曾被定义为“海量（Volume）、高速（Velocity）和多样（Variety）的信息资产，需要应用符合成本效益比的创新处理模式，以获得更强的洞察力、决策力和过程自动化。”
- 如今，数字化医疗已经成为行业常态。通过医疗记录、行业监管和医学研究，大量的医疗数据正在产生。在全球范围内，数据量预计将以每两年翻倍的速度增长；到2020年，数据量将比2011年增长50倍<sup>4</sup>。

### 医疗大数据的来源

- 行业监管性数据和医疗保险数据；
- 常规人口统计和重大疾病监测数据
- 真实世界数据，包括电子医疗档案、医学影像和健康体检数据等
- 研究数据、来自临床试验或队列研究等（包括生物学标志物和多组学信息）；
- 登记系统（如器械、操作、疾病）；
- 移动医疗设备；
- 患者自我报告的数据

# 大数据应用于健康医疗领域已成为国家发展的重点

医疗数据的多样性来自数据来源的多元性；这些数据来源反映了健康和疾病的全过程，包括结构化数据和非结构化数据。

此外，医疗数据正在与社交媒体、职业信息、地理位置、经济数据和环境数据等结合。这种对于不同来源数据的整合分析是发挥大数据力量的关键。另外，大数据生成和处理的速度还应该能够满足预防和治理疾病的时效需求。

最近，真实性（veracity）作为大数据的目标之一，被纳入到大数据的定义之中。尽管怀疑论认为大数据难以逐一验证，并且永远不可能“绝对准确”；但毋庸置疑，为了最大限度地利用大数据，数据质量的持续改进是不可或缺的。

与大数据相关的另一个重要概念是，汇集数据并非我们的最终目的。相反，需要分析、解读数据，并基于以上进行决策支持。因此，为了获取大数据的全部价值（value），需要不断地发展新兴技术和分析方法（如机器学习）；并且评估新生成信息的临床效果、并将其转化为临床实践。

- 在中国，将大数据应用于健康医疗领域已成为国家发展的重点。
- 政府和学术界已采取多种措施提升大数据应用。
- 应用大数据与创新数据技术可能有助于提升医学研究和对疾病、健康的理解水平。

促进大数据在医疗领域的应用已经成为我国的重要战略。2016年7月，国务院办公厅印发了《**促进和规范健康医疗大数据应用发展的指导意见**》。意见指出，健康医疗大数据是国家重要的基础性战略资源，其应用发展将提升中国的健康医疗水平。意见还为医疗大数据发展规划了目标、任务和组织框架。

## 目前在中国应用健康医疗大数据的项目案例

### 政府牵头

- 在福建和江苏两省四城市开展的区域医疗平台试点项目 ([http://en.nhfpe.gov.cn/2016-10/24/c\\_70420.htm](http://en.nhfpe.gov.cn/2016-10/24/c_70420.htm))。
- 将已有的国家卫生监管数据平台、医疗保险数据、死亡登记系统等供学术使用。
- 通过支持队列研究和综合数据平台推广精准医学 ([http://www.most.gov.vn/tztg/201603/t20160308\\_124542.htm](http://www.most.gov.vn/tztg/201603/t20160308_124542.htm))。

### 学界发起

- 中国队列共享平台 ([chinacohort.bjmu.edu.cn](http://chinacohort.bjmu.edu.cn))。
- 中国肾脏疾病数据网格 ([kidney.net.cn](http://kidney.net.cn))。
- 其他疾病，包括心血管疾病（如中国心血管手术注册系统）、卒中（如中国国家卒中登记系统）和癌症（如中国国家癌症登记系统）。

- 在过去的十年中，医学领域中产生、采集和储存的数据量呈现几何级增长；分析和解读这些数据的能力也快速增长。我国人口庞大、具备全国统一的医疗卫生保障体系，这些都为发展健康医疗大数据提供了优势。
- 促进健康大数据在医疗领域的运用是我国重要的国家战略，目前有很多国家支撑的正在进行的重要项目。
- 除了政府引导的研究项目以外，中国的医学研究团队和机构也已经启动了数据共享学术项目。
- 为了更好地应用健康大数据，需要解决不同医疗机构间的有效数据交互、构建更符合我国临床实践的医学术语体系、系统性改善对于人群的随访、改善数据质量、并尽快就保护个人隐私和数据安全出台相关政策法规。

日前，国务院办公厅发布《科学数据管理办法》，不少科学家表示，合理、完善的科学数据管理办法，将有助于我国的科技投入发挥出更巨大的能量。

当中国科学家的论文频频登上《自然》《科学》《细胞》等国际一流学术期刊，当中国的SCI论文发表数量上升到全球第二位，你可曾想到，这同时意味着大量一手、原始的科学数据也流到了国外？

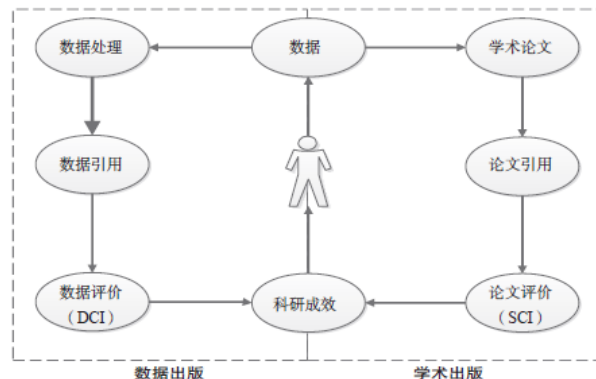
- ◆ 目前，在生命科学领域，世界上有三大数据中心，分别位于美国、欧洲和日本。
- ◆ 过去30年，中国学者为这三大中心贡献了**30%**以上的基因组数据，其中大部分与中国人群有关。
- ◆ “根据国际学术期刊论文发表惯例，一篇论文所有的原始数据都必须上传到权威的专业数据库。”中国科学院上海生命科学研究院生物医学大数据中心副主任张国庆表示，因为同行重复实验、杂志编辑审核论文，都需要调用原始实验数据。

- 医学资源可以创造更多的科研价值、经济价值和社会价值；
- 减少不必要的重复研究，节约资源，促进重用；
- 提供强大的数据支撑，使得医学决策更加精准；
- .....

# 数据出版 · 数据期刊 · 数据论文

**数据出版**是近几年由出版界和数据共享界共同提出和积极实践的新概念，通过“数据提交、同行审议、数据发布和数据永久存储、数据引用和数据影响评价”等基本环节搭建数据出版体系框架，解决数据知识产权、数字资源标识、数据（知识）发现等核心问题，最终构建包括文献、数据和科学家的知识管理体系。

科学数据的出版是一种全新的数据共享模式，通过科技期刊方式发布已经获取的、经过专业质控的、规范化的科学数据，是一种便捷地发现、获取数据的途径，也是引导科研人员分析、研究和再利用数据的索引；是在数据生产者的知识产权和所有权得到充分保障的前提下，对科学数据共享的一种探索性尝试，力求科学数据资源的最大化使用。



# 为什么要出版数据



## Resolving the publishing bottleneck for biodiversity

*Science is a combination of gathering facts and making theories; neither can progress on its own. In the history of science, the laborious accumulation of facts is the dominant mode, not a novelty.*

Peter Norvig



### Making "small" data big

- No lower/upper limit of manuscript size
- Publish all kinds of biodiversity related data
- Reduced page charges affordable by all



### More than just data journal!

- Integrated text and data publishing
- Completely online revisions and editing
- Community ownership of data



### Community peer review

- 7 weeks from submission to decision
- 3 days from acceptance to publication
- Public peer review on author's choice

ISSN 1314-2828 (online) Indexed by Scopus, DOAJ and Google Scholar Archived in PubMedCentral and CLOCKSS

### Why publish my data?

- Citable publication
- Establish scientific priority
- Increase collaboration
- Link data to a bigger network
- Re-use and multiply effect
- Respond to funding requirements



- 让小数据汇成大数据
- 方便引用
- 保护数据知识产权
- 增加合作机会
- 数据重用，增加数据多效性
- 回馈项目资金



# • 为什么要出版数据

## • 从数据工作者的角度

- 保护知识产权，数据产权清晰
- 便于他人验证数据的真实性
- 提高知名度，获得学术声誉、认可
- 可对数据引用情况进行跟踪统计与分析，从而客观评价其对科学发现的价值和影响（Impact）
- 有利于对从事数据加工分析的人员的科研成果评价
- 全球统一标识，可在学术出版物中正式引用
- 持久访问和数据存放安全
- 细化科研分工（数据生产、学术分析）
- 提高跨机构、跨学科合作机率



## 独立的数据出版

数据中心对本机构保存的数据的独立出版

- 各国政府数据中心  
(如data.gov)
- 专业数据中心和机构  
数据中心等



## 论文的辅助资料

数据资料作为支撑论文学术思想的佐证，是论文本身需要展示的不可或缺的部分。目前主要有两种形式：

- 数据以附件形式存在
- 数据存储在专业数据库或出版机构网站上，  
如：Dryad、Figshare、  
ScienceDB



## 数据期刊与论文

描述科学数据本身  
论文+数据

- 正式的学术论文
- 接受严格的同行评议  
(peer-review)
- 如*Earth System Science Data*、*Scientific Data*、  
中国科学数据、全球变化数据学报

## • 关于数据期刊的一些数据

15

### 数据期刊出版商

数据期刊的出版者主要是**15**个出版社，发行110多种数据期刊。其中：

BMC出版85种，  
ChemistryCentral出版3种，  
Pensoft出版7种，  
SpringerOpen出版8种，  
Ubiquity Press出版3种；

110<sup>+</sup>

### 数据期刊

数据期刊主要在**医学、生命科学、物理科学**领域；大部分的数据期刊为混合型期刊，即出版各种类型论文包括数据论文。

7

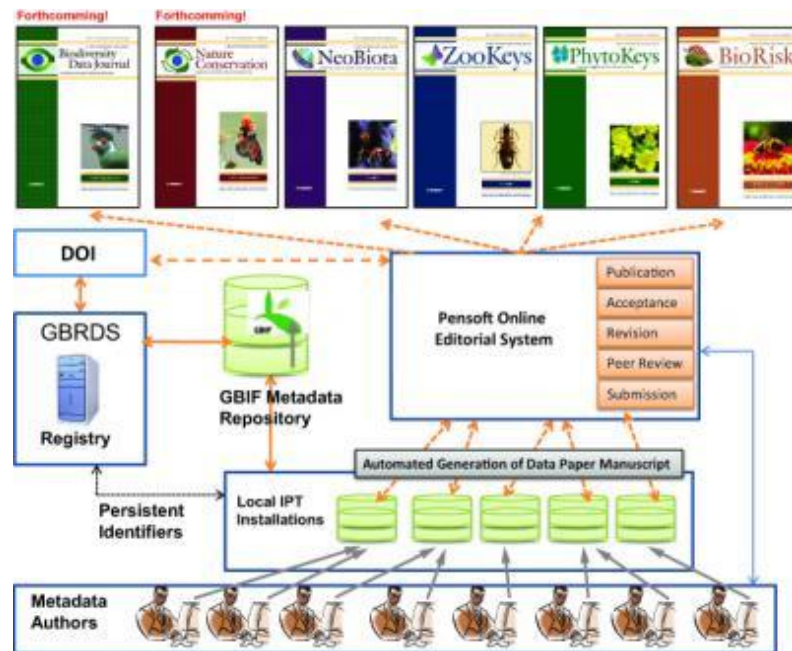
### 单一型数据期刊

15个出版社中只有5个专门出版数据论文的期刊，包括了**7**种期刊，其中**只有5种是开放获取期刊**：

*Dataset Papers in Science,*  
*Earth System Science Data,*  
*Genomics Data,*  
*Geoscience Data Journal,*  
*Scientific Data;*

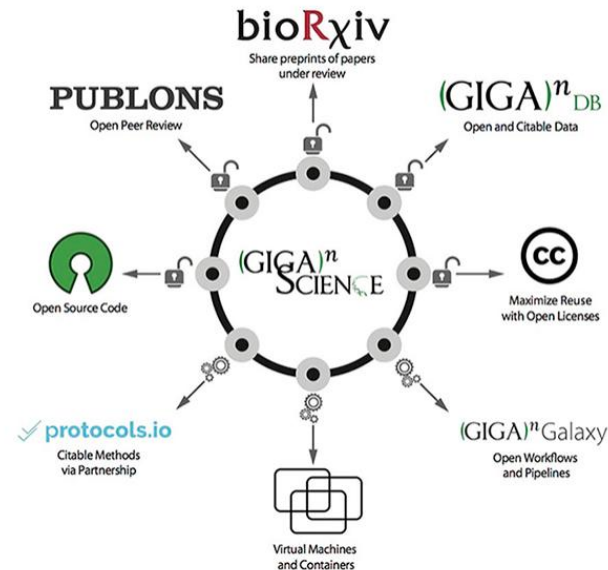
- 国际
  - *Scientific Data* ( 2014 , 自然科学/综合 , Nature )
  - *Earth System Science Data* ( 2009 , 地球科学 , Copernicus Publications )
  - *F1000 Research* ( 2012 , 生命科学/医学 , F1000 Research )
  - *Biodiversity Data Journal* ( 2013 , 生物多样性 , Pensoft )
  - *Ecology* ( Data paper , 1999 , 生态学 , Ecological Society of America→Wiley )
  - *Geoscience Data Journal* ( 2014 , 地球科学 , Wiley )
  - *Genomics Data* ( 2013 , 基因组学 , Elsevier )
  - *Gigascience* ( 2012 , 生物学 , BioMed Central→Oxford academic )
  - Springer Plus ( 2014 , 综合 , Springer )
  - **BMC research notes** ( 2008 , 生命科学 , BMC )
  - Dataset Papers in Science ( 2013 , 综合 , Hindawi Publishing )
  - **Open Health Data** ( 2013 , 生命科学 , Ubiquity Press )
  - Journal of Open Archaeology Data ( 2013 , 考古学 , Ubiquity Press )
  - Journal of Open Psychology Data ( 2013 , 心理学 , Ubiquity Press )
- 国内
  - **中国科学数据** ( 2016 , 自然科学/综合 , 中科院网络中心 )
  - 全球变化数据学报 ( 2017 , 地球科学 , 中科院地理所 )

