

基于非规范化和数据字典的地学元数据管理

周 敏¹, 汪新庆²

(1. 中国地质调查局 西安地质调查中心 测绘遥感处, 陕西 西安 710054;
2. 中国地质大学 地质过程与矿产资源国家重点实验室, 湖北 武汉 430074)

摘 要:针对目前比较主流的地学元数据,对其管理存在的查询、编辑繁琐,存储管理混乱,后续应用空缺等问题进行改进。拟通过非规范化和数据字典技术管理地学元数据,非规范化技术主要解除元数据 XML 文本形式的多层逻辑结构,将其转为符合关系数据库的关系模型,并通过该技术无损还原其元数据的逻辑结构。数据字典技术主要用于优化地学元数据的结构定义,方便在关系数据库中的操作。通过改进,使得地学元数据在查询编辑效率方面有了明显的提高。

关键词:地学元数据;非规范化;数据字典;元数据定制

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2015)03-0175-05

doi:10.3969/j.issn.1673-629X.2015.03.040

Management of Geoscience Metadata Based on Denormalization and Data Dictionary

ZHOU Min¹, WANG Xin-qing²

(1. Office of Mapping and Remote Sensing Technology, Xi'an Geological Survey Center,
China Geological Survey, Xi'an 710054, China;

2. State Key Laboratory of Geological Processes and Mineral Resources, China University of Geology,
Wuhan 430074, China)

Abstract: For the mainstream geoscience metadata, propose some improvements about some problems in tedious query and editing, chaos storage management, subsequent application vacancies and so on. Using the denormalization and data dictionary technology to manage the geo-metadata, the denormalization technology is mainly released the form of multi-layered logical structure of metadata XML text, converting it into line with the relational database relation model, and through the technique to restore the logical structure of its metadata non-destructively. Data dictionary technology is mainly used to optimize geoscience metadata structure definition to make operation more convenient in database. Through the improvements, the geological metadata has the obvious increase in efficiency of the query and editing.

Key words: geological metadata; denormalization; data dictionary; metadata customization

1 地学元数据概念

元数据(Metadata)作为“关于数据的数据”,是数字信息组织和处理的基本工具。而地质学中所涉及的元数据,主要指导和推动地质调查工作的空间信息编目、管理、发布以及社会服务等,它不仅要描述传统纸质地图上必要的信息,还要描述与计算机、GIS 相关的信息,特别是“数字地球”理念的推广,元数据变得尤其重要^[1-2]。

由中国地调局发布的《地质调查元数据内容与结构标准》^[3]由 7 个元数据子集构成,如表 1 所示。

每个子集由若干个实体(UML 类)和元素(UML 类属性)构成。可重复使用实体(负责单位信息实体、引用信息实体)由其他子集调用,不单独使用。图 1 是根据表 1 描述的内容而展示的地质信息元数据的概念结构图。每个元数据包包含一个或多个实体以及元数据元素。

收稿日期:2013-12-16

修回日期:2014-03-21

网络出版时间:2014-10-23

基金项目:国土资源部中国地质调查科技计划项目(1212011085466)

作者简介:周 敏(1987-),女,四川巴中人,硕士,研究方向为地学数据库模型管理;汪新庆,副教授,研究方向为数学地质、地质矿产信息系统、地矿数据库系统、数据库模型。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20141023.1047.002.html>

表 1 地质信息元数据包和元数据实体对比表

序号	包/子集名	实体	定义
1	元数据信息	MD_元数据	元数据的全部信息
2	标识信息	MD_标识	地质数据集的基本信息
3	数据质量信息	DQ_数据质量	数据集数据质量总体评价信息
4	空间参照系信息	RS_参照系	空间参照系信息
5	内容信息	MD_内容描述	数据集的内容信息
6	分发信息	MD_分发	数据集分发者和获取数据的相关信息方法
7	引用和负责单位联系信息	CI_引用 CI 负责单位	提供引用资料信息和负责单位信息

