生态研究数据库系统的设计与开发

郭明航[1,2], 李够霞[1,2], 从怀军[1,2], 吴开超[3], 沈志宏[3], 张世维[3]
(1. 西北农林科技大学, 陕西 杨凌 712100; 2. 中国科学院 水利部
水土保持研究所, 陕西 杨凌 712100; 3. 中国科学院 计算机网络信息中心, 北京 100080)

摘 要: 列表类型数据是生态研究中最为常见的数据形式。在分析列表类型数据特征及其与元数据关系, 数据安全和共享策略等问题基础上, 提出了生态研究列表类数据管理系统设计和开发方案。研究认为数据集的元数据不仅是对数据集实体的说明, 而且一定程度上决定着数据集实体的内容和数量, 以及数据集实体之间的内在联系, 这种联系正是进行列表类型数据管理依据所在。

关键词: 元数据; 列表数据; 数据库; 设计和开发
文献标识码: B 文章编号: 1000-288X(2005)06-0059-04 中图分类号: TP392

Design and Establishment of Database System for Ecological Research

GUO Ming-hang[1,2], LI Gou-xia[1,2], CONG Huai-jun[1,2], WU Kai-chao[2], SHEN Zhi-hong[2], ZHANG Shi-wei[2]
(1. *Northwest University of Agriculture and Forestry, Yangling District* 712100, *Shaanxi Province, China*; 2. *Institute of Soil and Water Conservation, Chinese Academy of Sciences and Ministry of Water Resources, Yangling District* 712100, *Shaanxi Province, China*; 3. *Center of Computer Network Information, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract: Tabular data is a common format in ecological research data. Based on the analysis of characteristics of tabulardata, the relationship between dataset and associated metadata, and safety and share strategy, a tabular database system has been designed with Oracle soft package and B/S web databased techniques. The study shows that metadata is not only documents for datasets, but a determination of the content and data volume of datasets as well as the relationships among datasets. The design scheme has been applied in development and establishment of soil conservation database in the Loess Plateau.

Keywords: **metadata; tabulay data; database; design and development**

科学数据作为现代科学可持续发展的重要资源, 它与知识创新的密切关系在科学界已经成为共识。从 20 世纪开始的百余年来, 随着科学研究的发展, 科学数据的地位、作用以及管理方式也处在不断的变化之中。这种变化体现在: 科学研究对数据时空广度的依赖性更强, 科研过程中数据的使用密度更大。科学数据亦成为引领科技发展方向的重要依据, 在地球系统科学领域, 综合性科学研究计划都是基于长期观测系统科学数据的基础上开展起来的。如中国生态系统研究网络(CERN)、美国长期生态研究(LTER)等。在数据管理方面, 数据库技术、数据处理技术、数据应用和共享的法律政策问题等也在迅速发展和完善。适应这种变化的重要标志就是世界数据中心的组建, 国际科学数据委员会的成立, 数据科学作为一门独立学科的诞生[1]等。毋庸置疑, 数据作为信息时代、知识经济时代最活跃的因素, 正在剧烈的影响着当今世界的每个角落。

生态科学是指研究地球表面生物与环境关系, 生态系统结构功能的科学体系。生态科学研究中的过程表述乃至结果的呈现形式纷繁多样, 其中列表数据形式的数据应该说是最普遍的, 而且是主要的, 科研工作中对该类数据的引用也是最常见的。基于上述各种原因, 本文将以黄土高原水土保持数据库的建库实践为基础, 着重讨论列表数据作为一个整体实施各种操作的数据库管理方法, 为同类数据库的建设提供参考。

1 列表数据及其元数据

1.1 列表数据

列表类型数据就是常说的数据表格, 它是由一组属性(列)和记录(行)组成。对任何一个数据集特定的数据表格而言, 则被称之为数据集实体。表 1 给出了一个列表数据的示意性例子。

收稿日期: 2004-11-02
资助项目: 中国科学院知识创新工程信息化建设专项(INF105-SDB-1-31)
作者简介: 郭明航(1962-), 男(汉族), 陕西省乾县人, 高级工程师, 主要从事科学数据管理工作。电话(029)87012387, E-mail: mhguo@ms.iswc.ac.cn。