全球生物多样性信息网络(Global Biodiversity Information Facility, GBIF)和 Pensoft在2011年联合提出出版数据论文，主要目的是采用学术论文的形式将数据资源在线发表出来，从而为数据共享、数据质量控制和改善科研评价提供一个有效的方法。2013年9月，Pensoft创办了数据期刊 *Biodiversity Data Journal* (BDJ),用于发表生物多样性数据论文，论文的数据集可以无缝分享到GBIF。 <http://bdj.pensoft.net/>



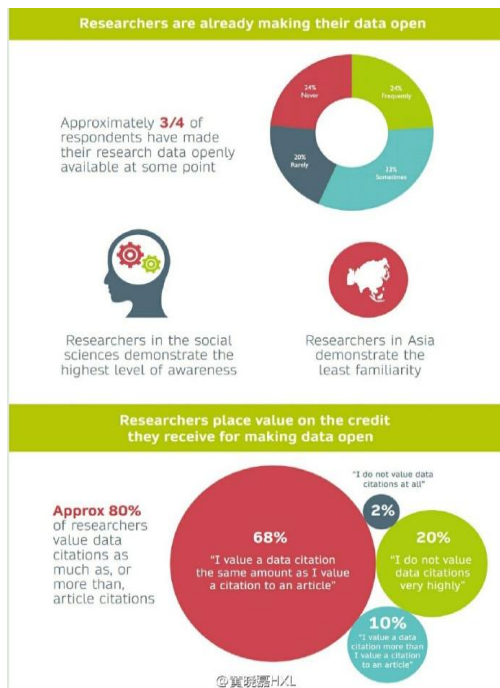
2012年7月，华大基因和生物医学中心 (BioMed Central)共同创办了*GigaScience*数据期刊，专注于生命科学和生物医学科学领域的“大数据”研究。*GigaScience*积极致力于数据开放的工作，该期刊采用标准全文文献、数据库信息以及信息分析工具相结合的崭新模式，为科研工作者提供免费公开的有效数据以及生物学发现等资源，从而促进信息的组织、理解和可重复性。*GigaScience*突破了传统出版业的束缚，为数据全面公开与共享提供了一个优秀平台，实现了科研人员对数据资源最大可能的合理利用，并提高了数据密集型科研项目的再现性。

<https://academic.oup.com/gigascience/pages/About>



# 科学数据的引用

据调查，大多研究者认为数据引用的价值至少等同于论文引用；  
(自figshare)



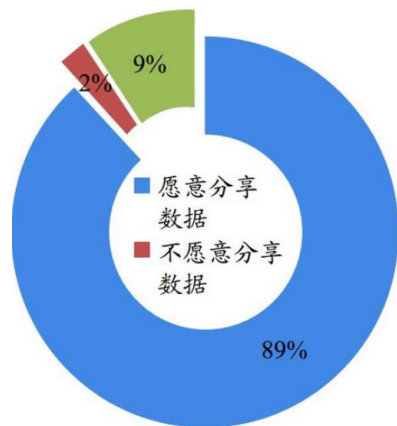
- 再次研究 ( Reproduce research )
- 重复科学发现 ( Replicate findings )
- 数据重用 ( Reuse data )
- 数据贡献的认可 ( credit )

If publications are the stars and planets of the scientific universe, data are the 'dark matter' - influential but largely unobserved in our mapping process\*

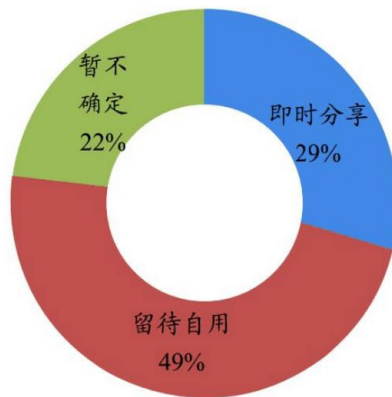
ESSD

| Journal metrics     |                |
|---------------------|----------------|
| Clarivate Analytics | IF 6.696       |
| Scopus              | CiteScore 7.28 |
| CWTS                | SNIP 2.755     |
| SJR                 | SJR 5.363      |
| CWTS                | IPP 6.509      |
| g1                  | h5-index 23    |
| Definitions         |                |

# 数据论文作者意向调查



科学数据共享意愿



是否立刻分享 (所获数据或实验结果)

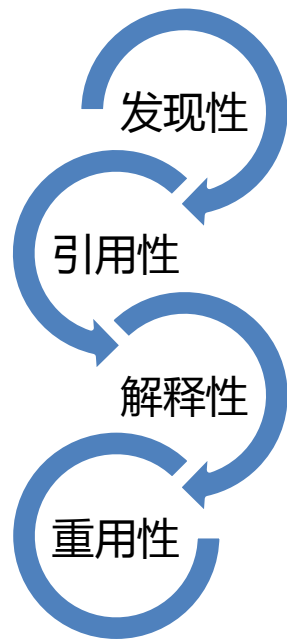
| 问题                                | 调查结果 ( % ) |      |      |
|-----------------------------------|------------|------|------|
|                                   | 是          | 否    | 不确定  |
| 学术声誉是推动数据出版的动力 ( Q3 )             | 46 %       | 39 % | 15 % |
| 对自己已发表数据的引用量表示满意 ( Q4 )           | 31 %       | 44 % | 25 % |
| 未来是否会再次发表数据 ( Q7 )                | 89 %       | 3 %  | 8 %  |
| 数据出版物和科研论文出版一样有学术意义 ( Q5 )        | 50 %       | 37 % | 13 % |
| 数据期刊出版者能否在科学共同体中获得较高声誉 ( Q8 )     | 45 %       | 19 % | 36 % |
| 数据发表于专门的数据出版物比普通科学出版物更有价值 ( Q14 ) | 51 %       | 26 % | 23 % |

# • 什么是数据论文？

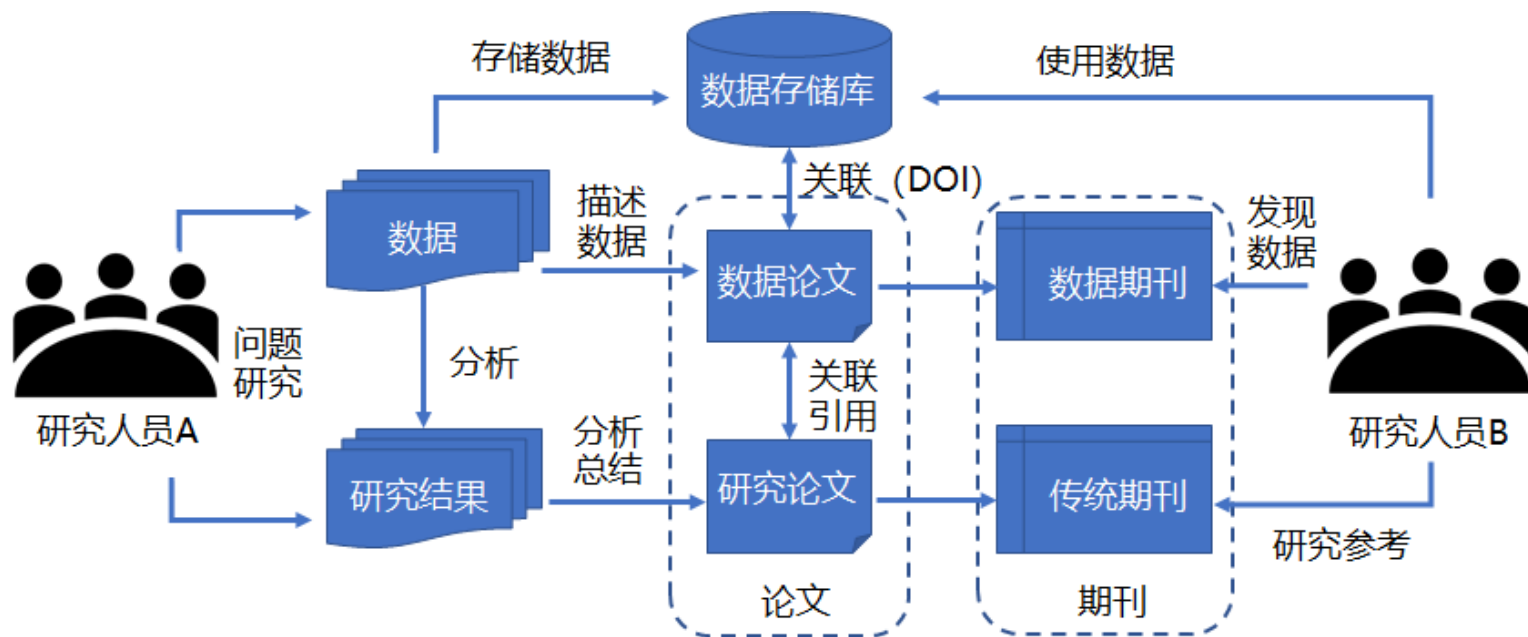
## Data paper, Data descriptor, Dataset/Database paper, Data note, Data in Brief, Data original article

- ✓ “scholarly publication of a searchable metadata document describing a particular on-line accessible dataset, or a group of datasets, published in accordance to the standard academic practices”. Their final aim being to provide “information on the what, where, why, how and who of the data”. The intent of a data paper is to offer descriptive information on the related dataset(s) focusing on data collection, distinguishing features, access and potential reuse rather than on data processing and analysis.
- ✓ Scientific Data primarily publishes Data Descriptors, a new type of scientific publication designed to promote an in-depth understanding of research datasets. Data Descriptors combine traditional scientific article content with structured information curated in-house, and are devised to maximize data reuse and enable searching, linking and data-mining. Each is peer-reviewed under the supervision of our Editorial Board.
- ✓ A data paper is a searchable metadata document, describing a particular dataset or a group of datasets, published in the form of a peer-reviewed article in a scholarly journal.
- ✓ A data paper is a short, peer-reviewed publication that is designed to raise awareness of a dataset and its reuse potential.





- **数据论文 ( Data Paper )** 是一种开放访问并经过同行评议的出版物，对具有科学价值的数据集进行描述，结合传统期刊论文内容和结构化描述模式，使数据更具**发现性**、**引用性**、**解释性**和**重用性**。
- 数据论文应当提供论文描述的数据集细节，包括：
  - 数据集收集、加工处理方法
  - 数据质量评估和验证的方法
  - 便于理解和使用数据的相关信息等但，
  - **不包含新的科学假设。**







## 相似

- 形式
  - 传统出版规范化流程
- 内容
  - 不重复原则
  - 额头出汗原则
  - ...
- 目标
  - 科学交流传播
  - 研究积累
  - ...

## 不同

- 形式
  - 在线、OA、两阶段评审
- 内容
  - 不强调观点创新
  - 不做推演分析
  - e.g. 数据分析——学术论文；数据描述——数据论文
- 目标
  - 科学交流传播（思想——数据）
  - 评价机制

## ● 通用存储库

| 存储库                    |                                                                                   | 费用                                       | 大小限制                  |
|------------------------|-----------------------------------------------------------------------------------|------------------------------------------|-----------------------|
| Dryad                  |  | 前20GB120美元，之后每增加10GB加50美元                | 未知                    |
| figshare               |  | 定制报价，如对Scientific Data来源的稿件免费100 GB，额外收费 | 5 GB                  |
| Science Data Bank      |  | 免费                                       | 无                     |
| Harvard Dataverse      |  | 超过1 TB后可能收费                              | 每个文件2.5 GB，每个数据集10 GB |
| Open Science Framework |  | 免费                                       | 每个文件5 GB，支持上传多个文件     |
| Zenodo                 |  | 根据个人意愿捐赠                                 | 每个数据集50 GB            |

