中国科技资源导刊 ISSN 1674-1544 CHINA SCIENCE & TECHNOLOGY RESOURCES REVIEW
2010 年 5 月 第 42 卷第 3 期 30-35 ISSN 1674-1544 Vol.42 No.3 30-35，May 2010

科学数据文档的研究

潘小多 李 新 南卓铜 吴立宗 王亮绪 李红星
（中国科学院寒区旱区环境与工程研究所，甘肃兰州 730000）

摘 要：数据文档是科学数据的必要补充，完善的数据文档对于未知的数据用户是非常有价值的，西部环境与生态科学数据中心为方便广大数据用户使用数据，尽量详细介绍与数据相关的背景知识，数据属性、数据的使用方法和数据的应用状况，同时也为了尊重和保护数据生产者的知识产权，在文档里详细给出数据的引用方法，并提供与数据生产相关的文献信息，针对特色数据集和“中国西部环境与生态科学研究计划”汇交数据撰写文档，在网络上实现无限制浏览与下载，并实现 Wiki 系统的协同写作。

关键词：数据文档；中国西部环境与生态科学数据中心；数据共享；Wiki；科学数据
中图分类号：TP3 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2010.03.005

Research on Data Description Document

Pan Xiaoduo, Li Xin, Nan Zhuotong, Wu Lizong, Wang Liangxu, Li Hongxing
(Cold and Arid Regions Environmental and Engineering Research Institute, CAS, Lanzhou, 730000)

Abstract: Data description document is a necessary complement to scientific data, high quality description of scientific data to unknown users is valuable. To make users convenient, the data description documents were written as detail as possible in the Environment and Ecological Science Data Center for Western China, such as, background information, data property, data usage, data applications and data limitations, but also to respect and protect intellectual property rights of data producers, the reference method is drawn up and literatures on data producing are provided as more as possible. Our data center provides perfect data description documents for special regional data sets and data sets from “The Western Environment and Ecology Research Programmer”, these documents can be viewed and downloaded from our website (http://westdc.westgis.ac.cn) without any limitation, and can be realized synergies writing via Wiki system.

Keywords: data description document, environment and ecological science data center for western China, data sharing, wiki system, scientific data

1 引 言

科学数据是人类社会从事科技活动所产生的原始观测数据、探测数据、试验数据、实验数据、调查数据、考察数据、遥感数据、统计数据、研究数据以及相关的元数据和按照某种需求系统加工的数据。具有科学价值和使用价值的科学数据是信息时代最基本、最活跃、影响面最宽的科技资源和一种战略性资源，它对于科技创新具有显著的基础支

作者简介：潘小多（1978– ），女，助理研究员，主要研究方向：陆面过程模拟及陆面数据同化研究中的大气驱动数据集制备及降尺度方法研究。

基金项目：“中国西部环境与生态科学研究计划”重点项目“中国西部环境与生态科学数据中心”（90502010）；中国科学院西部行动计划（二期）项目“黑河流域遥感—地面观测同步试验与综合模拟平台建设”（KZCX2-XB2-09）。
收稿日期：2010 年 5 月 6 日。

撑作用[1]。科学数据应该共享，科学数据只有在广泛应用过程中，才能实现数据拥有者对其获取最大效用的追求；同时，又可在应用过程中衍生出满足更高层次需求的新数据[2]。借助元数据能够实现网络共享，使得数据资源的用户可以迅速地发现与其需求匹配的数据资源，进而通过网络或其他途径取得这些数据资源并加以利用。元数据是对数据资源的规范化描述。它是按照一定标准(即元数据标准)，从数据资源中抽取出相应的特征属性，组成的一个特征元素集合(即元数据元素)。

但是元数据对于资源描述的特殊性和一般性的矛盾与生俱来，是其本身无法克服的。不仅科学数据的多样性与元数据的单一性存在矛盾，对于来自不同背景的用户，在数据使用中仅从元数据获取的数据基本信息也是不够的。他们希望了解数据产品的采集、处理、制备等详细过程，例如对从遥感数据得到的生物物理参数集，用户希望对从大气校正到参数反演算法的数据处理细节都能给予详尽说明，同时希望得到数据产品的制备背景、产品的质量和缺陷等辅助信息，完善的数据文档对于未知的数据用户是非常有价值的[3]。此外，数据中心往往只能对数据做出技术上的质量控制，科学数据可能需要经过类似科学论文的同行评议，才能更好地被科学家所接受，因此，必须提供完整的数据文档，用户才能对数据做出自己的取舍和判断。同时，在地球表层科学越来越成为实验科学的今天，数据文档也是科学实验可重复性的重要原始证据[4]。对于研究人员来说，主要精力应集中在研究工作本身，而不应将过多精力投入到收集数据，处理数据和如何使用数据上来。因此，西部环境与生态科学数据中心(以下简称“西部数据中心”)是以国际元数据标准 ISO 19115 作为核心进行数据共享，并利用 OGC CS-W 服务连接不同的数据中心，实现元数据的互通互联，将数据文档作为科学数据的必要补充，将数据文档与元数据、数据本身共同组成数据库提供给用户，方便用户使用科学数据。

