地球科学研究数据的分类与组织研究*

郭明航[1],田均良[1],李军超[2]
(1. 西北农林科技大学 水土保持研究所,陕西 杨陵 712100;2. 西北农林科技大学 生命学院,陕西 杨陵 712100)

摘 要:科学数据是重要的科技资源,科学数据的共享管理日益成为学术界和政府关注的前沿领域。地球科学门类众多,研究对象复杂且往往时空尺度大,在此过程中产生了大量数据结构形式各异的数据,诸如图型数据、表格数据、文本数据、影像数据等等。在数据库系统环境下如何对这些异构的数据进行存储、发布、显示是科学数据管理必须首先面对的问题。在分析研究地球科学数据特征的基础上,结合黄土高原数据中心的建设实践,以科学数据共享管理为目标,对地球科学研究数据的分类和组织进行了研究。阐述了地球科学研究数据的异构性、密集性、复合性等基本属性和特征,提出了关系类型、空间类型、文件类型等 3 种基本类型的数据集分类和组织方式,并提出了整编数据集的基本原则和方法以及科学数据分级、保护、共享的方式。实践表明:该数据分类与组织技术方案符合地球科学研究数据的特点,并将科学数据管理与计算机网络技术、信息技术有机结合,具有思路与技术的先进性和广泛的应用场合。

关键词:地球科学;科学数据;数据分类;数据管理
中图分类号:TP311.12 **文献标识码**:A **文章编号**:1005-3409(2009)04-0203-04

Classification and Organization of Data on Geoscience Research

GUO Ming-hang[1], TIAN Jun-liang[1], LI Jun-chao[2]
(1. *Institute of Soil and Water Conservation, Northwest A & F University, Yangling, Shaanxi* 712100 *China*; 2. *College of Life sciences, Northwest A & F University, Yangling, Shaanxi* 712100, *China*)

Abstract: Multi-discipline, complexity and big spatial-temporal scale resulted in isomerous, compression and composite character of data on geosciences including figure data, table data, file data, image and video data and so on. Thus, the management and share of data on geosciences is an important issue for manager of data. Combined with data construction practice of the Loess Plateau, based on data character research of geosciences, aimed at management in internet environment, classification and organization of data on geosciences was studied in this paper. The isomerous, compression and composite character of data on geosciences was expounded. Classification and organization of three types of data collection (relation type, spatial type and file type), basic principle and methods to establish data collection were presented. Data classification and share way were discussed based on science data share and harmonious protection principle. Practice application showed that technique scheme of classification and organization was accorded with character of geosciences research, and was of extensive applicability resulted from combination with computer internet technique and information technique.

Key words: geoscience; scientific data; classification of data; data management

地球科学是以地球系统(包括大气圈、水圈、岩石圈、生物圈和日地空间)的过程与变化及其相互作用为研究对象的基础学科。主要包括地理学(含土壤学与遥感)、地质学、地球物理学、地球化学、大气科学、海洋科学和空间物理学以及新的交叉学科(地球系统科学、地球信息科学)等分支学科。显然,地

* **收稿日期**:2008-05-15
基金项目:"973"项目"中国主要水蚀区土壤侵蚀过程与调控研究(2007CB407200)";国家科技基础条件平台建设项目"地球系统科学数据共享网建设与服务(2005DKA32300)"
作者简介:郭明航(1962-),男,硕士,高级工程师,主要从事科学数据管理和科研信息化研究工作。E-mail:mhguo@ms.iswc.ac.cn
通信作者:李军超(1960-)男,副教授,主要从事植物生态方面的教学与研究。E-mail:lijunchao57@sohu.com

球科学具有研究对象多、时空尺度大的特征,作为科学研究过程和结果表达的科学数据自然也表现出相当的复杂性。从地球科学研究数据管理的角度考虑,如何合理进行数据的分类和组织,从而有利于数据的存储、查找和使用就变得相当重要,其迫切性亦随着科研信息化环境的迅速发展与日俱增。论文以数据库建设的实践为基础,对地球科学数据在管理过程中的分类和组织问题进行了探讨。