# ● 生物科学类存储库

SCIENTIFIC DATA

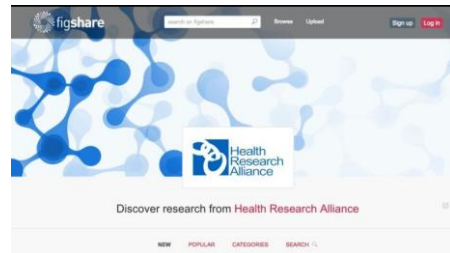
| 细分类别      | 推荐存储库                                                                                                                                                                                                                                                                                                                                                                                    |
|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 核酸序列      | DNA DataBank of Japan (DDBJ) ;European Nucleotide Archive (ENA) ;GenBank;dbSNP;European Variation Archive (EVA);dbVar;Database of Genomic Variants Archive (DGVA);EBI Metagenomics;NCBI Trace Archive;NCBI Sequence Read Archive (SRA);NCBI Assembly                                                                                                                                     |
| 蛋白序列      | UniProtKB                                                                                                                                                                                                                                                                                                                                                                                |
| 分子和超分子结构  | Protein Circular Dichroism Data Bank (PCDDb);Crystallography Open Database (COD);Coherent X-ray Imaging Data Bank (CXIDB);Biological Magnetic Resonance Data Bank (BMRB);Electron Microscopy Data Bank (EMDB);Worldwide Protein Data Bank (wwPDB);Structural Biology Data Grid                                                                                                           |
| 神经科学      | NeuroMorpho.org;Functional Connectomes Project International Neuroimaging Data-Sharing Initiative (FCP/INDI);OpenNeuro (formerly OpenfMRI);G-Node                                                                                                                                                                                                                                        |
| 组学        | <b>Functional genomics:</b> ArrayExpress;Gene Expression Omnibus (GEO);GenomeRNAi;dbGAP;The European Genome-phenome Archive (EGA);Database of Interacting Proteins (DIP);IntAct;Japanese Genotype-phenotype Archive (JGA);Biological General Repository for Interaction Datasets*;NCBI PubChem BioAssay.<br><b>Metabolomics:</b> MetaboLights.<br><b>Proteomics:</b> PeptideAtlas;PRIDE. |
| 分类学和物种多样性 | Integrated Taxonomic Information System (ITIS);KNB: The Knowledge Network for Biocomplexity;NCBI Taxonomy*;Global Biodiversity Information Facility (GBIF);Morphobank.org                                                                                                                                                                                                                |
| 数学与建模资源   | BioModels Database;Kinetic Models of Biological Systems (KiMoSys);The Network Data Exchange (NDEx)                                                                                                                                                                                                                                                                                       |
| 血细胞计数与免疫学 | FlowRepository;ImmPort                                                                                                                                                                                                                                                                                                                                                                   |
| 图像        | Image Data Resource;The Cancer Imaging Archive;SICAS Medical Image Repository;Coherent X-ray Imaging Data Bank (CXIDB)                                                                                                                                                                                                                                                                   |
| 有机体资源     | Eukaryotic Pathogen Database Resources (EuPathDB);FlyBase;Influenza Research Database;Mouse Genome Informatics (MGI);Rat Genome Database (RGD);VectorBase;Xenbase;Zebrafish Model Organism Database (ZFIN)                                                                                                                                                                               |

## ● 医学类存储库

| 推荐存储库                                                                  | 是否限制访问 |
|------------------------------------------------------------------------|--------|
| National Addiction & HIV Data Archive Program (NAHDAP)                 | 是      |
| National Database for Autism Research (NDAR)                           | 是      |
| The Cancer Imaging Archive                                             | 是      |
| ClinicalTrials.gov                                                     |        |
| SICAS Medical Image Repository (formally Virtual Skeleton Database)    |        |
| PhysioNet                                                              |        |
| National Database for Clinical Trials related to Mental Illness (NDCT) | 是      |
| Research Domain Criteria Database (RDoCdb)                             | 是      |
| UK Data Service                                                        | 是      |

HRA (Health Research Alliance) Open——卫生研究联盟开放平台，是由美国卫生研究联盟和美国国家医学图书馆合作资助、Figshare开发建设的学术研究云仓储开放平台。资助信息由HRA Reporter提供，HRA Open基于Web为受资助人员提供世界范围内25多个的研究资助者和1.3亿美元的基金信息，以及上传、分享资助项目相关资料如图像、数据集、报告、出版物等服务，是连接奖项、奖金、出版物、临床试验和专利数据的数字科学平台。

另外，通过与美国国家医学图书馆合作，HRA Open允许受资助成员将其研究成果链接到PMC中，促进成果公共访问。



## 资源地址:

<https://www.prnewswire.com/news-releases/health-research-alliance-hra-launches-new-platform-for-open-science-in-partnership-with-the-national-library-of-medicine-powered-by-figshare-and-dimensions-676518403.html> ;

<https://hra.figshare.com/>

# 不同期刊对数据论文的评议标准

| <i>Data of Brief</i>                                                                                                                                                                                                                                                                                                                                                                                                               | <i>GigaScience</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | <i>Scientific Data</i>                                                                                                                                                                                                                                                                | <i>Pensoft journals</i>                                                                                                                                                                                                                                                                                                                                                                                                            |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ol style="list-style-type: none"><li>1. Do the description and data make sense?</li><li>2. Do the authors adequately explain its utility to the community?</li><li>3. Are the protocol/ references for generating data adequate?</li><li>4. Data format (is it standard? potentially re-usable?)</li><li>5. Does the article follow the Data in Brief <a href="#">template</a>?</li><li>6. Is the data well documented?</li></ol> | <p>Reviewing of the manuscript.</p> <ol style="list-style-type: none"><li>1. Is the rationale for collecting and analyzing the data well defined?</li><li>2. Is it clear how the data were collected and analyzed?</li><li>3. Is it clear how the data are stored and shared?</li><li>4. Are accession numbers or links provided for the data?</li><li>5. Is the data and software available in the public domain under a Creative Commons license?</li><li>6. Are the data sound and well controlled?</li><li>7. Is the interpretation (Analysis)</li></ol> | <p><b>I. Experimental Rigor and Technical Data Quality</b>1. Method</p> <p>...e, and/or</p> <p>...the</p> <p>...combination</p> <p>...metadata,</p> <p>...ant</p> <p>...on or</p> <p>...reporting standards?</p> <p><b>III. Integrity of the Data Files and Repository Record</b></p> | <p><b>I. Quality of the Manuscript</b></p> <p><b>II. Quality of the Data</b></p> <p>completely and consistently Repeatable</p> <p>Are the data plausible, given the protocols?</p> <p>Authors are encouraged to report any tests undertaken to address this point.</p> <p>Is the repository to which the data are submitted appropriate for the nature of the data?</p> <p><b>III. Consistency between Manuscript and Data</b></p> |

**Peer Review 标准：**  
对论文的质量评审  
论文与数据集的一致性  
数据质量  
数据重用性  
数据使用价值

# 数据出版费 ( Data Publishing Charge , DPC )

- *Scientific Data* : ¥11200+
- *Geoscience Data Journal* : \$1500
- 中国科学数据: 优惠期暂免

## SD's article-processing charge (APC)

### APC Prices

These rates apply to Data Descriptors, Analyses, Articles and Comments.

#### 2017 rates

|                  |           |
|------------------|-----------|
| The Americas     | \$1,350   |
| Europe           | €1,050    |
| Japan            | ¥145,500  |
| UK/Rest of World | £890      |
| China            | RMB 8,300 |

#### 2018 rates

These rates will apply to all publications accepted after March 1st, 2018.

|                  |            |
|------------------|------------|
| The Americas     | \$1,675    |
| Europe           | €1,305     |
| Japan            | ¥180,000   |
| UK/Rest of World | £1,110     |
| China            | RMB 11,200 |



## Additional Data Hosting Charges

As part of our article processing charge, we cover 100GB of storage at [figshare](#). The following rates apply for large datasets uploaded to [figshare](#) through our integrated system:

|                | The Americas | Europe | Japan    | UK/Rest of World | China      |
|----------------|--------------|--------|----------|------------------|------------|
| >100 to 250 GB | \$1,000      | €775   | ¥110,000 | £660             | RMB 6,000  |
| >250 to 500 GB | \$2,000      | €1,550 | ¥220,000 | £1,320           | RMB 12,000 |
| >500 to 750 GB | \$3,000      | €2,325 | ¥330,000 | £1,980           | RMB 18,000 |
| >750 to 1 TB   | \$4,000      | €3,105 | ¥440,000 | £2,640           | RMB 24,000 |

# 如何撰写和发表数据论文 ——以《中国科学数据》为例



# ● 《中国科学数据 (中英文网络版)》 *China Scientific Data*

国家网络连续型出版物的**首批试点刊物**

**主管：**中国科学院

**主办：**中国科学院计算机网络信息中心

国内外公开发行人，中英文，**网络在线出版**。

国内统一刊号：CN 11-6035/N

国际标准刊号：ISSN 2096-2223。

- 2016年9月，JST中文数据库（日本科学技术振兴机构）
- 2016年，获得北京市音像电子网络出版物奖励扶持专项资金
- 2017年，入选“中国科学引文数据库(CSCD)来源期刊(2017-2018)"核心库

本刊致力于

- 科学数据的开放、交流和引用
- 推进科学数据的长期保存与数据资产管理
- 探索科学数据工作的有效评价机制
- 推动数据科学的发展



## ● 重点领域

- 学科领域
  - 地球系统科学
  - 生命科学与医学
  - 空间天文
  - 化学化工
  - 材料科学
  - 信息科学
  - 社会科学

重大科研项目产生/汇聚的基础性数据/数据集

大科学装置、大型仪器设备和各类野外台站等产生的数据集

科研院所、高等院校等组织机构长期积累的优质科学数据资源

对分散的、碎片化的数据进行系统加工整编形成的衍生数据集

在学术期刊已发表的重要学术论文相关联的基础数据、实验数据等

## 快速出版与引用

### I区论文

同行评议前发表的论文  
通过初审  
快速发表  
提供访问服务  
尽早使用数据

### II区论文

同行评议后出版的论文  
经过同行专家评议后正式收录

中国科学数据  
CHINA SCIENTIFIC DATA  
Volume 1, No. 2, 2017

2002 ~ 2016年高亚洲地区中大型湖泊微波亮温和冻融数据集  
A dataset of microwave brightness temperature and freeze-thaw for medium-to-large lakes over the High Asia region 2002 ~ 2016

推荐阅读

问渠哪得清如许, 为有源头活水来——《中国科学数据》发刊词

作者: 郭华欣  
出版年月: 2016年6月1日

人类正在进入大数据时代, 作为科学生产方式认识的重大革命性事件, 大数据为科学研究带来了新的方法论; 作为科学研究的革命性事件, 大数据正在改变人们认识世界的思维方式和发现, 在追求大数据驱动科学, 我国将大数据列为国家战略性新兴产业下, 《中国科学数据》应运而生。

中国有毒动物数据集

作者: 王林, 李之辉, 吴建强  
出版年月: 2017年11月17日

中国科学数据  
CHINA SCIENTIFIC DATA  
Volume 1, No. 2, 2017

A dataset of microwave brightness temperature and freeze-thaw for medium-to-large lakes over the High Asia region 2002 ~ 2016

Recommended Articles

"Pool mirroring the sky and clouds. From the source run fresh currents" — Introducing China Scientific Data

Authors: Guo Huaxing  
Published: June 1, 2016

A brand new era for mankind is dawning — the big data era. As a revolutionary innovation that transforms our understanding of scientific productivity, big data not only brings new methodologies into science.

Introduction: Snow, ice and environment over the Tibetan Plateau

Authors: Gao Yuehui  
Published: Nov. 22, 2017

中国科学数据  
CHINA SCIENTIFIC DATA  
Volume 1, No. 2, 2017

1960 ~ 2012年青藏高原极端气候指数数据集

周玉科<sup>1</sup>, 高琦<sup>2</sup>

**摘 要:** 青藏高原作为我国重要生态环境敏感区, 其气候变化及发展趋势一直是学术界的研究热点, 极端气候事件分析是其中一项重要研究领域内容。本文基于99个气象观测站1960 ~ 2012年的日逐气温(日最高气温、日最低气温)和降水观测资料(日降水), 经过数据质量控制和异常值剔除, 利用R语言中RCLIMdex包计算得到每季每型15项极端气候指数和8项极端气候指数, 其中包含15项极端气候指数和8项极端气候指数(包含15项极端气候指数和8项极端气候指数), 本数据集可以作为全球气候变化下青藏高原地区极端气候事件发生频率和趋势等分析材料, 也可作为基础数据探索极端气候事件对农牧业生产的影响。

**关键词:** 青藏高原; 极端气候指数; 气候变化; 气象站

中国科学数据  
CHINA SCIENTIFIC DATA  
Volume 1, No. 2, 2017

A dataset of climate extreme indices over the Tibetan Plateau (1960 ~ 2012)

Zhou Yuke<sup>1</sup>, Gao Qi<sup>2</sup>

**ABSTRACT** As an ecologically sensitive region with the highest altitude, the Tibetan Plateau has been a hotspot in the study of long-term climate changes. Extreme weather events are related to the regional climate change and are an important component of the study. Based on the daily temperature and precipitation observed at 99 meteorological stations over the Tibetan Plateau during 1960 ~ 2012, this study calculated a series of climate extreme indices. Firstly, quality control was performed on the raw meteorological observations to eliminate abnormal values. Then the R package RCLIMdex 1.0 was used to compute 15 temperature and 8 precipitation extreme indices. 10 indices were determined by the absolute observed value and 30 indices determined by the threshold approach. This dataset is useful for research on the occurrence frequency and developing trend of extreme events. When used together with regional statistical data, this dataset can also reveal the impacts of climate extremes on agricultural and livestock production.