2 数据文档的研究现状

早期数据文档依附于数据文档文件，所谓数据文档文件就是用于存档数据及其相应的数据说明部分的文件，数据文档文件包括两部分：一是数据的文字说明部分，即数据文档文件的文档部分，用于描述数据的背景、目的、测试方法、时间、地点、负责人等信息；二是观测记录的二维数值表。其中数据的说明部分，是描述数据的文字说明，目的是使没有参加收集或获取该数据的其他研究人员能够通过数据文档，了解数据的来源、测定方法和目的，数据产生的具体时间、地点、研究项目及其负责人等信息，以便能够准确地掌握数据使用的局限性，指导其他或者未来的科研人员正确使用数据。

上世纪 90 年代初，为了野外生态数据的规范化和共享，美国著名的生态数据管理专家 Walt 和 James[5] 共同创立了生态系统数据管理中的数据文档文件，又称站际间文档交换文件。中国生态系统研究网络(CERN)[6-8] 是采纳站际间文档交换文件结构，结合中国的国情积极采纳和推广数据文档格式，建立一个生态站的科学数据库，提供丰富的信息为来站工作的科技人员提供参考，实现 CERN 站际之间的数据通讯和资源共享。

野外台站数据在科研上的重要地位固然不会动摇，但是其存储的数据量在科学研究的数据量中比例却大大下降，随着遥感技术、地理信息技术以及生态水文模型的发展，大量的再生产数据也成为数据的重要来源，动辄以 MB、GB、TB 存储的数据已屡见不鲜，已非 KB 级存储的野外台站数据所能比拟的。除了存储量外，还在以下方面有所区别：

(1)读取方式上：传统的野外台站数据基本上以 .txt 格式存储，读取方式比较单一，未知的科研人员也能轻易读取数据，但是遥感数据和地理信息系统数据等，格式比较多，如有 HDF 格式、IMG 格式、e00 格式、shapefile 格式、coverage 格式、grid 格式等，还有分析资料数据的 grib 格式和 NETCDF 格式等。每一种格式都有最适合的软件来读取，使用单一的 .txt 格式不能满足这种多元化的海量数据存储的要求。

(2)制备方法上：野外台站获取主要依赖于野外仪器，而地球科学中日益突出的遥感数据来源于不同的传感器，误差处理、几何纠正、系统

误差等处理都不同。遥感反演产品更是多样化，采用不同的波段，采用不同的反演公式，获取不同的产品。这些详细的制备过程，在传统的数据文档文件中根本不能体现。

因此“数据文档文件”的结构已经不适合海量数据，我们需要将数据文档和数据分离开来，也不再拘泥于 .txt 格式的文件。

在国际上，尤其是欧美国家的一些数据中心，数据集介绍文档比较完备，如 ISLSCP[9]（The International Satellite Land Surface Climatology Project）计划撰写的说明文档。国内研究机构和学者对数据集介绍文档还不够重视，数据集介绍等同于元数据说明。元数据对于资源描述的特殊性和一般性的矛盾与生俱来，是其本身无法克服的。或许随着标准化的进程，DC 元数据（Dublin Core Metadata）等少数元数据格式将占据主导地位，然而永远不可能统一到仅有少数几种格式 [10]。科学数据不同于一般的空间数据，元数据所提供的数据描述通常是不够完备的，用户希望了解数据背景、数据采集、数据处理等细节 [4]。所以在文档说明部分应尽可能记载与数据来源有关的信息，包括研究项目的来源、名称、负责人、开始日期、结束时间，研究目的、内容和方法，采取的技术手段，数据采集的具体时间、地点，所用的仪器型号，样品处理过程，所用的数据分析方法，变量名和字段名的缩写及其准确含义，数据的应用情况，数据的限制等，以帮助后来者充分理解该数据集的文字说明，同时在数据文档中注明引用方式，以充分尊重和保护数据生产者的知识产权。如果有必要，在这一栏目中还可以增加有关该数据集的摘要，以便对文档文件进行检索。