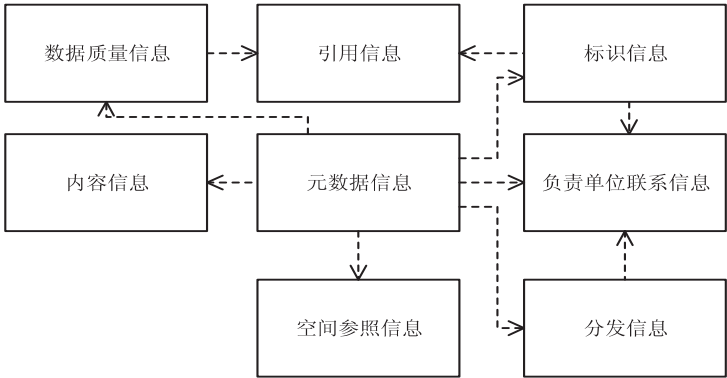


图 1 地质信息元数据概念结构图

2 问题分析及方案设计

2.1 地质元数据的存储管理问题

元数据也是一种数据,其形式与其他数据没有特别之处,它可以以数据存在的任何一种形式存在。一般存储形式有文本文件、超文本文件、元数据文件、关系数据库表等^[4]。

从元数据的形式上看,虽然曾经出现过多种标准,但可扩展标记语言(eXtensible Markup Language,XML)的流行,使其成为最主要的表示元数据的方式,其元数据存储方面主要利用RDBMS扩展的XML存储机制在数据管理能力包括存储、检索、修改等方面比较成熟,但是在数据库异构性的信息共享和数据交换性比较差;直接使用专门的XML数据库管理系统(Native XML Database),在具有非常复杂网络结构的数据及半结构面向文本的数据处理能力很好,但缺少关系数据库完善的关系理论,事物处理,数据一致,多用户访问等机制还没有规范,因此在存储海量数据、恢复灾难、整合事务等方面还不成熟;当然也有使用文件来组织管理元数据,这种方案一般只用于原型演示并不实用。

通过对比,在存储方面元数据还不成熟,它需要一种能满足海量数据存储,兼容异构数据,要求移植性好、维护性能强等要求的方案。而地质元数据本身是关于图件信息的进一步描述和补充,和图件是息息相关的,但是由于图件众多,对于其元数据文件的管理可能会出现混乱而又复杂的局面,没有一定的规范,从而

读取元数据文件也将会显得复杂而且出错的频率会大大增加。

2.2 地质元数据在数据库中的模型设计

基于上述元数据存储管理几种方式分析对比,关系数据库以它完善的数据库关系理论以及强大的数据处理能力管理元数据不失为一种折中的方式。根据元数据自身的特点合理设计逻辑结构,将有效优化数据共享,数据查询,甚至是数据为软件程序服务等性能。

数据模型(Data Model)是数据特征的抽象,是数据库管理的教学形式框架,数据库系统中用以提供信息表示和操作手段的形式架构。数据模型包括数据库数据的结构部分、数据操作部分和约束条件,其中数据操作根据用户自身对数据的需求不同而有所差异。本次数据结构设计,根据其元数据集或实体的UML图如图2所示。

数据集或实体在数据库中表现为数据表,如元数据信息表、数据质量信息表、引用信息表、标识信息表、内容信息表、负责单位联系信息表、空间参照信息表、分发信息表这8个数据表。而数据集或实体中的元素在数据库中对对应表中的字段,这样依次类推,其在数据库中的物理模型图如图3所示。

2.3 非规范化技术

在关系数据库中,数据库是否规范化在数据库设计的环节中占非常重要的地位,它与数据库性能密切相关。但是规范化程度越高,随即复杂查询时连接库表的操作也会增加,降低了查询速度,因此可以适当地

应用非规范化技术来提高数据库执行效率^[5]。非规范化(denormalization)——常用作规范化的反向,通过逆向规范化等方法降低粒度^[6-7]。一般有以下几个方面会使用到非规范化技术:

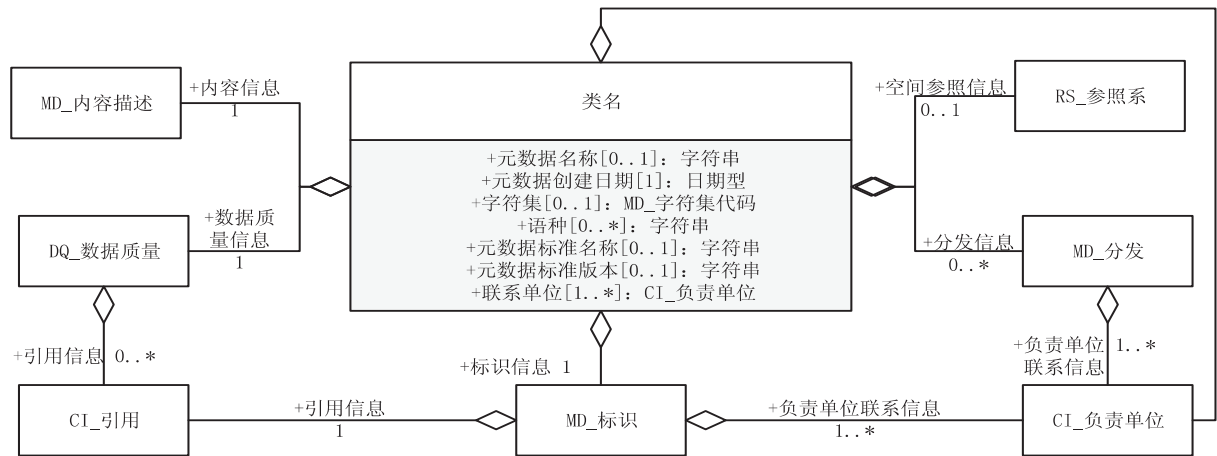


图 2 元数据 UML 图

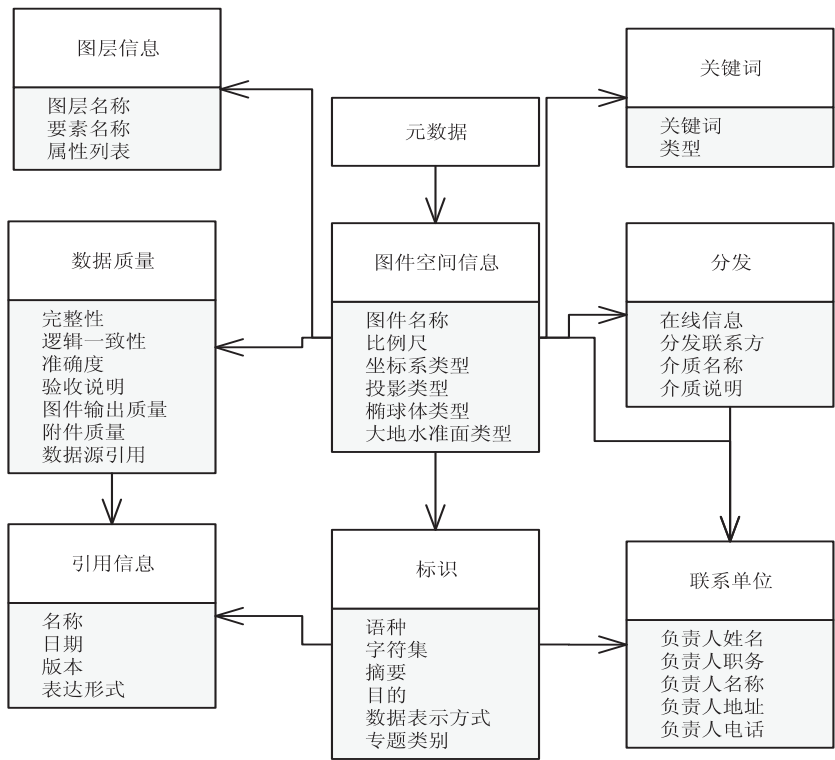


图 3 元数据在数据库的物理模型图

- (1) 查询过程中频繁涉及表连接操作;
 - (2) 应用程序执行时需将表连接查询;
 - (3) 在数据计算时需要进行复杂查询或临时表;
 - (4) 简化应用逻辑。
- 在数据库操作中为了生成报表或屏幕可能会涉及到十几个甚至更多的表检索一条特定信息,如果在数据库中存储这条信息的冗余,便能降低程序的逻辑复杂度^[8]。前面已经介绍了地质元数据的结构,了解其逻辑关系,一旦应用关系数据库管理元数据,一条元数据记录,被“拆分”为字段值存储在关系数据库中,这样在操作过程中则势必会存在多个表的链接,查询,以及生成临时数据表等。基于使用非规范化技术的上述