**KEYWORDS** Tibetan Plateau; climate extreme indices; climate change; meteorological stations

| Dataset Profile      |                                                                                                   |
|----------------------|---------------------------------------------------------------------------------------------------|
| Chinese title        | 1960 ~ 2012年青藏高原极端气候指数数据集                                                                         |
| English title        | A dataset of climate extreme indices over the Tibetan Plateau (1960 ~ 2012)                       |
| Corresponding author | Zhou Yuke (zhouyuke@igpp.ac.cn)                                                                   |
| Data authors         | Zhou Yuke, Gao Qi                                                                                 |
| Time range           | 1960 ~ 2012                                                                                       |
| Geographical scope   | The Tibetan Plateau (26°00'N ~ 39°47'N, 73°19'E ~ 104°47'E), including 99 meteorological stations |
| Temporal resolution  | Year                                                                                              |
| Data volume          | 2.30 MB                                                                                           |



HOME

EDITORIAL BOARD

FOR AUTHORS

ABOUT

BROWSE






Published time

☒ All time
 ☐ Last 7 days
 ☐ Last 30 days
 ☐ Last 90 days

Paper status

☒ All time
 ☐ Zero (0)
 ☐ Zero (0)
 ☐ New Submission

Article type

☒ All time
 ☐ Data Paper(s)
 ☐ Comment & Response(s)
 ☐ Interview(s)
 ☐ Preprint(s)

Category

☒ All time
 ☐ Case Report(s)
 ☐ Comment & Response(s)
 ☐ Special Issue: Chinese Ecosystem Research Network(s)

Special Issue

☐ Special Issue: Emergency Management Operational Conceptual Application(s)
 ☐ Special Issue: Food, Safety and Environment Over The Tropic Tropics(s)
 ☐ Special Issue: Special Issue: National Citizens to Disaster Mitigation and Research(s)
 ☐ Special Issue: Bioenergy in China(s)

Subject

☐ All time
 ☐ Bioacoustic(s)
 ☐ Geoacoustic(s)
 ☐ Acoustic(s)
 ☐ Special Issue: Physics & Chemistry of The Marine Science(s)
 ☐ Information Science(s)
 ☐ Social Science(s)
 ☐ Climate(s)

Sort by downloads

Sort by views

Sort by published time



 [Special Issue: Biodiversity in China] Published: Nov 17, 2017 | Volume: 2017, Issue 4 | ● 1030 | ▲ 22

A dataset of venomous animals in China

Wang Lin, Li Wenhui, Zhu Jiansui

Keywords : Venomous animal, Species, Toxin, Active ingredient, Treatment

doi : 10.1155/jcsdata.2017.13.m

**Abstract :** This dataset of venomous animals in China integrates the data of taxonomy and toxin active ingredients of terrestrial venomous animals in China, as well as envenomation, poisoning and treatment.



**QUINTAN** [Special Issue: Snow, Ice and Environment Over the Tibetan Plateau] Published: Nov 22, 2017 | Volume:2017, Issue 2 | # 65 | 7

**Introduction: Snow, ice and environment over the Tibetan Plateau**

Qiu Yibao et al.

**Keywords:**

doi : 10.11912/qi.v2017.0002.zh

Paper Title | Zone II - Version ENL

# Daily fractional snow cover dataset over High Asia (2002 – 2016)

Qin Yubin · Wang Xingqun · Shi Lixin

Dataset title

Ver 2.0 (2017)  
DOI: 10.1155/2018/1732017  
Data DOI: 10.1155/2018/1732017-410

Reviewed: 2017-07-11  
Peer Review: 2017-07-31  
Published: 2017-08-20

16 0 0 0

## Abstract & Keywords

**Abstract:** High Asia, defined as the high altitude area (including mainly the Tibetan Plateau), is an important area of snow distribution in low- and mid-latitude regions. The dynamic changes of snow in this area have notable effects on water, energy balance and regional climate. As the assurance source in the Tibetan Plateau, the daily snow fraction and extends short duration. It is necessary to establish a change in snow cover on a daily time scale. This article describes a novel snow cover dataset named Daily Fractional Snow cover (FSC) data set over High Asia (2002 – 2016). The daily snow fraction and its short duration are derived from the MODIS normalized snow index data with a spatial resolution of 500 meters. Data processing involves the application of additional near-real time and a variety of snow cover cloud elimination algorithms. The new dataset provides a cloud-free estimate (cloud cover = 1%) of FSC over High Asia, which meets the real requirements of remote models. Binary snow products under clear sky conditions are selected as reference data to construct an inter-comparison of cloud distribution and total cloud cover in a series analysis. Results show that fractional snow cover products are mostly consistent with binary snow products in both space and time during snow accumulation, though some deviation exists in the snow ablation period. In the case of winter 2013, when the fractional snow cover was greater than 50%, the correlation coefficient of the two products was at 0.9268. This data set will be research to provide environmental, hydrological and energy balance and disaster assessment.

**Keywords:** Tibetan Plateau, fractional snow cover, MODIS, cloud cover

## Dataset Profile

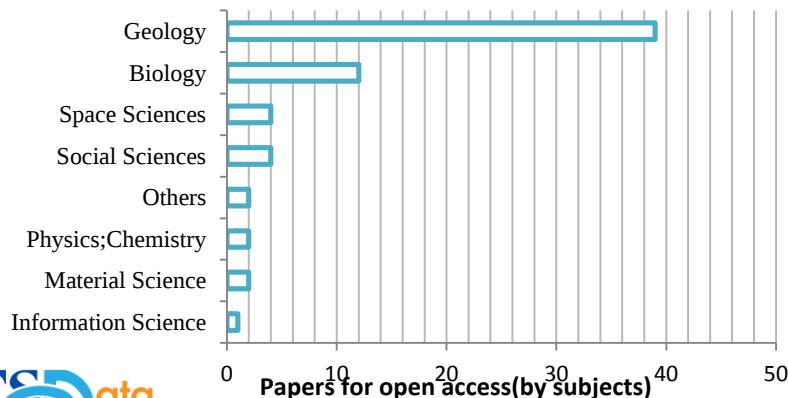
Download

|                      |                                                                                                                                                                                                                                                                                                                                                                  |             |         |  |  |
|----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|---------|--|--|
| Citation title       | 2002 – 2016年高亚洲区日积雪覆盖数据集 (2002 – 2016)                                                                                                                                                                                                                                                                                                                           |             |         |  |  |
| English title        | Daily fractional snow cover dataset over High Asia (2002 – 2016)                                                                                                                                                                                                                                                                                                 |             |         |  |  |
| Audience             | Qin Yubin, Wang Xingqun, Han Li, Chang Lu, Shi Lixin                                                                                                                                                                                                                                                                                                             |             |         |  |  |
| Corresponding author | Qin Yubin (yubin@niglas.ac.cn)                                                                                                                                                                                                                                                                                                                                   |             |         |  |  |
| Time span            | July 2002 – June 2016                                                                                                                                                                                                                                                                                                                                            |             |         |  |  |
| Geographical scope   | High Asia (28°–40°N, 62°–100°E) is a high altitude area of Asian comprising of the Tibetan Plateau, Kailash or Gargil-Potat Plateau, the Helianshan, Pamirs, the Tatarian mountains, and the Gobi Altai Basin. The following countries and regions include China, Nepal, Bhutan, India, Russia, etc.                                                             |             |         |  |  |
| Spatial resolution   | 500 m                                                                                                                                                                                                                                                                                                                                                            | Data volume | 22.9 GB |  |  |
| Native format        | GeoTIF                                                                                                                                                                                                                                                                                                                                                           |             |         |  |  |
| Data service system  | < <a href="#">http://www.knowledgetoearth.com/dataset/410&gt;</a>                                                                                                                                                                                                                                                                                                |             |         |  |  |
| Source of funding    | Supported by the Scientific Fund for Universities Scientific Research in the Public Interest Constructing a Remote Sensing Product Dataset for Snow Cover State Tibetan Plateau (No.10YJC920040), the International Partnership Program of the Chinese Academy of Sciences (No.11C030821000) and the National Natural Science Foundation of China (No.41371251). |             |         |  |  |
| Related components   | The dataset consists of 16 packages (2002_2016), stored by year: HMA_FMODIS_FSC_2002-06, HMA_FMODIS_FSC_2003-06, ..., HMA_FMODIS_FSC_2016-06, etc., with a raw data amount of 22.9GB.                                                                                                                                                                            |             |         |  |  |

## 1. Introduction

Snow cover is one of the fastest changing natural elements on earth, which has important effects on the circulation of freshwater,

# 出版情况



新

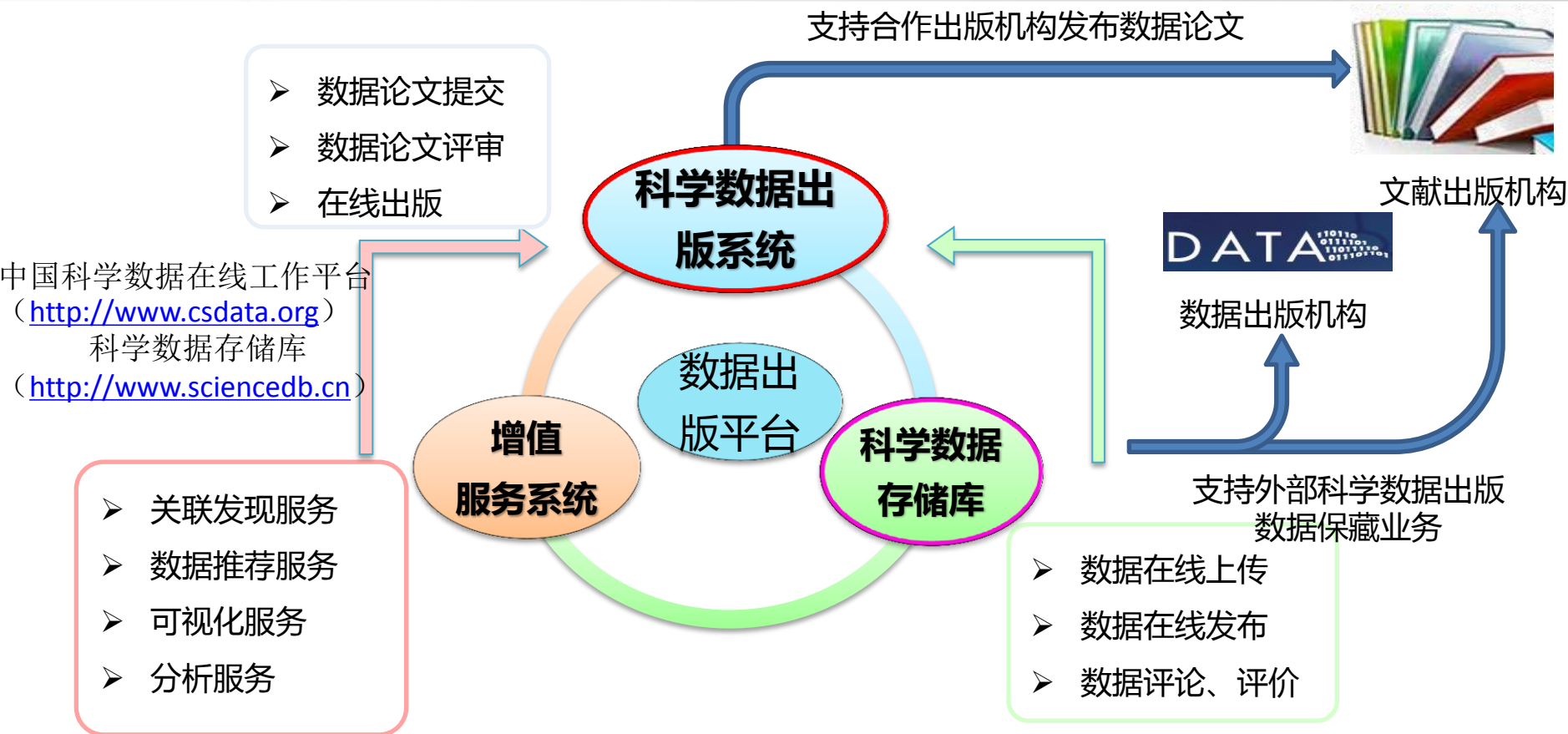
快

- ✓ **98.11%**文章数据集本刊首发;
- ✓ 首次全文发布时间：**34.1**天
- ✓ **70%**ScienceDB Web, 3.33%FTP, 26.67%其他;
- ✓ **100+**家单位来稿, **700+**位作者;
- ✓ **Speical issues**
  - ✓ 中国生态研究网络专题数据 (2017.1)
  - ✓ 青藏高原雪、冰和环境专刊 (2017.2)
  - ✓ 中国生物多样性数据集 (2017.4)