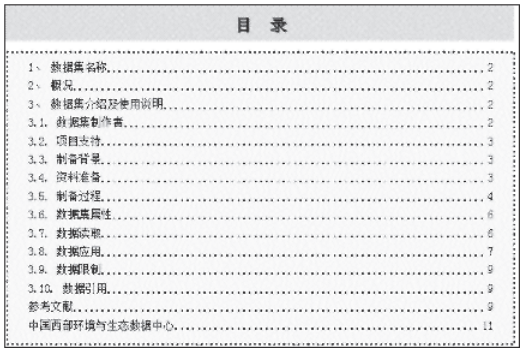
3　西部数据中心的数据文档

西部数据中心是国家自然科学基金委员会“中国西部环境与生态科学研究计划（简称为西部计划）”以建设集成化、体系化、规范化科学数据共享平台为目标而设立的重点研究项目，承担“西部计划”项目数据产出的收集、管理、集成，并面向西部环境与生态科学的各个领域提供科学数据服务；以共建共享的指导思想集成针对西部环境和生态现状与变化的基础背景数据、环境与生态观测数据、模型数据集、数据工具及数据文档；并以完全与开放的数据共享原则为用户提供推送式数据服务。数据文档是科学数据的必要补充。跟随西部数据中心建设的进展，撰写特色数据集文档和西部计划汇交数据集文档。

3.1　文档案例

科学数据文档的目标是方便广大数据用户使用数据，尽量详细介绍与数据相关的背景知识、数据属性、数据的使用方法和数据的应用状况，同时也为了尊重和保护数据生产者的知识产权，在文档里详细给出数据的引用方法，并提供与数据生产相关的文献信息。

西部数据中心的数据文档内容包括：（1）数据制备背景、项目支持等背景信息；（2）数据处理过程，包括资料准备、数据的采集方法和取样方法，仪器的相关说明，数据处理的方法和详细过程；（3）数据属性信息，包括元信息，时空分辨率、投影信息和数据格式等信息；（4）数据使用说明，包括数据的读取、数据使用建议等；（5）数据来源及引用信息，为了突出对数据产权的保护，特别在数据文档中提供明确的数据来源和数据引用的相关信息。其案例如图 1 所示。


目　录

1、数据集名称……2
2、概况……2
3、数据集介绍及使用说明……2
3.1. 数据集作者……2
3.2. 项目支持……3
3.3. 制备背景……3
3.4. 资料准备……3
3.5. 制备过程……4
3.6. 数据集属性……6
3.7. 数据读取……6
3.8. 数据应用……7
3.9. 数据限制……9
3.10. 数据引用……9
参考文献……9
中国西部环境与生态数据中心……11

图 1　西部数据中心的文档案例

西部数据中心的数据引用方式首先尊重数据生产者的意愿。如果数据生产者没有提供相应的引用信息，数据中心将提供以下的统一引用方式：

致谢：数据下载于国家自然科学基金委员会“中国西部环境与生态科学数据中心”（http://westdc.westgis.ac.cn）。

Acknowledgements: This data set was downloaded from “Environmental & Ecological Science Data Center for West China, National Natural Science

Foundation of China”(http://westdc.westgis.ac.cn)。

3.2 主要的数据文档

由于时间和人力投入关系，西部数据中心数据文档的撰写主要集中于特色数据集和西部计划汇交数据文档，主要的数据文档见表 1。

表 1 西部环境与生态科学数据中心的主要数据文档

特色属性	数据文档
特色数据：遥感	被动微波SMMR & SSM/I亮温数据集
	长时间序列中国植被指数数据集——GIMMS AVHRR NDVI
	长时间序列中国植被指数数据集——Pathfinder AVHRR NDVI
	长时间序列中国植被指数数据集——SPOT VEGETATION NDVI
特色数据：冰冻圈	中国冰川信息系统数据库（1：10万）
	中国雪深长时间序列数据集
	中国长序列地表冻融数据集
特色数据：区域科学数据集	HEIFE观测实验数据库
	石羊河流域信息系统专题数据集
特色数据：生态环境	中国地区土地覆盖数据集
	中国1：100万植被数据集
特色数据：沙漠与沙漠化	中国1：10万沙漠分布数据集
西部计划汇交数据	中国西部地区2002年地表气候要素再分析数据集
	中国西部陆面同化系统输出数据集
	中国地震目录
	古气候数据集
	西部荒漠植被生态效应研究数据集
	塔里木河下游生态数据集
	冻土环境对青藏铁路工程建设的影响及工程的环境效应
	三江源科学数据集
	河西走廊内流区生态资料
	青藏高原1∶100万地貌数据集
	中国西部1∶10万土地利用数据
	干湿指数序列数据