原因,系统将利用非规范化技术设计数据表,以简化录入、读取操作。根据非规范化思想设计数据表 geometa (见表 2)。

在 geometa 数据表中,table_name,存储元数据库中的数据表名,应用非规范化的合并表方法。对于频繁涉及表链接操作的表,将这些表的表名作为数据项混合存储在 table_name 列,达到降低表链接操作代价的目的。field_name,存储元数据库的结构信息,即整个库中所有表的字段名,举例该库共有两个关系表,table1 中共有 A,B,C 三个字段,table2 中有 D,E 两个字段,此 5 个字段全部作为数据项存储在 field_name 列中。field_capt,即字段的中文名,现各行业或专题元

数据制定标准基本采用英文或字母表示,因此将其中文说明存储在 field_capt 列方便用户阅读。meta_value,引入冗余列,元数据值存储在该列。另外值得一提的是,虽然 field_capt 列存储的是冗余数据,但是它们并不是固定不变的,随着程序机制不同,数据值随即会发生变化,但始终来源于元数据库,因此巧妙地避开了冗余数据一致性难以维护的缺点。通过实践表明,在数据库建模的过程中,利用非规范化平均能够在 10 个性能因素中提高一个。

表 2 非规范化表 geometa

名称	长度	类型	注释
table_name	16	文本	数据表名
field_no	3	文本	字段编号
field_name	32	文本	字段名
field_capt	48	文本	字段说明
recod_no	3	文本	记录编号
meta_value		备注	元数据值
attcode	3	文本	数据表名

2.4 数据字典技术

数据字典(Data Dictionary)是一种用户可以访问的记录数据库和应用程序源数据的目录,是系统中各类数据描述的集合,是进行详细的数据收集和分析所获得的主要成果。具体上,数据字典是通过一定的专业知识对某个系统数据库中所有用户数据的抽象和概括,以及把相关准则转化为数据库中的数据,并在程序中有所体现,具有灵活性和易操作性等优点^[9-12]。

地学元数据的结构,逻辑关系,数据类型等都由定义于它的元数据标准产生,即 Schema 语言。Schema 是用于描述和规范 XML 文档的逻辑结构的一种语言,由它定义数据元素的数据类型,包括字符串、数字、日期等,元素是否为必须元素,或者元素值只能为标准提供的数据范围,即枚举类型,这些都是定义元数据的基本要素同时也是元数据在数据库中数据约束的基准。因此将定义元数据标准的这些数据约束应用数据字典管理,将其固化。如数据类型,在数据库设计中,用户自定义字典涵盖了 90% 以上的数据类型,因此只需在字典中挑选符合要求的数据类型。如枚举类型数据,最直接方法之一就是为用户提供一组预先设定的选项,这样既可保证在数据库中不存储无效数据,又可提高用户数据输入效率^[13]。合理运用数据字典可确保数据库数据的准确性。将所有枚举值都存入数据字典中,限定范围,高效利用。

3 实践研究

地学元数据管理系统涵盖了元数据的导入、导出

及编辑等一般操作,还包括数据质量检查,数据转换,注册标准,批量编辑,数据模板等其他非常实用的功能。下面介绍几个主要的功能。

3.1 数据录入

根据元数据的 UML 图了解到一条元数据在关系数据库对应的存储结构和逻辑关系。如对某省的图件的元素录入,会涉及元数据实体 8 个表以及隶属各子集的表的操作。通过设计唯一标识码 ID 列保证元数据的逻辑关系和完整性。在元数据入库的常规操作中,通过标识列的 ID 值遍历表进行相应的操作,在这种频繁链接表操作的情况下,应用非规范化表 geometa,从 XML 文档获取的父节点,子节点结构信息和元数据值,存储在 geometa 中的 table_name、field_name、meta_value 三列中。这样便拆开了元数据的逻辑关系,可理解为:将连续的数据按照一定的规则分解成离散的数据。因为 geometa 中存储了库中的表名、字段名、表关系等信息,只需要一条命令即可将刚读入的元数据值插入对应的表对应的列中,而不再需要通过标识码 ID 去依次遍历或者递归各个数据表,这个过程也可视为具有数层的数据操作转变为 2 层数据操作。对于批量录入元数据在效率上将会有很大的提高。