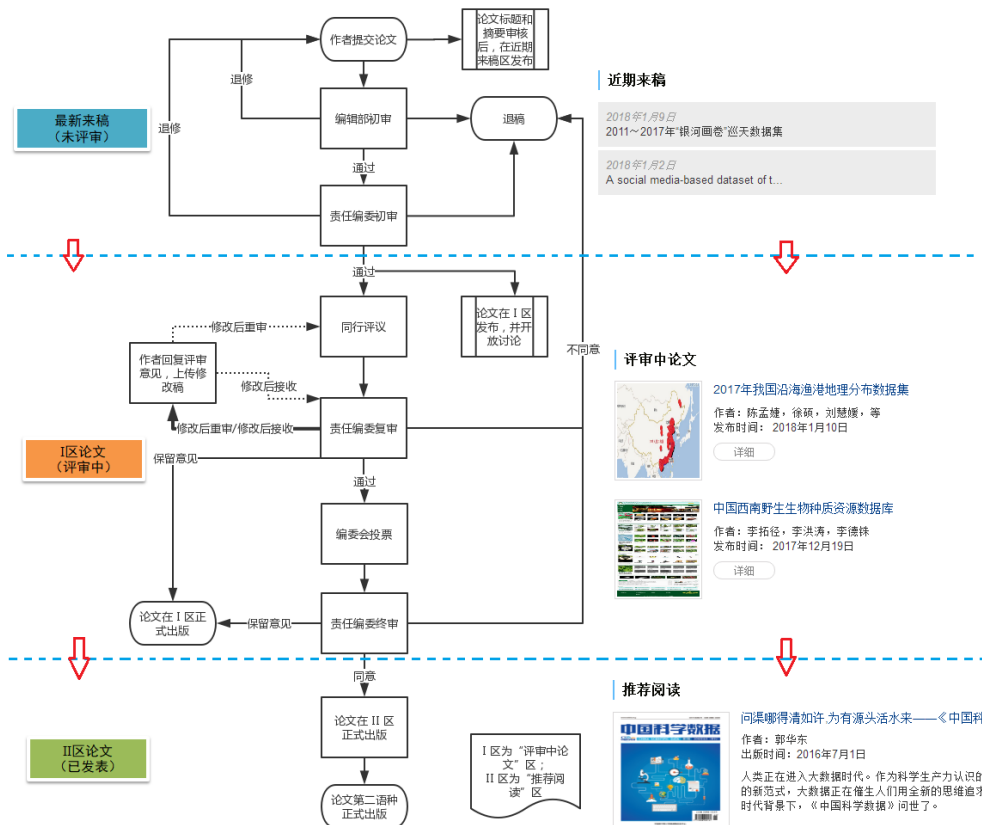
## • 三步一站式提交



# 中国科学数据出版平台



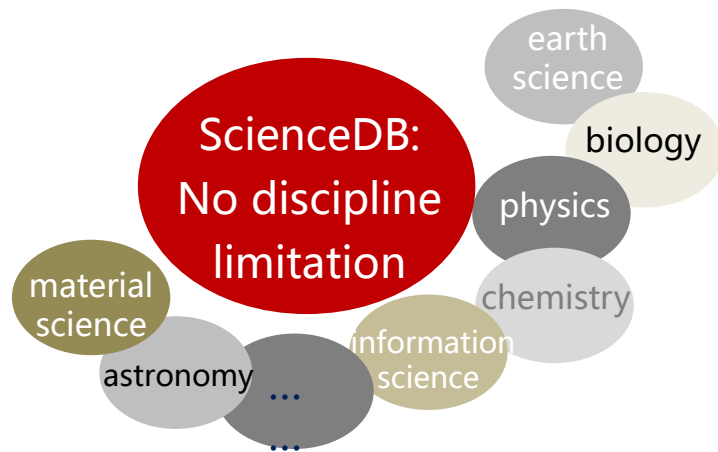
2015年8月上线，2016年9月，进行了系统迁移和流程优化,2017年11月，新系统



|       |                                                             |
|-------|-------------------------------------------------------------|
| 近期来稿  | 摘要通过编辑部形式初审<br>部分发表<br>为作者争取首发权                             |
| I区论文  | 同行评议前发表的论文<br>全文发表<br>通过编辑部内容初审<br>快速发表<br>提供访问服务<br>尽早使用数据 |
| II区论文 | 同行评议后出版的论文<br>全文发表<br>经过同行专家评议后正式收录<br>收录至CSCD              |

# ● 数据存储库ScienceDB

- **在线通用存储库** [www.sciencedb.cn](http://www.sciencedb.cn)
- **组织模式**
  - 与期刊进行合作，配合其工作流程，存储相应数据。
  - 重大项目和科研团队的数据可设置一定的保护期。
- **主要功能**
  - 支持多数据格式，部分格式可在线预览。
  - 数据关联：为数据分配DOI，也可以与对应期刊和专业数据存储库建立或识别其他关联(如基因库等)。
  - 数据使用：内容检索、引用和评价，提供API。
  - 标准规范：策划文件和元数据内容，遵循国际标准。
  - 长期保存：对数据长期保存，无限期保证访问。
  - 个性定制：为合作刊提供特定配置。
  - 语种服务：中英文双语版本，支持国际化。
- **基本规则**
  - 遵循CC0许可协议，可免费下载。数据集最终开放





- 数据论文和数据集遵循CC BY 4.0协议（Creative Commons Attribution 4.0 International License，署名共享）
- 元数据遵循CC0协议
- 其他CC种类：CC BY-NC-ND-SA

| Item       | License   |
|------------|-----------|
| Data paper | CC BY 4.0 |
| Dataset    | CC BY 4.0 |
| Metadata   | CC0       |

- data paper
  - metadata
  - full-text
- dataset
  - metadata
  - data files or a valid DOI

- 同行专家评议
  - 评审意见透明公开
- 读者在线评议

论文评审里程



## 读者评论信息

暂无讨论

提交问题或建议：

输入意见

提交

## 同行评议信息

2017-07-13  
来自稿件 版本 1

责编:

一、对数据集的意见

1. 在论文的“数据使用方法和建议”里面介绍“本套数据集共享了高亚洲地区2002~2016年每天的积雪覆盖产品”，但是在数据集中并不是有每天的数据，比如2002年，是第188天-365天的数据，2016年是第001天-118天的数据，如果是论文中的数据量的话，建议作者补充数据；

2017-07-18  
来自稿件 版本 1

作者:

一、数据集修改

1. 回答：已将数据库（集）基本信息简介中数据时间范围填写为：2002年7月至2016年6月。该数据产品为每日产品，由于MODIS上下午星同时观测时间为2002年7月开始，在处理时按照积雪年为基准。为节省空间，在保存时采用了每年度一个压缩包的方式做压缩存储。

2017-08-04  
来自稿件 版本 2

该数据集提供了最新的高亚洲区域主日积雪覆盖度，对于第三级气候、水文水文及环境相关研究具有重要意义。论文结构合理、论述清晰。因为“高亚洲”已在题目中，建议用“青藏高原”替换“高亚洲”作为关键词，以利于检索。

网评专家二：

本数据是逐日积雪数据，云对积雪提取影响很大，建议将相同时期的云数据一同提取出来。

2017-09-04  
来自稿件 版本 2

作者:

1. 回答：已将关键词中“高亚洲”替换为“青藏高原”。

2. 回答：数据集侧重于对积雪数据的提取，由于云数据变化速度快、类型多、情况复杂，所以本数据集不将云数据作为产出结果，只作为数据质量评估的参考依据。

3. 针对“建议作者补充2016年的数据。”回答：已在数据集部分、数据集论文对应部分进行补充与修

2017-09-20  
来自稿件 版本 2

责编:

经责编投票，同意录用本文。



## 我们**关注**



数据的复用价值



数据生产方法



数据对相关研究结论的支持性



数据的质量和加工处理



## 我们**不**关注



研究的创新性



不对与数据集相关联的结果进行主观判断



多大size的数据  
可以发？

# 同行评议

## 论文

- 论文质量与可读性
- 论文所描述的数据集的共享与利用价值
- 论文对数据加工处理方法描述的完整性
- 论文价值
  - 数据重用
  - 数据加工处理方法的新意及借鉴价值
  - 数据对科学研究成果的支持验证作用

## 数据与论文 - 一致性

| 评价项目          | 主要评价指标                                                                    | 评分 (0-10) |
|---------------|---------------------------------------------------------------------------|-----------|
| 一、论文          |                                                                           |           |
| 数据的可重用价值      | 数据可重用价值高；方法和数据处理的每一步的步骤描述都足够详细清晰，足以使他人复制这些步骤；提供了让他人重用数据集或者与其他数据整合所需的所有信息。 | ☆☆☆☆      |
| 数据加工处理方法有参考价值 | 数据加工处理方法严谨合理，有新意和借鉴意义                                                     | ☆☆☆☆      |

### 2000-2014年中国地表温度月值数据集

赵江华<sup>1</sup>

同行评议: 参考建议: 1、摘要中“目前中国尺度1km分辨率地表温度月值数据还是空白”建议慎重, 毕竟还是有同类产品的(例如基于站点数据空间化的月值数据等)。2、摘要中“农林牧渔”可改为“农业”。3、参考文献中, “杨振刚, 武亚军, 闫振峰. 基于地表温度植被指数特征空间的区域土壤干湿状况[J]. 生态学报, 2009, 29(3): 1205-1216.”的作者和文章名称写错了, 需修改。

提交问题或建议:

输入意见

您还没有登录, 请[登录]或[注册]!

|              |                                                                                              |        |
|--------------|----------------------------------------------------------------------------------------------|--------|
| 数据加工与丰度(完整性) | (或)完整性是否充分覆盖? 数据值是否落在预期范围内?<br>该数据是否含有明细的错误?<br>(如果需要)是否提供了关于数据质量方面可信的技术验证实验、数据质量统计分析以及误差分析? | ☆☆☆☆   |
| 数据一致性        | 数据集是否与数据论文中的描述一致?                                                                            | ☆☆☆☆   |
| 合 计          |                                                                                              | (分数统计) |
| 处理意见         |                                                                                              |        |
| 发 表          | 1. 全文发表( )<br>2. 不必退修, 编辑加工后发表( )                                                            |        |
| 退 修          | 1. 退修后发表( )<br>2. 退修后再审( )                                                                   |        |
| 再 审          | 建议送 _____ 单位 _____ 教授<br>再 审                                                                 |        |
| 退 稿          | 1. 不适合本刊发表( )<br>2. 建议转投他刊( )                                                                |        |

数据审核专员 + 专家评审 + 开放公众评议

| 评价项目          | 主要评价指标                                                                                                                                              |
|---------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>一、论文质量</b> |                                                                                                                                                     |
| 题目            | 能准确反映文章内容，简明扼要。                                                                                                                                     |
| 摘要&关键词        | 简明扼要，涵盖主要信息。                                                                                                                                        |
| 论文的可读性        | 论文行文流畅，包含了数据的处理方法、数据的描述及规范性等主要内容，有助于对数据的理解。                                                                                                         |
| 数据的可重用性描述     | 数据价值高，方法和数据处理的每一步的步骤描述都足够详细清晰，足以使他人复制这些步骤；提供了让他人重用数据集或者与其他数据整合所需的所有信息。                                                                              |
| 数据加工处理方法描述    | 数据加工处理方法的完整性，是否严谨合理，有新意和借鉴意义。                                                                                                                       |
| 论文价值          | 数据论文具有较高价值，如学术价值及社会价值等。                                                                                                                             |
| <b>二、数据质量</b> |                                                                                                                                                     |
| 数据存储          | 数据存储在可靠且适合的存储库中；作者存放的数据文件齐全，并与数据论文描述相符。                                                                                                             |
| 数据查看          | 该数据集可以访问和查看（便于读者快速阅览，且包含主要元数据信息的预览展示页面）；用于查看数据的软件应提供了包括版本信息在内的相关信息等。                                                                                |
| 数据质量与丰度(完整性)  | 数据的生产方法严谨、合理；<br>数据格式恰当合理，符合业内标准；<br>根据作者的研究内容，数据的深度、范围、大小及（或）完整性应充分覆盖，数据值应落在预期范围内；<br>该数据不应含有明细的错误；<br>（如果需要）提供了关于数据质量方面可信的技术验证实验、数据质量统计分析 & 误差分析。 |
| 数据一致性         | 数据集与数据论文中的描述一致                                                                                                                                      |

- **明晰的数据产权关系**
- **严谨的实验与较高的数据质量**
  - 数据的产生与加工方法是否严谨
  - 数据的可靠性和质量控制
  - 完整充分的背景信息，如特定领域数据集的时间/空间属性信息以及数据应用建议等
- 提交到《中国科学数据》的数据论文应当提供论文描述的数据集细节，包括数据集收集、加工处理方法，数据质量评估和验证的方法，便于理解和使用数据的相关信息等。但是，论文中**不应当**包括新的科学假设、新的观点和新的科学方法等内容（一般情况下，此类信息应体现在相关学术论文中）。
- **规范一致的描述**
  - 对研究方法、数据处理步骤以及数据资源、过程、辅助软件、数据文件类型等信息进行充分而具体的描述，确保数据的可重复试验
  - 数据重用或数据整合所必要的信息
  - 元数据、数据集、其他数据描述以及数据论文等的协调一致与规范性
- **数据存储准备得当**
  - 数据论文中所涉及数据集的完整性
  - 数据应存放于最合理的存储库中，确保数据的可访问性和可靠性
  - **数据也应无政治敏感问题（如国界）**
- 数据论文在开放评议后，其描述的数据集将**公开共享**（涉及人类的隐私问题或公众安全等信息除外）。

- 理想时机-受众角度
  - 热点数据（地区、话题、领域、方法、工具）
  - 稀缺数据（技术优化、算法改进、领域空白）
  - 持续数据（已有规模的扩展、连续与完整性）
- 成熟时机-作者角度
  - 基于数据的研究论文已发表
  - 已有数据要进入休眠期
  - 需要第三方稳定的数据服务环境/技术支持
  - 希望通过同评把握数据质量

- 标题
- 作者及单位信息
- 摘要
- 数据库（集）基本信息简介
- 引言/概述
- 数据采集和处理方法
- 数据样本描述
- 数据质量控制和评估
- 数据价值（可选）
- 数据使用方法和建议（可选）
- 致谢（可选）
- 数据作者分工职责
- 参考文献
- 数据引用格式
- 论文引用格式