1 地球科学研究数据的特征

1.1 异构性

异构性是指在计算机环境下,数据存储结构的差异,或者称为存储格式的差异。这种结构的分异是由于所描述对象的不同决定的。例如,描述一个土壤剖面湿度状况的数据通常采用表格形式;描述地貌类型可以是遥感影像、航空照片、矢量图、栅格图等形式,即使在同一种形式内,还会因仪器设备的不同、时空尺度的不同也会产生数据的异构性。现行数据格式主要包括:.zip、.rar、.dbf、.mdb、.xls、.doc等等。数据异构是描述客观事物的需要,但另一方面,对异构数据的一体化管理无疑将变得十分的复杂。面对异构数据类型管理的实际需要,人们采用了元数据的概念和技术,这种技术是通过建立结构化的元数据标准、数据库体系对异构数据进行管理,而不考虑各种数据的具体格式[1-2]。

1.2 密集性

密集性是指从数据-信息-知识的转化过程中数据的使用量较大的特性。生态学研究具有时间序列长、空间尺度大以及数据密集的特点。典型的事例譬如全球变化研究,它需要全球尺度、地球系统的不同圈层(大气圈、土壤圈、水圈、生物圈)以及长期的观测数据才能解答诸如碳循环、气候变暖这样的重大科学问题。

1.3 复合性

复合性是指科研工作的数据来源、数据类型的多样性特性。这一特性表现在:生态学研究常常需要其他学科的数据,例如水文、气象、测绘、遥感等学科;由于任何一个科研项目只能取得一定空间范围和一定时间段的某个特定对象的观测资料,因而,任何一个科研项目都需要其他科研项目数据的支持,并且这种需求是互相的、多向的;从数据的获取方式来看既有调查数据、观测数据、试验(实验)数据,也有文献、标本等实物数据;从数据类型来看,既有数值型数据、文本型描述数据、空间矢量数据、栅格数据,也有影像、图片、图形数据等等。

2 地球科学研究数据集的类型

数据集是基于元数据进行科学数据管理数据的基本单元或对象,分析大量的生态研究数据,主要区别其数据结构的差异性、数据管理系统的差异性和数据描述对象的差异性,可以归纳出关系数据类型、空间数据类型和文件数据类型3种基本的数据集类型。

2.1 关系类型数据集

关系类型数据集就是传统意义上的关系数据库,它是由若干二维表组成的数据体。每一个二维表的字段(亦即表头)是相对固定的,也就是说,在建立数据表时是能够使用数据库结构描述语言(Data Define Language,DDL)进行数据表结构的定义。由于关系类型数据集所具有的固定关系,便可利用数据库操纵语言(Data Manipulation Language,DML)对数据表中的数据实施各种操作,这种操作可以定位到数据表中的任意行和列,即数据表的记录和字段。

显然,与下文将要论述的文件类型数据集比较,关系类型数据集操作数据的粒度要小得多,定位数据的精度要高得多。因此,在进行数据管理的过程中,对于关系类型数据集应该尽量的建立关系数据库管理系统,这样才能提高数据管理的能力。

2.2 空间类型数据集

空间类型数据集是地图数据所构成的数据集。空间数据的基本数据模型包括点、线、面;地图管理的基本单元是图层(Coverage)。通过商用的GIS软件,如ARC/info、Map/info等可对空间类型数据集进行管理,实现强大的空间数据管理功能。例如:空间数据的数字化和图形编辑、空间数据的查询和分析、制图和输出。除了这些空间数据管理的基本功能外,较新版本的GIS软件还提供了地表模型生成、显示、分析模块(TIN)。该模块可根据等高线、高程点、地形线生成不规则的三角网并进一步生成地表模型,并对地表模型三维显示分析[3-4]。

要实现对空间数据的上述管理功能自然需要相应的系统环境。与关系数据类型数据集一样,任何一个空间数据管理的基本单元-图层,都可封装成一个数据文件,纳入文件类数据集管理的范畴。

2.3 文件类型数据集

文件类型数据集是以一个数据文件为数据集的基本单元,在数据库管理系统中以此基本单元对数据进行存储、描述、添加、删除、显示、交换等操作。数据文件本身可以是无结构的,其内容也可以不具有任何相似性。例如,一个径流小区年径流量的观