3.2 定制元数据模板和元数据标准

在地质元数据中,各个省市或专题元数据在某些方面有相似之处,如影像轨道标识、空间分辨率、数据表示等一些标准或规范信息相似或相同,这样通过关系数据库管理便体现出了它的优势。很容易地将数据库中存储的数据值按照出现的概率进行分类,或根据其他分类标准进行分类,如专业分类、区域范围分类等。系统将分类的数据根据元数据标准定义的结构生成多种元数据模板,并填充数据值。

而对元数据名称、地理坐标范围等信息,因为研究区域、目的、专业等差别而存在明显差异的元数据值,通过系统提供的模板再根据差异信息进行修改或赋值,从而达到通过定制模板输出对应的地质信息元数据文件的目的。这说明建立一个完整的元数据库很有必要,它可以根据已有的元数据信息为后续使用提供参考和依据,而且随着相关技术的发展和元数据信息的积累,会存在通过元数据来检查相关地质图件的质量、数据完整性以及正确性的可能,而不需要直接打开图件进行相关操作。

在数据字典中有各种功能的数据表控制数据表结构信息,其中的 tablereport 数据表,是核心数据表,里面存储了整个元数据库中所有数据表的属性结构及隶属的数据表信息。当然元数据标准无疑会发生变化,如林业、海洋等其他领域的元数据,一旦标准发生变化,其数据结构也会发生相应的改变,这时只需将不同

的标准读入 `tablereport` 中,字典将会根据此表在数据库中生成新标准的数据结构,这样新的数据结构就会应运而生,相当于完成了初始化工作,这样体现了数据架构变化程序不变的便捷之处。

3.3 元数据批量处理

将元数据应用关系数据库管理的另外一个优势便是批量处理。它跟元数据模板定制的思想如出一辙。如一百幅图件的元数据由普通的文本文件存储管理,如果查询、修改则需要打开文件上百次,相当耗费精力。一旦采用关系数据库管理,便可按照一定的检索条件呈现符合要求的元数据,其查询、更改等操作也可按照用户自行设置,这样的管理效率更高,数据的准确性也会大大提升。另外一个优势是将所有元数据集中管理有利于数据挖掘技术的使用。当存储大量数据后,由专业人员进行简单的查询就能反应其中一些比较明显的规律。如各省的地质背景情况,某区域与某种矿物存在的关系以及延伸到空间上的联系等规律,这些特质在以往的文本文件管理中是体现不出来的,因此可进一步利用这些规律来反向检查图件质量。

4 结束语

地质元数据管理系统,以一种新的管理方式即通过关系数据库来管理元数据,同时通过非规范化技术和数据字典技术来弥补数据录入输出的效率问题,这样巧妙地抵消了输入输出的时间消耗问题,反而在后期体现出它的价值,利用已有数据为后续元数据编辑,定制提供了部分数据源和参考。它不仅仅是元数据管理软件,同时也是一款元数据编辑器,为不同行业的数

据进行元数据编制,方便而快捷。

参考文献:

[1] 万江波,邹逸江. 空间数据仓库元数据的认知过程[J]. 测绘科学,2008,33(1):58-61.

[2] 陈惠荣,游 雄. 地理空间元数据及其相关技术的探讨[J]. 测绘学院学报,2003,20(4):290-292.

[3] 地质信息元数据标准[S]. 北京:中国地质调查局,2006.

[4] 赵改善,曹邦功. 元数据:勘探开发数据管理的一种新工具[J]. 石油物探,2002,41(2):236-242.

[5] Abraham S, Henry F K. Database system concepts[M]. New York:McGraw Hill Higher Education,2010.

[6] Powell G. Beginning database design[M]. Birmingham:Wrox,2007.