X 标题大，内容小

✓ 标题简明、必要时体现时空范围

✓ 数据作者可以和论文作者不一致

✓ 数据价值、数据使用方法和建议二选一

✓ 数据和论文引用格式可由编辑完成

# ● 数据库（集）基本信息简介

|                                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 数据库(集)中文名称<br>English title              | 2011 ~ 2013夏季南黄海大型底栖软体动物数据集<br>A dataset of macrobenthic mollusks of southern Yellow Sea in summer 2011 ~ 2013                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| 数据作者<br>Data author(s)                   | 徐勇、李新正、王洪法、张宝琳<br>Xu Yong, Li Xinzheng, Wang Hongfa, Zhang Baolin                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| 通讯作者<br>Corresponding author             | 李新正 Li Xinzheng,<br>(lixzh@qdio.ac.cn)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| 数据时间范围<br>Time range                     | 2011 ~ 2013                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| 地理区域<br>Geographical scope               | 南黄海东经124°以西海域(北纬37°00'00" ~ 33°00'00", 东经120°30'00" ~ 124°00'00")。<br>West of 124°E, southern Yellow Sea (37°00'00" ~ 33°00'00"N, 120°30'00" ~ 124°00'00"E)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| 数据量<br>Data volume                       | 0.17 MB                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| 数据格式<br>Data format                      | *.xls                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| 数据服务系统网址<br>Data service system          | <a href="http://www.sciencedb.cn/dataSet/handle/479">http://www.sciencedb.cn/dataSet/handle/479</a>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| 基金项目<br>Source(s) of funding             | 中科院先导科技专项(A)“黑潮及其变异对中国近海生态系统的影响”(XDA11020303)、海洋公益性行业科研专项(201505004-1)、青岛海洋科学与技术国家实验室鳌山科技创新计划项目(2015ASKJ01)<br>The Strategic Priority Research Program of the Chinese Academy of Sciences (XDA11020303), the Ocean Public Welfare Scientific Research Project of State Oceanic Administration of the PRC (No. 201505004-1), the Scientific and Technological Innovation Project funded by Qingdao National Laboratory for Marine Science and Technology (No. 2015ASKJ01)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| 数据库(集)组成<br>Dataset/Database composition | 南黄海采泥底栖软体动物资料201108.xls, 数据量54K;<br>南黄海采泥底栖软体动物资料201208.xls, 数据量52K;<br>南黄海采泥底栖软体动物资料201308.xls, 数据量70K。<br>3个文件包括29个站位(3个航次共获得70个沉积物样品, 每个沉积物样品包含一个或多个生物样品, 每个生物样品对应一条数据), 172条数据。数据包括的参数有: 站号、站经度、站纬度、深度、类别、科中文名、科拉丁名、种中文名、种拉丁名、采集日期、标本个数、密度、标本重量、生物量、底质、海区、箱式采泥器规格、采泥次数、取样面积、调查船。<br>Benthic mollusk data collected by box corer in southern Yellow Sea in August 2011.xls, with a data volume of about 54 K;<br>Benthic mollusk data collected by box corer in southern Yellow Sea in August 2012.xls, with a data volume of about 52 K;<br>Benthic mollusk data collected by box corer in southern Yellow Sea in August 2013.xls, with a data volume of about 70 K.<br>A total of 29 stations (70 sediment samples were collected during 3 cruises; each sediment sample included one or more biological samples; each biological sample corresponded to one item) were investigated and 172 items were recorded in three files. The parameters in the dataset include station number, longitude, latitude, depth, group, family name in Chinese, family name in Latin, species name in Chinese, species name in Latin, sampling date, number of samples, density, weight of samples, biomass, sediment type, sea area, corer specification, sampling times, sampling area, and investigation vessel. |

✓ 中英文双语，与数据的元数据一致

✓ 数据作者可以为团体

✓ 压缩文件可将数据格式表述为“\*.shp、\*.xls（压缩为zip格式）”

✓ 填写数据集存放的网址，可由编辑填写

✓ 可根据研究领域特征，修改条目名称（如空间分辨率）

✓ 数据集组成是相对独立、完整的部分；应定量描述，如包括几个文档、多少个参数。

## 内容要求

- 简述同类工作的研究进展、研究意义
- 简述本数据集产生的背景/支撑条件/方法
- 简述与现有数据集相比，本数据集的优势、创新点、价值体现
- 引用参考文献

- 详细描述获得目标数据所需进行的研究设计、完成实验、生成数据所需的全部处理步骤中所使用的方法、程序、算法与分析等，利于读者理解和重复试验
- 鼓励作者提供图例，展示相关研究和实验设计的概况
- 可根据需要划分多个子标题
- 原则上不鼓励基于数据集的衍生研究，如新的研究方法等的讨论
- 如果是对数据集进行加工，建议列出源数据名称或信息（如在线可获取，需提供链接地址，保证可访问）

度则是表现人口空间分布最主要的形式和衡量人口分布地区差异的主要指标<sup>[2]</sup>。准确、精细的人口密度数据能够为区域规划与开发、灾害风险救援、健康科学和资源环境承载力评估等研究和应用提供支撑。

当前，县级平均人口密度是人口问题研究中常用的指标。乡镇（街道）处于中国行政体系的基层位置，也是全国人口普查数据公开发布的最小单元。较县级平均人口密度而言，乡镇（街道）级平均人口密度能够更加客观细致地刻画中国人口分布的格局和态势，也可作为区域、县级尺度的人口问题研究提供数据支撑。具有重要的科学价值和广泛的应用领域。目前，中国尚无共享的较大范围的乡镇（街道）人口密度空间数据集。主要原因包括乡镇（街道）界线数据的限制使用、乡镇（街道）人口统计数据的缺乏及乡镇行政区划调整等<sup>[3]</sup>。国家科技基础条件平台——地球系统科学数据共享平台收集了中国 2000 年的乡镇（街道）级行政界线数据，国家统计局发布了第五次人口普查乡镇（街道）级人口统计数据，本文基于这两个基础数据库，经过标准化处理之后形成中国 27 省乡镇级人口空间基础数据，计算获得各乡镇（街道）级平均人口密度，形成了本数据集。

## 1 数据采集和处理方法

中国 27 省乡镇（街道）级人口密度数据集利用的基础数据集包括国家科技基础条件平台——地球系统科学数据共享平台提供的 2000 年全国 27 省乡镇（街道）行政界线数据集（<http://www.geodata.cn>），及国家统计局发布的第五次全国人口普查乡镇（街道）级人口统计数据（<http://www.stats.gov.cn/tjsj/ndsj/remk.asp?ch=2000jiedao/jiedao.htm>）。前者通过离线共享获得，后者为在线下载资源。处理流程包括乡镇（街道）界线数据处理、乡镇（街道）人口统计数据处理和平均人口密度计算 3 个部分。乡镇（街道）界线数据的处理包括原始数据分析、空间参考统一和几何纠正。乡镇（街道）人口统计数据的处理主要是虚拟乡镇标定。在这两项工作基础上，综合处理乡镇（街道）界线数据和乡镇（街道）人口统计数据，

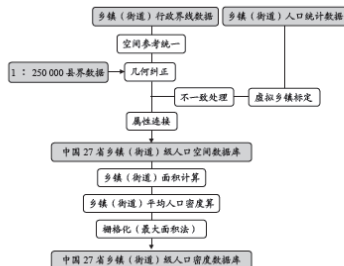


图 1 数据集生产技术路线图

消除二者不一致的情况，经过编码、属性连接等形成中国 27 省乡镇（街道）级人口空间数据库。计算各乡镇（街道）面积，进而计算各乡镇（街道）平均人口密度，栅格化形成本数据集。主要的技术流程如图 1 所示。

## 1.1 乡镇（街道）界线数据处理

1. 初始资料获取与分析。首先向国家科技基础条件平台——地球系统科学数据共享平台申请乡镇界线数据。空间范围覆盖辽宁、吉林、内蒙古（部分地区）、北京、天津、上海、河北、河南、陕西、宁夏、山西、山东、安徽、江苏、湖南、湖北、江西、浙江、福建、广东、海南、云南、贵州、四川、重庆、青海、西藏共 27 个省（直辖市、自治区）。该数据集以 1:25 万、2000 年的数据为参考原则，数据质量良好，但其数据来源、时间和比例尺并不完全相同。

2. 空间参考统一。将各省数据集的空间参考统一为 Albers Equal Area 投影，地理参考为 WGS84 坐标系。

3. 几何纠正。以 1:25 万全国县界数据（2000 年）为标准数据，分省对原始乡镇界线数据进行几何纠正。根据省区面积各省约选取 300~600 个配准点，保证纠正完成输出的乡镇界线和县界空间偏差不得超过 500 m。

## 1.2 乡镇（街道）人口统计数据处理

1. 虚拟乡镇标定。对于普查数据中的经济开发区、虚拟普查片区等无明确行政边界范围的虚拟乡镇单元，通过查阅文献及空间位置判定两种手段，将其人口归并到具体乡镇（街道）单元中。

2. 乡镇空间单元与统计数据不一致情况的处理。分省逐县区核对统计乡镇空间单元与统计数据对应情况，对于二者不一致的情况，分以下三种情况处理：

- 1) 乡镇界线数据表征时间晚于 2000 年且该县在 2000 年之后发生过乡镇撤并和拆分，查阅资料，依据调整情况合并乡镇人口统计数据以匹配乡镇界线数据；
- 2) 乡镇界线数据表征时间早于 2000 年且该县在该表征时间至 2000 年之间发生过乡镇撤并和拆分，查阅资料，依据调整情况合并乡镇界线数据以匹配人口统计数据；
- 3) 某些县（区）乡镇（街道）由于缺少详细的边界数据、划分过程，查阅资料，合并统计数据以匹配乡镇界线数据。这种情况多发生在面积较小且人口密度或边界模糊的街道。

## 1.3 平均人口密度计算

1. 分别对乡镇空间界线数据和人口统计数据按照省名称—县名称—乡镇（街道）名称顺序排序，并按顺序赋予各乡镇空间单元及统计属性唯一乡镇代码；

2. 利用乡镇代码为关键字，将各乡镇的人口统计属性赋予其空间多边形单元；

3. 计算各乡镇单元的面积；

4. 依据公式（1），计算各乡镇平均人口密度：

$$PD = \bar{N}_p / A \quad (1)$$

式中： $PD$  为人口密度（人/平方千米）； $\bar{N}_p$  为乡镇（街道）人口数量（人）； $A$  为乡镇（街道）面积（平方千米）。

## 1.4 栅格化处理

在 ArcGIS 中，基于上一步计算的乡镇平均人口密度，计算 100 m×100 m 格网尺寸上的人口分布，将人口密度图层栅格化为百米格网人口密度图层；基于百米格网人口密度图层，利用栅格聚合功能，生成 1 km×1 km 的乡镇（街道）级平均人口密度数据。

## 2 数据样本描述

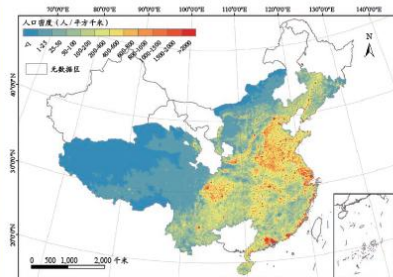


图 2 中国 27 省 2000 年乡镇（街道）级平均人口密度图

经过处理，中国 27 省 2000 年乡镇（街道）级平均人口密度数据如图 2 所示；数据样本：北京市 2000 年乡镇（街道）级平均人口密度数据（图 3）。

## 3 数据质量控制和评估

主要通过以下手段对数据集进行质量控制：

1. 原始中国 27 省乡镇界线数据在生产时具有严格的质量控制规范（参见其数据文档）；第五次人口普查乡镇（街道）级人口统计数据由中国统计局发布，保证了其可靠性和规范性。

2. 分省对乡镇界线进行几何纠正所参考的 1:25 万县界数据集来源于中国测绘与地理信息局，纠正过程中，选择了充足的特征点（各省边界面积在 300~600 个），保证纠正后的乡镇边界与县界的误差不得超过 500 m。

3. 严格匹配了乡镇单元空间界线与人口统计属性，对于不一致情况进行了规范化处理，处理依据中国民政部或各级政府组织所发布的乡镇行政区划变迁文件。

数据质量评估：  
经统计，本文研究区的原始人口普查数为 1 103 880 920，栅格化后的总人口数为 1 103 493 334.3，栅格化过程损失人口为 387 586，损失量约为 0.035%。可以认为，该数据集质量可靠，精度较高。

# 《中国 27 省乡镇（街道）级人口密度数据集（2000 年）》

- 具体给出数据集的典型样本，描述来源、数据结构等，帮助读者溯源定位、理解数据集内涵
- 如果数据集包含多种数据结构的数据，应当对每种数据结构的数据进行数据样本举例说明
- 数据样本的选择应尽可能选取富有典型代表意义且体量适当的数据样本，便于在有限的篇幅中做最充分而精炼例证说明

### 2 数据样本描述

西双版纳民间传统利用植物信息数据集的内容涉及西双版纳地区民间常用植物的科学分类、民间分类、民间对植物的具体用法等数据，收集了民族植物学引入中国以来 30 多年间的研究数据，主要包括修、哈尼、基诺、拉祜、布朗等 8 个民族关于 1071 种植物（含藻类与真菌）的传统利用数据。为了增强数据的可用性，本数据集对原始数据进行了规范化处理：将植物名录进行整合，统一采用现行的 APG III 系统；对西双版纳地区民族植物学研究的主要成果——植物编目信息进行统一化处理。

西双版纳民间传统利用植物信息数据集提供原始参考文献，供用户查询数据的来源及其他信息。该数据集中植物学名均为植物的合法拉丁名，植物科名与中文名主要参考《中国植物志》（中英文版）、《云南植物志》、《西双版纳高等植物名录》等专著。

对于英文文献中所存在的中文名不全或有误的情况，参考《中国植物志》、《云南植物志》等专著对其进行了增补，但仍有 33 个物种原产地非中国，因而未查到中文名，数据集中以“-”代替中文译名。

地方名为各民族名的音译，以汉语拼音的形式展现，实际发音可能因地区差异略有不同。用途类型分为食用、药用、他用三类，各类型中又细分为具体用途，如食用植物可分为蔬菜、香料、水果、坚果等。

民间的传统利用信息以利用部位、利用方法、具体用途等字段来体现。其中利用部位主要以植物的根、茎、叶、果实、种子、花等器官来区分，也包括一些具体植物组织（树皮）或细胞后含物（树脂树胶等）；利用方法如食用植物主要有炒食、生食等；药用植物主要有捣烂外敷、煮水内服、煮水擦洗等。具体用途则反映民间对植物的具体利用，如蔬菜的食用，或草药具体治疗的病症等。