测记录可以做成一个表格形式的文件类型数据集;一个记录不同施肥处理作物生长状况的照片也可以是一个文件类型数据集。而这两个数据文件所记录的内容没有任何相似之处。该类数据文件可以是数据表格、文本文件、图片、视频、一张图形、一景遥感影像等等。

通过对关系数据类数据集和空间类数据集的分析,如果舍弃其内部数据管理的功能而从数据的发现和交换这个层面上考虑,二者也可纳入文件类型数据集管理的范畴。

3　数据集的标准化处理

3.1　数据集与数据实体的一对多关系

几乎所有科研项目都会产生一个数据集对多个数据集实体的研究结果。对于特定的科研项目,研究者会根据自己的科研目的设计若干的实验、试验、调查等科研过程,而每一个过程都会产生相应的数据实体。由于这些过程是围绕一个科研项目设计的,所以,数据实体间是有联系的,这种联系是为了揭示同一事物的不同方面或者不同阶段的个别现象,如果取消了这种联系,那么,对事物的认识、分析就是片面的,甚至是错误的。所以一个科研项目产生的多个数据实体必须组织在一个数据集中。另一方面,由于每一次科研过程(实验、试验、调查等)具有针对性,所以、观测到的数据项,即数据实体中的属性是不同的,而且这些属性往往不能组织到一个数据实体当中。这就是说,一个科研项目的实施必然产生多个数据实体。

3.2　数据实体与元数据的关系

(1)对一般关系数据库而言,数据表的数量是有限的,而且每一个数据表的属性字段是固定的,因此,对其说明的元数据就变得相对简单。而对文件类型数据而言,每一个数据集所产生的数据实体的内容、数目是随着不同科研项目而变的,数据实体的属性字段也是不固定的,因此,要能正确地使用数据实体,就需要关于数据实体的产生条件、过程等情况的详细说明,对数据实体的属性字段也同样需要准确的解释说明,这些说明信息就是所谓的数据集的元数据。否则,如果数据实体离开了这些特定的背景条件,数据实体是毫无意义的,对数据使用者而言将不知所云。

(2)元数据不仅仅是对数据实体的说明,更重要的是数据实体的产生依赖于元数据[5-6]。从科学研究的过程来看,科研计划、方案决定科学研究的结果;而从数据管理的角度来看,所有的科研计划、方案又都是元数据,科学研究的结果都是数据实体。在这个意义上,元数据对于数据实体具有重要的决定作用。

(3)从使用数据的角度分析,用户总是借助元数据寻找数据实体。这一过程表现为:数据库用户根据自己的需要,在数据库的查询条件中设置合适的查询条件,进而得到查询结果。在此,所谓的查询条件都是元数据的范畴。所以,元数据可以将具有某种逻辑关系的数据实体联结起来,使得数据库的数据组织更为科学。

3.3　数据集的粒度

数据集是具有相似意义数据的集合,在实际应用的过程中,其"相似意义"的界定在不同的场合有着很大的差异,或者说这个"相似意义"的界定可以分不同的层次,其结果就是把数据集分成了在物理存储上不同的大小,这里所说的数据集大小就是数据集的粒度。

数据集的粒度没有确定的数值度量,一个庞大的数据库可以称为一个数据集,一次观测得到的数据表也可称为一个数据集;一幅地图可以称为一个数据集,一个声像片段也可以称为一个数据集等。数据集的粒度是一个理论概念,其用途在于为进行数据整编提供理论指导。以下因素需要在实际确定数据集粒度的时候给予考虑。

3.1.1　数据的逻辑关系　围绕同一主题的数据其逻辑关系密切,反之则疏散,数据间的逻辑关系强调的就是数据间的这种亲疏联系,显然,数据间关系密切者就应组织在一个数据集中,反之则应建立不同的数据集。

3.1.2　数据的存储结构　不同观测对象、观测方法、观测活动等会得到不同的观测数据,这些数据的存贮结构自然会多种多样。但从数据管理的角度来说,不同存储结构的数据一般会采用不同的管理技术方案。所以,数据的存储结构是划分数据集的重要基础。

3.1.3　数据的引用方式　科学数据管理的目的在于应用。所以,数据集的粒度确定要能方便数据的引用,而方便引用的度量可以是数据获取的便捷程度、数据内容的完整性和系统性。