表 1 不同土壤的萎蔫湿度测定结果

观测序列号	土壤类型	土壤萎蔫湿度/%
001	轻壤土	4.47
002	中壤土	5.63
003	重壤土	8.80

在本数据库系统中数据集实体是基本的操作单元,即表 1 将作为一个整体进行数据库的各种操作,而不涉及表 1 中的属性字段,这是与一般关系数据库在操作对象方面存在的不同。

1.2 数据集

数据集可从狭义和广义 2 个方面解释,狭义的数据集就是数据集实体;广义的数据集是多个数据集实体及其元数据的集合。本文所使用的是广义数据集的概念。考虑到数据集和数据集实体二者在管理、使用上的不同,所以数据集中的元数据是完全开放共享的,而数据集实体则是有条件共享的。

1.3 元数据

对元数据的定义和使用在不同行业间存在一定的差异,但一般都认可元数据是“关于数据的数据”,亦即,元数据是用于描述数据的内容、质量、产生过程和方法、管理和使用说明的数据。元数据的特点和作用在不同的行业和时期在不断地发展和完善。对于面向列表类型数据的元数据可归结为以下几个方面。

(1) 元数据描述的基本单位是数据集。根据数据集的复杂程度可扩展数据集中所包含的数据集实体以及数据集实体的属性。

(2) 元数据的每一个元素是具有特定语义规则的结构化文本。为了提高元数据标准的适用性、数据的安全性,对每一个元数据元素可定义相应的属性。

(3) 元数据的使用通过建立 Web 数据库实现对数据集及其数据集实体的客户端访问。

(4) 元数据对数据的涵义、产生方法、使用方法等要进行完整准确的表述,以避免数据引用过程中对数据理解的二义性。

(5) 提供适当的数据管理元素,帮助数据的交换和传输,促进数据共享。

2 列表类型数据的特征

利用元数据管理列表类型数据的思路主要来源于生态科学研究中数据的特点,这些特点是科学研究自身的规律所决定的,所以,就数据管理而言,认识这些特点并根据这些特点制定适当的数据管理方法,才能更好地管理,进而更好地使用数据。

2.1 数据集与数据集实体的一对多关系

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集实体。由于这些过程是围绕一个科研项目设计的,所以,数据集实体间是有联系的,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集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,即数据集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集实体当中。这就是说,一个科研项目的实施必然产生多个数据集实体。这种数据集的组织方法有如下优点。(1) 将围绕同一研究项目(或同一研究目标)所取得的数据集从逻辑上联结在一起,提高了数据组织的完整性。(2) 减少数据集的元数据文档冗余,提高数据库建设速度。(3) 有利于不同数据集实体之间的相互参考和综合分析。

2.2 数据集实体与数据集元数据的关系

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。而对列表类型数据而言,每一个数据集所产生的数据集实体的内容、数目是随着不同科研项目而变的,实体数据集的属性字段也是不固定的,因此,要能正确地使用数据集实体,就需要关于数据集实体的产生条件、过程等情况的详细说明,对实体数据集的属性字段也同样需要准确的解释说明,即数据集的元数据。否则,如果实体数据集离开了这些特定的背景条件,实体数据集是毫无意义的。

另一方面,元数据不仅仅是对实体数据集的说明,更重要的是数据集实体的产生依赖于元数据。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、方案都是元数据,科学研究的结果都是数据集实体。在这个意义上,元数据对于数据集实体具有重要的决定作用。

从使用数据的角度分析,用户总是借助元数据寻找数据集实体。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集实体联结起来,使得数据库的数据组织更为科学。

鉴于科学研究过程中元数据与数据集实体间存在的上述关系, 建立基于元数据管理列表类型数据集实体的数据库方法符合该类数据自身的规律性。因而, 这种方法对于列表类型的数据是一种简单有效的管理方法。

2.3　列表类型数据的标准化处理

2.3.1　*数据集的归并与划分*　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, 保持一个数据集内各数据集实体内容之间逻辑关系上的完整性是应该遵循的基本原则。也就是说, 一个相对独立的科研过程及其结果, 应完整地编入一个数据集中, 即所谓内容的完整性。