3.3 文档共享与编辑平台

为了实现文档编撰者和数据生产者以及文档修改者之间的协同合作，西部数据中心采用 Wiki 系统[11]。Wiki 系统支持面向社群的协作式写作，为协作式写作提供必要帮助。我们可以在 Web 的基础上对 Wiki 文本进行浏览、创建、更改，而且创建、更改、发布的付出远比 HTML 文本要小，同时 Wiki 有使用方便及开放的特点。西部数据中心的 Wiki 系统中，普通用户可以浏览“页面”、“讨论”、“源码”和“历史”等信息。在“页面”中可以浏览所有文档的标题，也可以进入下一级浏览各个数据集的最新文档说明；在“讨论”中可以浏览文档编辑者之间的讨论信息；在“源码”中可以浏览网页代码；在“历史”中可以浏览数据文档不同版本的更新者、更新的内容。更新的时间，点击任一版本还能浏览当前版本的数据文档内容。文档编辑者登录进入以后，除了“页面”、“讨论”、“历史”外，还有“编辑”、“移动”和“监视”。“编辑”便是文档编辑者撰写、更新数据文档的地方。“移动”可以实现“重命名一个页面，并将其修订历史同时移动到新页面”；“监视”可以实现将页面加入到个人制定的“监视列表”，有关此页面的任何修改都将在那里列出。

另外，西部数据中心的 Wiki 系统添加了导航系统，有“首页”、“社区”、“当前事件”、“最近更改”、“随机页面”和“帮助”等链接，还添加了搜索功能，同时提供了工具箱，有“上传文件”和“特殊页面”。“上传文件”的文件类型有 png、gif、jpg、pdf、doc 和 xls 等；“特殊页面”可以定制网页的显示方式。其案例如图 2 所示。

图 2 西部数据中心的 Wiki 系统案例

4　问题与讨论

4.1　数据文档撰写

现阶段，西部数据中心根据数据生产者提供的文章和其他文档信息，撰写数据文档，其中特色数据集大部分数据来自西部数据中心，所以在撰写、编辑过程中得到数据生产者的大力支持，从而完成了内容详尽的数据文档。西部计划汇交数据的数据文档，则存在一定程度的欠缺。但无论是特色数据集还是西部计划汇交数据，如果数据生产者能够按照西部数据中心的数据文档案例（或者在不断的实践摸索中，总结出更好的数据文档案例），在提交数据的同时也能提供数据文档，将会大大提高数据文档的质量，这样也更方便了数据使用者，发挥科学数据的最大效用。

4.2　数据文档发布

西部数据中心的数据下载分为在线下载和离线下载，其中在线下载的方式是：用户登录进入方可下载。而伴随科学数据的数据文档则是直接点击即可下载或者浏览，这样有利于数据使用者在未下载数据之前就已经对数据的属性等有全方位的了解，以免下载了不合适的数据。

4.3　数据文档更新

数据文档的更新伴随着数据更新而至，Wiki 系统在数据文档更新方面发挥重要的作用，提倡数据生产者能够登录文档编辑页面及时对数据的更新内容在文档里体现出来，尤其是涉及准备资料、制备方法、数据属性发生变更时，数据生产者能够亲自对文档进行更新，西部数据中心文档编撰者也能辅助数据生产者做简单的数据文档更新。

5　结　论

数据文档是实现文档文件的完整和准确的重要手段之一，并使其他研究人员能够通过数据文档，了解数据的来源、测定方法和目的，数据产生的具体时间、地点、研究项目及其负责人等信息，以便能够准确地掌握数据使用的局限性，指导其他或者未来的科研人员正确使用数据，而且对于研究工作本身的及时总结、中间结果的取得、研究进度的合理安排和有效管理、科技档案的建立都有很大帮助。西部数据中心结合数据特点，区别于元数据说明，详细撰写数据文档，包括数据集描述、数据制备方法和过程、数据使用说明（如何读取等），并明确说明数据集的贡献者，引用了相关作者的文献，同时要求在发表来自该数据的研究成果时对数据来源及贡献者进行说明或引用。

参考文献

[1] Huang D C, Li X B, Wang J L. Study on the Construction Strategy of National Scientific Data Sharing Program[J]. Chinese Basic Science, 2005(5): 29-35. (in Chinese)
〔黄鼎成 , 李晓波 , 王卷乐 . 浅谈科学数据共享工程建设的战略取向 [J]. 中国基础科学 , 2005 (5): 29-35. 〕

[2] Huang D C. Theoretical Foundation and Mechanism of the Scientific Data-sharing[J]. Chinese Basic Science, 2003 (2): 22-27. (in Chinese)
〔黄鼎成 . 科学数据共享的理论基础与共享机制 [J]. 中国基础科学 , 2003(2): 22-27. 〕

[3] Parsons M A, Duerr R. Designating User Communities for Scientific Data: Challenges and Solutions[J]. Data Science Journal, 2005(4): 31-38.