3.4　数据集的完整性及其归并与划分

数据集的归并是针对一个数据集而言的,其目的是将不同的数据实体按照某种联系整编到一个数据集中。数据集的划分是针对不同的数据集而言的,其目的是按照某种差异性将不同的数据实体整编成不同的数据集。数据集的归并与划分是数据整编和实施进一步管理的基础,保持数据集之间和一个数据集

内各数据实体之间在知识层面上逻辑关系上的系统性是数据集归并与划分应该遵循的基本原则。也就是说,不同研究目的、不同研究内容、不同数据内容的数据要分别建立数据集;而一个相对独立的科研过程及其结果,则应完整的编入一个数据集中,即所谓的内容的系统性和完整性。例如,一个科研过程得到 N 个不可合并的数据实体,假设第 N_i 个数据实体是作物产量,第 N_i+1 个数据实体是土壤湿度,第 N_i+2 个数据实体是土壤养分测定结果等等。那么,这些数据实体应收编到一个数据集中,否则,围绕同一科研目的所取得的不同数据集实体若分属不同的数据集,不仅无谓地增加了数据集的个数,而且破坏了数据集的完整性,给数据引用带来困难。

在一个数据集内各个数据实体的划分也存在像数据集的归并与划分同样的问题,对此,除要遵守一个数据集内各数据实体之间在知识层面上逻辑关系上的系统性原则外,还应遵守数据实体中数据属性的一致和结构简化原则,也就是说,一个数据实体应该尽可能是相同意义的数据。为了做到这一要求,必要时可增加数据集实体的数目,这一点类似于关系数据库中的利用 $E-R$ 图进行数据库设计的原理。

4 小结

生态研究的内容宽泛而复杂,作为科学研究过程和结果表达的科学数据自然也表现出相当的复杂性。通过对生态研究数据及管理需要的分析研究,提出了生态研究数据的异构性、密集性和复合性等特性是生态学研究科学数据管理时必需考虑的因素。

数据集是基于元数据进行科学数据管理数据的基本单元或对象。经过对大量的生态研究数据分析,主要区别其数据结构、数据管理系统和数据描述对象的差异性,归纳出了3种基本的数据集类型,即关系类型数据集、空间类型数据集和文件类型数据集。文件类型数据集在生态研究中是大量存在和频繁使用的,研究对该类型数据集的管理进行了重点研究。而对关系数据类型数据集和空间类型数据集,如果从数据的发现和交换这个层面上考虑,二者也可纳入文件类型数据集管理的范畴。而对其数据实体实施更进一步的管理则可使用其他的方法,也依赖其他的软件环境。

数据集粒度的概念以及数据集与数据实体、数据集元数据的关系是进行数据集组织的理论指导,它对于数据管理、数据发现、数据的系统性和完整性都有重要意义,但目前对其研究还很不够。

在寻求数据提供者、数据管理者和数据使用者之间利益的平衡点的基础上,提出了现阶段科学数据共享管理的策略和方案,从而实现不同身份的用户对不同状态的数据严格而灵活的访问控制,从而有效的保护数据所有者的利益,同时又为数据共享提供了有效的途径。这种用户管理策略和数据组织策略在当前对于数据共享认识的大环境下具有现实意义。

参考文献:

[1] ARC/INFO - 专业地理信息系统平台[ZB/OL]. http://www.gisky.com/.

[2] Understanding GIS. The ARC/INFO Method[S]. Redland, California: Environmental System research Institute, Inc.

[3] 孙九林,李爽. 地球科学数据共享与数据网格技术[J]. 中国地质大学学报:地球科学,2002,27(5):539-543.

[4] 孙九林. 异构数据共享与网格计算[J]. 地理信息世界,2005,3(1):1.

[5] 王卷乐,游松财,谢传节. 元数据技术在地学数据共享网络中的应用探讨[J]. 地理信息世界,2005,3(2):36-40.

[6] 李军,周成虎. 地理空间数据元数据标准初探[J]. 地理科学进展,1998,17(4):55-63.

[7] 郭明航,李够霞,从怀军. 生态研究数据库系统的设计和开发[J]. 水土保持通报,2005,25(6):59-62.

[8] 郭明航,李够霞,吴开超,等. 科技文献摘录数据库的建设与应用[J]. 水土保持研究,2006,13(2):186-188.