例如, 一个科研过程得到 N 个不可合并的数据集实体, 假设第 N_i 个数据集实体是作物产量, 第 N_i+1 个数据集实体是土壤湿度, 第 N_i+2 个数据集是土壤养分测定结果等等。那么, 这些数据集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。

而在一个数据集内, 各个数据集实体的划分应遵循数据集实体中数据属性的一致和结构简化原则, 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。

2.3.2　*数据集元数据的编写*　基于元数据管理数据集实体的方法对元数据的要求比一般意义上的元数据(关于数据说明的数据)有更多的要求, 这种要求可归纳为对数据集实体的标识、说明和访问控制等内容, 并通过列表类型的元数据标准体现出来。本系统所采用的生态研究元数据标准包括 7 个部分, 46 个元素: 数据集基本信息 16 个元素; 数据集实体基本信息 6 个元素; 实体属性信息 7 个元素; 研究项目信息 10 个元素; 研究方法信息 2 个元素; 研究场地信息 9 个元素; 人员信息 6 个元素。

标准中的每一个元素是编写元数据文档的依据, 对每一个元数据元素的属性描述, 参考了生态学研究元数据标准(V1.1)[2] 的有关内容。包括: 中文名称、英文名称、标识、定义、类型、值域、可选性、最大出现次数等。

3　列表类型数据管理系统功能的设计

3.1　对系统功能的需求分析

对系统功能的需求分析可以从数据库用户和系统管理等方面考虑[3]。

(1) 数据著录。提供数据集的著录模板及其著录规则说明文件, 使得数据库用户能够在数据库系统的客户端编写自己的数据集元数据文档, 并自动添加到数据库中。(2) 基于元数据的查询、检索功能。(3) 提供必要的数据安全策略。数据的安全管理是数据共享最重要的条件和保障, 在目前的数据管理法规政策、人们对数据共享的观念认识等社会条件下, 数据的共享仍然是一个复杂的事情。根据作者的调研, 可行的数据安全策略应当是数据集元数据的完全共享和数据集实体的有条件共享。所谓有条件共享是指数据集实体对其提供者、数据提供者所在的直接组织单位和普通社会用户采用不同的共享政策。所以, 在数据库系统中, 应该具有用户管理功能, 对不同用户赋予相应的数据库使用权限。