[7] 孙 睿,刘 磊,邵明珠. 数据库的规范化理论与非规范化设计[J]. 电脑知识与技术:学术交流,2006(4):23-24.

[8] Dorsey P, Hudicka J R. Oracle8 database design using UML object modeling[M]. New York:Oracle Press,1998.

[9] 萨师煊,王 珊. 数据库系统概论[M]. 北京:高等教育出版社,2000.

[10] 马小刚,汪新庆,毋丽红,等. 应用数据字典实现多源地质空间数据的通用管理[J]. 矿业研究与开发,2007,27(1):37-40.

[11] 杨圣伟,汪新庆. 数据字典在煤炭数据发布平台中的应用[J]. 煤田地质与勘探,2008,36(6):17-19.

[12] 葛 艳,汪新庆. 用 ASP 和数据字典技术解决网络数据库中通用动态查询的问题[J]. 计算机与现代化,2004(5):75-77.

[13] 韩志军,汪新庆. 数据库系统数据字典的设计与实现[J]. 微机发展(现更名:计算机技术与发展),1999,9(2):30-32.

[9] 刘 昱,胡晓爽,段继忠. 新一代视频编码技术 HEVC 算法分析及比较[J]. 电视技术,2012,36(20):45-49.

[10] Texas Instruments. TMS320C6000 optimizing compiler v7.4 User's guide[EB/OL]. 2012. <http://www.ti.com/lit/ug/spru187u/spru187u.pdf>.

[11] Texas Instruments. Hand-tuning loops and control code on the TMS320C6000[EB/OL]. 2010. <http://www.ti.com/lit/an/spra666/spra666.pdf>.

[12] Texas Instruments. TMS320C66x DSP CPU and instruction set reference guide[EB/OL]. 2012. <http://www.ti.com/lit/ug/sprug7/sprug7.pdf>.

[13] Pescador F, Garrido M J, Juarez E, et al. On an implementation of HEVC video decoders with DSP technology[C]//Proc of IEEE international conference on consumer electronics. [s. l.]; IEEE,2013:121-122.

[14] 曾接贤,郑大芳,符 祥. 基于运动矢量空间相关性的 H. 264 分像素运动估计[J]. 计算机工程与应用,2013,49(15):175-178.

(上接第 174 页)

[3] 梁 凡. AVS 视频标准的技术特点[J]. 电视技术,2005(7):12-15.

[4] Sullivan G J, Ohm Jens-Rainer, Han Woo-Jin, et al. Overview of the High Efficiency Video Coding (HEVC) standard[J]. IEEE Transactions on Circuits and Systems for Video Technology,2012,22(12):1649-1668.

[5] 蔡晓霞,崔岩松,邓中亮,等. 下一代视频编码标准关键技术[J]. 电视技术,2012,36(2):80-84.

[6] Texas Instruments. TMS320C66x CorePac user's guide[EB/OL]. 2011. <http://www.ti.com/lit/ug/sprugw0c/sprugw0c.pdf>.

[7] Pescador F, Chavarrias M, Garrido M J, et al. Complexity analysis of an HEVC decoder based on a digital signal processor[J]. IEEE Transactions on Consumer Electronics,2013,59(2):391-399.

[8] 郭晓珉. 基于 FPGA 的 H. 264 运动估计算法优化与实现[D]. 南京:南京航空航天大学,2012.

基于非规范化和数据字典的地学元数据管理

作者：[周敏](#)，[汪新庆](#)，[ZHOU Min](#)，[WANG Xin-qing](#)
作者单位：[周敏, ZHOU Min\(中国地质调查局 西安地质调查中心 测绘遥感处, 陕西 西安, 710054\)](#)，[汪新庆, WANG Xin-qing\(中国地质大学 地质过程与矿产资源国家重点实验室, 湖北 武汉, 430074\)](#)
刊名：[计算机技术与发展](#)[ISTIC](#)
英文刊名：[Computer Technology and Development](#)
年，卷(期)：2015(3)

引用本文格式：[周敏](#). [汪新庆](#). [ZHOU Min](#). [WANG Xin-qing](#) [基于非规范化和数据字典的地学元数据管理](#)[期刊论文]-[计算机技术与发展](#) 2015(3)