[4] Li X, Nan Z T, Wu L Z, et al. Environmental and Ecological Science Data Center for West China: Integration and Sharing of Environmental and Ecological Data[J]. Advances in Earth Science, 2008, 23(6): 628-637. (in Chinese)
〔李新 , 南卓铜 , 吴立宗 , 等 . 中国西部环境与生态科学数据中心 : 面向西部环境与生态科学的数据集成与共享 [J]. 地球科学进展 , 2008, 23(6): 628-637. 〕

[5] John B G. Data Management at Biological Field Stations and Coastal Marine Laboratories[R]//Walt, James W B. Intersite Archival and Exchange File Structure. Michigan: Michigan State University,1992.

[6] Liu D G, Cai Y T, Lin Z D. The Data Document Structure Design of Ecological Station Information Management System (CERN / SIMS) [J]. Research Progress on Resources and Ecological Environment, 1995, 6(3): 1-8. (in Chinese)
〔刘德刚 , 蔡玉悌 , 林志磊 . 生态站信息管理系统（CERN/SIMS）数据文档结构的设计 [J]. 资源生态环境网络研究动态 , 1995, 6(3): 1-8. 〕

[7] Tang W L, Xu F A, Li X, et al. The Data Document File Format of Fengqiu Experimental Station of Agricultural Ecology, Chinese Academy of Sciences[J]. Research Progress on Resources and Ecological Environment, 1995, 6(3): 9-12. (in Chinese)
〔唐万龙 , 徐富安 , 李欣 , 等 . 中国科学院封丘农业生态实验站数据文档文件格式 [J]. 资源生态环境网络研究动态 , 1995, 6(3): 9-12. 〕

[8] Tang W L, Liu Y B. Metadata in Ecosystem Data Management[J]. Rural Eco-Environment, 1996, 12(3): 57-59. (in Chinese)
〔唐万龙 , 刘元波 . 生态系统数据管理中的数据文档文件 [J]. 农村生态环境 , 1996, 12(3): 57-59. 〕

[9] Hall F G, Collatz G, Los S, Brown de Colstoun E, Landis D, eds. ISLSCP Initiative II [R]. NASA, 2005.

[10] Liu W, Li D L, Xia C J. Ontology-based Metadata Application for Digital Libraries[J]. Library Journal, 2004, 23(6): 50-54. (in Chinese)
〔刘炜 , 李大玲 , 夏翠娟 . 元数据与知识本体 [J]. 图书馆杂志 , 2004, 23(6): 50-54. 〕

[11] Ozok A A, Zaphiris P. Online Communities and Social Computing[M]//Hoisl B, Aigner W, Miksch S. Social Rewarding in Wiki Systems - Motivating the Community. New York: Springer Berlin / Heidelberg, 2007: 362-371.

（上接第 23 页）

〔黄鼎成 . 科学数据共享的理论基础与共享机制 [J]. 中国基础科学 , 2003(2):22-27. 〕

[10] Raymond Eric S. The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary[M]. O'Reilly Media, 1999.

[11] Reid J, Martin F. The Open Source Movement and Its Potential in Implementing Spatial Data Infrastructures [C]//International Symposium on Spatial Data Infrastructure. Australia: Melbourne, 2001.

[12] Steiniger S, Hunter A J S. Free and Open Source GIS Software for Building a Spatial Data Infrastructure. Working Paper, University of Calgary, 2010.

[13] Jolma A, Ames D, Horning N, et al. Free and Open Source Geospatial Tools for Environmental Modeling and Management[G]//Environmental Modeling, Software and Decision Support. 2008, 3: 163-180.

[14] Steiniger S, Hay G. Free and Open Source Geographic Information Tools for Landscape Ecology[J]. Ecological Informatics, 2009, 4(4): 183-195.

[15] Steiniger S, Bocher E. An Overview on Current Free and Open Source Desktop GIS Developments[J]. International Journal of Geographical Information Science, 2009, 23(10): 1345-1370.

[16] Leng Shuying, Li Xiubing, Cheng Guodong, et al. The Progress of Studies on the Environmental Change and Ecological Issues in Western China[J]. Science Foundation in China, 2005(5):262-267. (in Chinese)
〔冷疏影 , 李秀彬 , 程国栋 , 等 . 中国西部环境和生态科学重大研究计划阶段性进展及深入研究的问题 [J]. 中国科学基金 , 2005(5):262-267. 〕