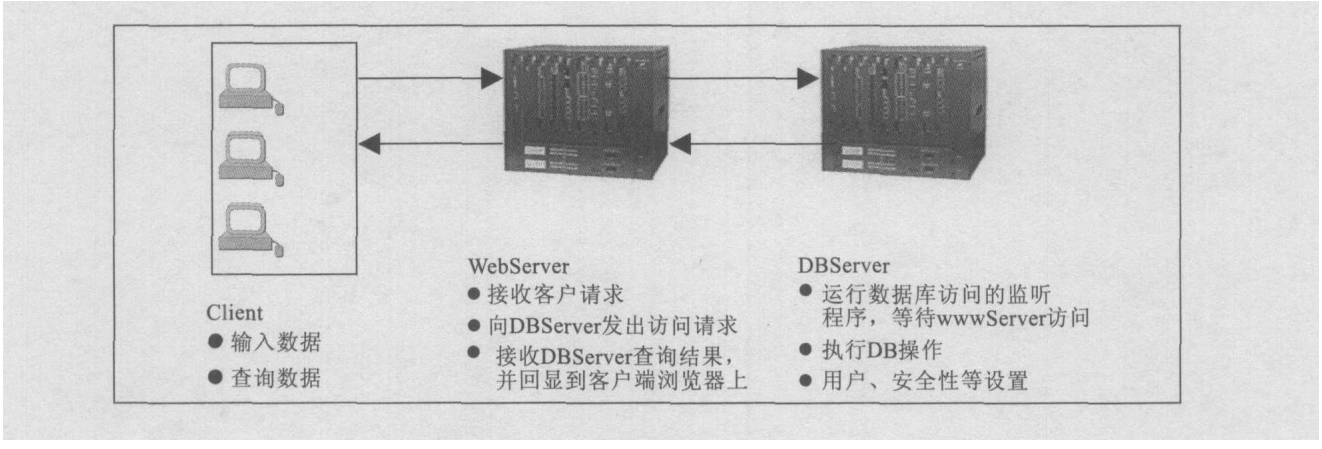

图 1　数据库系统功能逻辑结构图

3.2　数据库系统的实现

3.2.1　*采用关系数据库系统管理元数据*　对于元数据的操作而言, 采用关系数据库的方式管理可以使用关系数据库的许多成熟的技术, 提高系统的整体性能。再则, 关系数据库的应用比较普及, 这样数据库系统的开发就变得相对容易, 即体现了简单、实用的宗旨。

在进行数据库的设计开发中, 前述 7 个部分的元

数据标准, 对应 7 个功能模块, 也对应数据库的 7 个关系数据表, 数据表的实体关系模型如图 2。

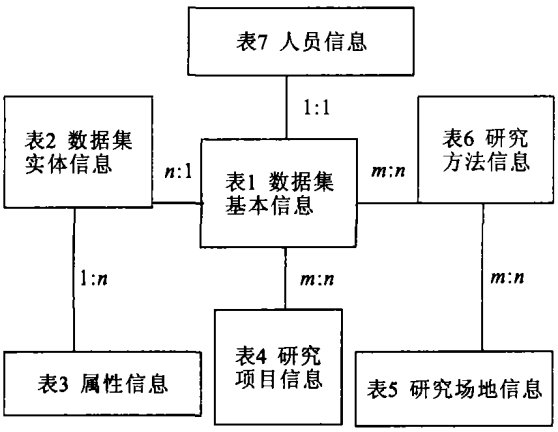

图 2 数据库的实体关系模型

3.2.2 *数据库系统采用 B—S 体系结构* 采用浏览器—服务器构架的体系结构。用户通过 Internet 直接访问的是 Web 服务器, Web 服务器再与数据库服务器进行数据交换, 由数据库服务器对数据进行处理后, 再把数据传递给 Web 服务器, 最终显示在客户端的浏览器上。

3.2.3 *用户及其权限的设计* 本系统设置 4 类用户, 即系统、高级授权、普通授权和匿名用户。

(1) 系统用户。用户管理(增加、修改、删除)等其它修改系统的权限。

(2) 高级授权用户。可由系统用户授权, 通过录入界面向数据库中添加元数据, 同时也可修改、删除元数据, 但这种权限只应限定到该用户自己所添加的元数据范围内。此外, 高级授权用户拥有一般授权用户的权限。

(3) 普通授权用户。可由系统用户或高级授权用户授权。一般授权用户除了拥有匿名用户的权限外, 还可以下载检索数据库中授权范围内的数据集实体文件。

(4) 匿名用户。用户只能够检索元数据, 并且对于检索结果, 只能够浏览非授权字段的信息。

4 结 语

首先, 元数据对于列表类型数据的管理和使用具有非常重要的意义, 在应用实践中, 元数据标准的详略取舍始终存在着矛盾: 对于元数据的编写而言, 元数据标准越简单越好; 而从使用角度而言, 元数据标准越详细越好。掌握什么样的度才能合理的兼顾以上两种需要, 对于不同的专业领域需要根据专业数据的特征区别对待。

其次, 数据管理技术也多种多样。例如网格技术、XML 技术等等。本文侧重的是针对小型数据库, 使用简单、成熟和实用的技术, 面向基层科研单位的数据管理。对于这种技术路线的利弊也希望与同行们交流。了解本数据库系统的详细情况, 请登陆 http://www.loess.csdb.cn。

[参 考 文 献]

[1] 孙鸿烈, 刘闯. 国际科学技术数据前沿领域发展研究[J]. 中国基础科学, 2003(1): 13—18.

[2] 中国科学院科学数据库生态研究元数据标准[S]. 2003.

[3] 李永武. 中国生态系统研究网络元数据标准和操作工具[M]. 中国地理信息元数据标准研究. 北京: 科学出版社, 1999.