典型数据集的字段详见表 1。

表 1 植物传统利用信息数据集

| 数据项  | 示例                          |
|------|-----------------------------|
| 植物学名 | <i>Acorus calamus</i> L.    |
| 科名   | Acoraceae                   |
| 中文名  | 菖蒲                          |
| 地方名  | Hanhao (傣族), Longxia (哈尼族)  |
| 利用部位 | 全株: 根、叶                     |
| 利用方法 | 煮水                          |
| 具体用途 | 全株煮水治疗胃病; 根、叶煮水治疗腹泻、消化不良、胃病 |
| 利用民族 | 傣族, 哈尼族                     |
| 参考文献 | 潘玉梅, 2006; Ghorbani, 2011   |

《西双版纳民间传统利用植物信息数据集》

- 提供资料证明数据的可靠性，包括数据质量控制方法和验证过程，列明误差数据和相关精度
- 对于连续性数据中的数据缺失、断点，数据的取舍等情况进行必要说明

### 3 数据质量控制和评估

在聚合物溶液汽-液平衡数据库中收入的数据均为在文献中公开发表的实验数据。尽管目前尚没有对聚合物溶液汽-液平衡数据进行热力学一致性检验的方法，但仍有一些规律可循，如聚合物溶液中由于过量焓是个很大的正值，因而与Raoult定律形成负偏差，这使得溶剂活度小于1。该数据库对数据的筛选主要根据经验及已掌握的规律。对有明显错误（如印刷错误）或不符合规律的数据予以剔除。

| 聚合物的溶液            | (续表)      |                  |                |
|-------------------|-----------|------------------|----------------|
|                   | 数据项       | 实例               |                |
| 文献原始数据<br>及溶剂校正活度 | 溶剂质量分数 w1 | a1 (uncorrected) | a1 (corrected) |
|                   | 0.0570    | 0.2210           | 0.2222         |
|                   | 0.0727    | 0.2630           | 0.0243         |
|                   | 0.0973    | 0.3360           | 0.3375         |
|                   | ...       | ...              | ...            |
|                   | 1.0000    | 1.0000           | 1.0000         |

以任何其他方式（如内插或外推得到的数据）以及由其他途径（如书籍、手册、综述文章）为来源的数据均被认为是二手数据，因其作者对数据加工的可信性无从掌握，因而不被收入聚合物溶液汽-液平衡数据库。

此外，为保证数据质量，对出现以下情况的汽-液平衡数据不予加工处理：

1. 实验数据点的数量 < 6；
2. 聚合物的定义不明确；
3. 溶液组成采用体积分数表达，但未提供聚合物密度的数据；
4. 溶剂处于超临界状态；
5. 优化过程不收敛或结果不合理。

### 4 数据使用方法和建议

如果聚合物溶液汽-液平衡数据库中恰有符合使用者要求的数据，或与使用者要求相近的数据，使用者可直接采用该数据库提供的数据。使用者也可采用该数据库提供的汽-液平衡数据，用于模型参数优化和模型验证。对于工程应用，使用者则可根据该数据库提供的 UNIQUAC 模型参数，通过计算直接获得数据的内插和外推结果，同时还能掌握采用 UNIQUAC 模型及其参数所能达到的计算精度，这对工程应用是十分便利的。

实际上，聚合物溶液汽-液平衡数据库早在上世纪 90 年代就已经应用于 Sørensen 等人对聚乙二醇的气相色谱研究<sup>[10]</sup>、Chen 等人对基团贡献法 Flory 模型的研究<sup>[11]</sup>、Zhong 等人对简化胞腔理论的研究<sup>[12]</sup>。

访问网址：

溶剂无限稀释数据：[http://enginchem.chemdb.csdb.cn/sdb\\_2004/pdb\\_infinity.html](http://enginchem.chemdb.csdb.cn/sdb_2004/pdb_infinity.html)；

汽-液相平衡数据：[http://enginchem.chemdb.csdb.cn/sdb\\_2004/pdb\\_vle.html](http://enginchem.chemdb.csdb.cn/sdb_2004/pdb_vle.html)。

《聚合物溶液汽-液平衡数据库》

### 3 数据质量控制和评估

传统的数据质量控制主要根据气象学、天气学、气候学原理，以气象要素的时间、空间变化规律和各要素间相互联系的规律为线索分析气象资料是否合理。其方法包括：范围检查、极值检查、内部一致性检查、空间一致性检查、气象学公式检查、统计学检查、均一性检查。这些方法被普遍应用到气象资料的质量控制中<sup>[7]</sup>。

数据质量控制过程包括对数据集的检查、时间序列检查、单个数据点的检查。对数据的检查包括明显的数问题（如文件被切断、格式化错误、存储损坏）等。时间序列检查时，检测平均值变化使用了累积总和检测法，通过找出时间序列平均数的变化来检查。方差变化检查使用了尺度累积量（SCUSUM），该量是CUSUM的一个变量，用来查找站时间序列的方差变化。在单个数据点的检查中，辨别时间异常值使用了双权重标准差用作计算方差，这种方法既能很好地抵抗异常值，又能保持整体的功效<sup>[8]</sup>。

在野外，仪器易被停电和雷击等因素干扰，造成数据缺失，或者因突然断电和供电造成的数据突然异常。在这样情况下，我们往往采取剔除异常数据，和毗邻的气象站相互对比，拟合恢复数据，保证数据连续性（图3）。

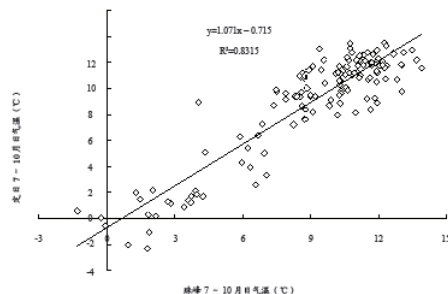


图3 数据拟合

针对青藏高原研究所开展的常规观测项目，在数据入库阶段建立了完善的ETL（Extract-Transform-Load：数据抽取-转换-装载）过程（图4），保证已入库数据的完整性和一致性。观测阶段，由长期工作在台站的专业人员进行观测和观测仪器维护工作，数据整理和入库过程的质量控制方面，主要分为两个步骤：（1）进行了各种数据集的集成、整理、转换、统一格式；（2）去除随机及系统误差。通过一系列质量控制的计算法则，进行自动和人工的质量控制。使用的质量控制方法包括变量检查、极值检查、内部一致性检查，以保障数据的质量。元数据由台站数据观测和处理人员根据数据中心制订的元数据标准格式（包含数据名称、时间范围、字段说明、观测方法、知识产权等）进行填写，并聘请相关领域的研究人员对元数据描述进行审核。

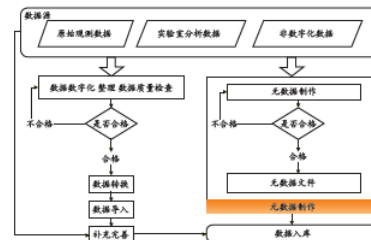


图4 数据质量控制

## 《青藏高原野外台站大气物理观测数据》

## 内容要求

- 通过和国内外相关数据集的比较，从数据覆盖范围、数据加工方法等角度说明本数据集的价值，体现数据来源、数据整理加工过程、质量控制等方面的创新性
- 尽可能多地挖掘数据潜在的应用价值
- 应比引言中的介绍更详尽

## 内容要求

- 包括数据打开/分析软件的相关使用描述、数据的深加工步骤、数据集成或数据比较的说明
- 如有必要，鼓励作者上传代码、程序或数据处理的工作流程作为补充资料，以帮助其他研究人员重用数据
- 对于涉及人类隐私或公众安全等问题的数据，本部分应详细描述相关使用限制

## 内容要求

- 包含文字感谢非论文作者的贡献
- 致谢应该是简短的，不应包括感谢匿名审稿和编辑等内容
- 资助或提供的相关资金如已在数据集简介表中提到，不必再提

- 每个作者对成果的贡献应在本部分以单独行的方式进行简要描述
- 包含所有的论文作者和数据作者
- 格式：

作者姓名（出生年-），性别，×省（市）×市人，学历，职称，研究方向为……。

主要承担工作：……。



China Scientific Data Vol. 1, No. 1, 2016

MODIS daily cloud-free snow cover products over Tibetan Plateau (2002 – 2015)

Qiu Yubao<sup>1</sup>, Guo Huadong<sup>1</sup>, Chu Duo<sup>2</sup>, Zhang Huan<sup>1</sup>, Shi Jiancheng<sup>1</sup>, Shi Lijuan<sup>1</sup>, Zheng Zhaojun<sup>3</sup>, Laba Zhuoma<sup>2</sup>

**ABSTRACT** Snow cover over Tibetan Plateau plays an important role in regional water and energy circulation. Snow ablation also affects downstream rivers. Snow parameters and their long-term changes are sensitive factors affecting and responding to regional climate, influencing ecology and disasters. Moderate-resolution imaging spectrometer (MODIS) is widely used for remotely sensing snow due to its high spatio-temporal resolution. However, snow over Tibetan Plateau is distributed patchily and changes rapidly with unexpected atmospheric convection and precipitation. Also, because optical remote sensing is influenced severely by clouds, daily snow cover monitoring is a challenge requiring to remove cloud cover instances. Engaged in Tibetan Plateau's terrain

**Dataset Profile**

|                      |                                                                                                                                                                                                                                                                                                                                                                                                                       |             |        |
|----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|--------|
| Chinese title        | 青藏高原 MODIS 逐日无云积雪面积数据集 (2002 – 2015 年)                                                                                                                                                                                                                                                                                                                                                                                |             |        |
| English title        | MODIS daily cloud-free snow cover products over Tibetan Plateau (2002 – 2015)                                                                                                                                                                                                                                                                                                                                         |             |        |
| Corresponding author | Qiu Yubao (qiuyb@radi.ac.cn)                                                                                                                                                                                                                                                                                                                                                                                          |             |        |
| Data author(s)       | Qiu Yubao, Guo Huadong, Chu Duo, Zhang Huan, Shi Jiancheng, Shi Lijuan, Zheng Zhaojun, Laba Zhuoma                                                                                                                                                                                                                                                                                                                    |             |        |
| Time range           | July 2002 – April 2015                                                                                                                                                                                                                                                                                                                                                                                                |             |        |
| Geographical scope   | The study spans an area of 25°–45°N and 67°–107°E, including the entire Tibet Autonomous Region and Qinghai Province, parts of Sichuan, Yunnan, Xinjiang and Gansu provinces, as well as parts of foreign territories in southern and western Tibetan Plateau                                                                                                                                                         |             |        |
| Spatial resolution   | 500-meter                                                                                                                                                                                                                                                                                                                                                                                                             | Data volume | 6.9 GB |
| Data format          | Geotiff                                                                                                                                                                                                                                                                                                                                                                                                               |             |        |
| Data service system  | < <a href="http://www.sciencedb.cn/dataset/handle/55">http://www.sciencedb.cn/dataset/handle/55</a> >                                                                                                                                                                                                                                                                                                                 |             |        |
| Source(s) of funding | Special Fund for Meteorological Scientific Research in the Public Interest "Constructing a Remote Sensing Product Dataset for Snow Pack over Tibetan Plateau" (No. GHY201206040); State Key Program of National Natural Science Foundation of China (ABCC Grant No. 41120114001); the National Natural Foundation of China (No. 41371331); and "One-Three-Five" Planning Projects of the Chinese Academy of Sciences. |             |        |

The dataset consists of two parts: the "MODIS daily

## Authors and contributions

Qiu Yubao, PhD, Associate Professor; research area: applied research of environmental remote sensing. Contribution: algorithms and process modeling.

Guo Huadong, Academician of CAS; research area: applied research of environmental remote sensing. Contribution: geo-processing guidance.

Chu Duo, PhD, Professor; research area: applied research of environmental remote sensing. Contribution: snow process modeling and verification.

Zhang Huan, MSc; research area: applied research of remote sensing. Contribution: data pre-processing and data production.

Shi Jiancheng, PhD, Professor; research area: applied research of remote sensing. Contribution: data modeling.

Shi Lijuan, PhD; research area: applied research of remote sensing. Contribution: data pre-processing and data production.

Zheng Zhaojun, PhD, Associate Professor; research area: applied research of snow remote sensing. Contribution: data pre-processing and input.

Laba Zhuoma, MSc; research area: applied research of remote sensing. Contribution: data collection and file editing.

# 数据论文模板

## 标题/作者/摘要/关键词 (中英文)

## 正文

- 数据集简表
- 引言/概述
- 数据采集和处理方法
- 数据样本描述
- 数据质量控制和评估
- 数据价值 (可选)
- 数据使用方法和建议 (可选)
- 致谢 (可选)
- 作者分工职责

## 参考文献

## 引用数据

Dataset Profile

|                      |                                                                                                                                                                                                                                                                                                                                                                                                                        |             |        |
|----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|--------|
| Chinese title        | 青藏高原 MODIS 逐日无云积雪面积数据集 (2002 – 2015 年)                                                                                                                                                                                                                                                                                                                                                                                 |             |        |
| English title        | MODIS daily cloud-free snow cover products over Tibetan Plateau (2002 – 2015)                                                                                                                                                                                                                                                                                                                                          |             |        |
| Corresponding author | Qiu Yubao (qiuyb@radi.ac.cn)                                                                                                                                                                                                                                                                                                                                                                                           |             |        |
| Data author(s)       | Qiu Yubao, Guo Huadong, Chu Duo, Zhang Huan, Shi Jiancheng, Shi Lijuan, Zheng Zhaojun, Laba Zhuoma                                                                                                                                                                                                                                                                                                                     |             |        |
| Time range           | July 2002 – April 2015                                                                                                                                                                                                                                                                                                                                                                                                 |             |        |
| Geographical scope   | The study spans an area of 25° – 45°N and 67° – 107°E, including the entire Tibet Autonomous Region and Qinghai Province, parts of Sichuan, Yunnan, Xinjiang and Gansu provinces, as well as parts of foreign territories in southern and western Tibetan Plateau                                                                                                                                                      |             |        |
| Spatial resolution   | 500-meter                                                                                                                                                                                                                                                                                                                                                                                                              | Data volume | 6.9 GB |
| Data format          | Geotiff                                                                                                                                                                                                                                                                                                                                                                                                                |             |        |
| Data service system  | < <a href="http://www.sciencedb.cn/dataSet/handle/55">http://www.sciencedb.cn/dataSet/handle/55</a> >                                                                                                                                                                                                                                                                                                                  |             |        |
| Source(s) of funding | Special Fund for Meteorological Scientific Research in the Public Interest “Constructing a Remote Sensing Product Dataset for Snow Pack over Tibetan Plateau” (No. GYHY201206040); State Key Program of National Natural Science Foundation of China (ABCC Grant No. 41120114001); the National Natural Foundation of China (No. 41371351); and “One-Three-Five” Planning Projects of the Chinese Academy of Sciences. |             |        |
| Dataset composition  | The dataset consists of two parts: the “MODIS daily cloud-free snow cover products over Tibetan Plateau from 2002 to 2015”, and the vector data of the research area. They are, respectively:<br>(1) MODIS_Dysno_Cloud-free_2002-2015.zip (6.9 GB), i.e., the daily cloud-free snow data;<br>(2) Tibet_Range.zip (24 KB), i.e., the auxiliary vector data storing the boundary of the study area over Tibetan Plateau. |             |        |

DOI: 10.11922/csdata.170.2016.0003

SUBJECT CATEGORY: Earth sciences

RECEIVED: March 3, 2016

PUBLISHED: June 1, 2016

### MODIS daily cloud-free snow cover products over Tibetan Plateau (2002 – 2015)

Qiu Yubao<sup>1\*</sup>, Guo Huadong<sup>1</sup>, Chu Duo<sup>2</sup>, Zhang Huan<sup>3</sup>, Shi Jiancheng<sup>3</sup>, Shi Lijuan<sup>3</sup>, Zheng Zhaojun<sup>3</sup>, Laba Zhuoma<sup>3</sup>

**ABSTRACT** Snow cover over Tibetan Plateau plays an important role in regional water and energy circulation. Snow ablation also affects downstream rivers. Snow parameters and their long-term changes are sensitive factors affecting and responding to regional climate, influencing ecology and disasters. Moderate-resolution imaging spectrometer (MODIS) is widely used for remotely sensing snow due to its high spatio-temporal resolution. However, snow over Tibetan Plateau is distributed patchily and changes rapidly with unexpected atmospheric convection and precipitation. Also, because optical remote sensing is influenced severely by clouds, daily snow cover monitoring is a challenge requiring to remove cloud cover instances. Engaged in Tibetan Plateau's terrain complexity and snow spatio-temporal characteristics, this paper presents a compound method by combining different cloud removal algorithms, giving a MODIS daily cloud-free snow cover algorithm for Tibetan Plateau, as well as MODIS daily cloud-free snow cover products (2002 – 2015). The accuracy of the snow cover products is then verified against experimental data observed from 145 ground stations during two winter periods from October 1, 2009 to April 30, 2011. Results show that, when snow depth exceeds 3 cm, the general classification accuracy is 96.6% and the snow classification accuracy is 89.0%. Accuracy was well controlled in each step, which provided a good algorithm for removing clouds from the MODIS snow cover imagery. A multi-language operational process was developed and the daily, cloud-free climatological snow cover products over Tibetan Plateau were released as a free utility online.

**KEYWORDS** daily snow cover products, cloud removal algorithm, cloud free, Tibetan Plateau, MODIS

#### Dataset Profile

|                      |                                                                                                                                                                                                                                                                                                                                                                                                                        |             |        |
|----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|--------|
| Chinese title        | 青藏高原 MODIS 逐日无云积雪面积数据集 (2002 – 2015 年)                                                                                                                                                                                                                                                                                                                                                                                 |             |        |
| English title        | MODIS daily cloud-free snow cover products over Tibetan Plateau (2002 – 2015)                                                                                                                                                                                                                                                                                                                                          |             |        |
| Corresponding author | Qiu Yubao (qiyub@radi.ac.cn)                                                                                                                                                                                                                                                                                                                                                                                           |             |        |
| Data author(s)       | Qiu Yubao, Guo Huadong, Chu Duo, Zhang Huan, Shi Jiancheng, Shi Lijuan, Zheng Zhaojun, Laba Zhuoma                                                                                                                                                                                                                                                                                                                     |             |        |
| Time range           | July 2002 – April 2015                                                                                                                                                                                                                                                                                                                                                                                                 |             |        |
| Geographical scope   | The study spans an area of 25° – 45°N and 67° – 107°E, including the entire Tibet Autonomous Region and Qinghai Province, parts of Sichuan, Yunnan, Xinjiang and Gansu provinces, as well as parts of foreign territories in southern and western Tibetan Plateau                                                                                                                                                      |             |        |
| Spatial resolution   | 500-meter                                                                                                                                                                                                                                                                                                                                                                                                              | Data volume | 6.9 GB |
| Data format          | Geotiff                                                                                                                                                                                                                                                                                                                                                                                                                |             |        |
| Data service system  | <http://www.sciencedb.cn/dataSet/handle/55>                                                                                                                                                                                                                                                                                                                                                                            |             |        |
| Source(s) of funding | Special Fund for Meteorological Scientific Research in the Public Interest "Constructing a Remote Sensing Product Dataset for Snow Pack over Tibetan Plateau" (No. GYHY201206040); State Key Program of National Natural Science Foundation of China (ABCC Grant No. 41320114001); the National Natural Foundation of China (No. 41371351); and "One-Three-Five" Planning Projects of the Chinese Academy of Sciences. |             |        |
| Dataset composition  | The dataset consists of two parts: the "MODIS daily cloud-free snow cover products over Tibetan Plateau from 2002 to 2015", and the vector data of the research area. They are, respectively:<br>(1) MODIS_Dynso_Cloud-free_2002-2015.zip (6.9 GB), i.e., the daily cloud-free snow data;<br>(2) Tibet_Range.zip (24 KB), i.e., the auxiliary vector data storing the boundary of the study area over Tibetan Plateau. |             |        |

#### 1. Introduction

As an important part of the cryosphere, snow is one of the most active natural elements on the Earth's surface, profoundly impacting regional and even global climate change, energy balance, and water cycle<sup>1</sup>. For Tibetan Plateau – the highest plateau in the world which is known as the "roof of the world" and "the third pole", snow is an important feature of the land surface. With strong reflective and near thermal insulative snow changes the radiation balance of the atmosphere by absorbing a s in atmospheric circulation, thus s of snowmelt also have profound land conditions<sup>1</sup>. Snow's impact on South and East Asia has been a focus of domestic and international climatologists for more than a century<sup>4-6</sup>.

1. Key Laboratory of Digital Earth Science, Institute of Ren 100094, P. R. China;  
2. Institute of Tibetan Plateau Atmospheric and Environment 850000, P. R. China;  
3. National Satellite Meteorological Center, Beijing 100081, P \*Email: qiyub@radi.ac.cn

## Abstract/Data profile

Online Reading

Comments

Metrics

Related Articles

### MODIS daily cloud-free snow cover products over Tibetan Plateau (2002 – 2015)

暂无讨论

**Advice&Answers :**  
输入意见

讨论版PDF 讨论版评论

**Editor:** 3.3部分的图存在以下问题：（1）有两幅标识为第357天的图，而没有标识为第359天的图。（2）有两幅表示为第364天的图，而没有标识为第365天的图。请予以修改。

**Editor:** 引用数据和引用论文的信息都会由系统自动生成，不需要在论文中写上这两个信息，所以请将原文中最末尾处的引用数据和引用论文信息删除后将论文重传一下，谢谢！

**Peer Reviewer:** 该文利用去云算法和技术处理，获取青藏高原原无云积雪面积长时间序列数据，原理、方法、数据、文档等描述清晰，是一篇较好的数据论文。总体同意发表，同时提出一个小意见如下：英文论文的标题翻译不合适，是以算法先进性为中心的译法，建议改为为本数据论文所遵循的以数据为中心的译法。

**Peer Reviewer:** 该数据集质量很高，数据论文非常规范。还有部分技术和科学问题没有阐述清晰，可建议作者增加说明：1）为什么采用连续三天数据组合成，三天是如何确定的？通常MODIS自身合成很少使用三天来融合判定时间。2）作者单位，是否应该使用中科院数字地球科学重点实验室的全称？3）该数据集基于MODIS冰雪产品进行后处理形成，在青藏高原地区的一个核心问题就是山地落差造成的正射校正。分类判断中是否已经考虑了坡向、河谷等可能改变自然地理特点的影响因素？

**Executive Editor:** 根据同评专家的意见，建议下载论文最新版本，在此基础上做如下修改：1. 英文论文的标题翻译不合适，是以算法先进性为中心的译法，建议改为为本数据论文所遵循的以数据为中心的译法。2. 建议增加说明：为什么采用连续三天数据进行合成，三天

## Comments

### Authors and contributions

**Qiu Yubao**, PhD, Associate Professor; research area: applied research of environmental remote sensing. Contribution: algorithms and process modeling.

**Guo Huadong**, Academician of CAS; research area: applied research of environmental remote sensing. Contribution: geo-processing guidance.

**Chu Duo**, PhD, Professor; research area: applied research of environmental remote sensing. Contribution: snow process modeling and verification.

**Zhang Huan**, MSc; research area: applied research of remote sensing. Contribution: data pre-processing and data production.

**Shi Jiancheng**, PhD, Professor; research area: applied research of remote sensing. Contribution: data modeling.

**Shi Lijuan**, PhD; research area: applied research of remote sensing. Contribution: data pre-processing and data production.

**Zheng Zhaojun**, PhD, Associate Professor; research area: applied research of snow remote sensing. Contribution: data pre-processing and input.

**Laba Zhuoma**, MSc; research area: applied research of remote sensing. Contribution: data collection and file editing.

## Citation

**How to cite this article:** Qiu Y, Guo H, Chu D et al. MODIS daily cloud-free snow cover products over Tibetan Plateau (2002 – 2015). *China Scientific Data* 1 (2016), DOI: 10.11922/csdata.170.2016.0003

青藏高原MODIS逐日无云积雪面积数据集 (2002 – 2015年)

祁玉宝<sup>1\*</sup>, 郭华东<sup>1</sup>, 触多<sup>2</sup>, 张欢<sup>3</sup>, 石利娟<sup>3</sup>, 郑超军<sup>3</sup>, 拉巴卓玛<sup>3</sup>

浏览下载总计

浏览量

3610

浏览次数统计

下载次数统计

3610PageViews

126Downloads



## 开放获取

数据论文与数据集开放获取



## 方便重用

用户友好，全生命周期数据管护



## 高效迅速

全在线流水作业，最快可半月内开放同评、三个月内正式出版



## 质量保障

严格的学术出版评审全流程（同行评议/大众评议/编委会投票/主编负责制）



“数据出版是科研共同体的防腐剂，通过出版数据，我们将清楚地看到公共经费支持项目所捕获“大数据”的真实情况；数据出版是科研共同体的试金石，通过数据论文与数据集的开放获取，科学数据的价值外延将逐步扩展；数据出版是科学前沿的探测器，通过推广数据出版，我们也将更进一步感知国内外不同领域科学数据共享的创新进展。‘问渠哪得清如许，为有源头活水来。’为唤醒沉睡的数据，《中国科学数据》希望为科研共同体输送源头活水，承载数据，问路未来。”

郭华东. 问渠哪得清如许，为有源头活水来——《中国科学数据》发刊词[J/OL]. 中国科学数据, 2016, 1(1). <http://www.csdata.org/paperView?id=55>. DOI: 10.11922/csdata.0.2016.0014.

# Thank you !

# csdata@cnic.cn

孔丽华

-----

010-58812761  
klh@cnic.cn

Photograph by Jeffrey



## Submission instructions

Make online submissions via [www.csdata.org](http://www.csdata.org), and deposit the datasets in an accredited database (e.g., [www.sciencedb.cn](http://www.sciencedb.cn)).

Manuscript submission: [www.csdata.org](http://www.csdata.org) Data repository: [www.sciencedb.cn](http://www.sciencedb.cn)  
ISSN 2096-2223 CN 11-6035/N

2004年创刊,是一本面向医药行业经营管理人员的综合性医药卫生专业刊物,中国科技核心期刊,以解读医药政策法规、报道产业发展走势、展示医药科研成果、传播经营理念为主,内容强调理论与实践相结合。



2008年创刊,围绕我国卫  
和卫生改革发展的最新成果与  
供重要学术支撑。中文核心期刊、中国科技核心期刊、中国科学引文数据库(CSCD)来源  
期刊、RCCSE中国核心学术期刊(A)。



中国医学科学院医学史研究所 中国医学科学院医学史研究所

ISSN 1673-8018

**医学研究杂志**

中国医学科学院医学史研究所

1972年创刊，为国家级医  
统计源期刊，中国科技核心期刊  
医学发展趋势，适时追踪医学热  
水平较高。