

地球科学数据的元数据研究

李 军

(陕西师范大学旅游与环境学院 西安 710062)

陈崇成

(福州大学环境与资源工程系 福州 350002)

摘 要、元数据是数据库领域中一个基本的概念,地球科学数据的元数据系统的建立有助于地学数据的开发和利用,文中说明了元数据在地学数据中的应用,论述了地学元数据的分类、获取、管理等问题。

关键词 元数据 地学元数据 元数据分类 元数据管理 数据共享

分 类 中图法 TP311.13

随着科技进步,特别是信息系统的发展和应用,社会对各种数据的需求急剧增加。从科学研究、各行业、各部门应用到政府决策都离不开大量详实、准确的数据。到1994年底全球已建成有一定规模的数据库已有3 000多个。用户对不同类型数据的需求,以及数据库的大量出现,要求数据库的内容、格式、说明等符合一定的规范和标准,以利于数据的交换、更新、检索、数据库集成以及数据二次开发利用等,而这一切都离不开元数据(Metadata)。地球科学数据是指地学研究领域的数据,而空间特征是地学数据区别于其它数据的最显著的特性,为了有效地生产和利用地球科学数据,要求地学数据规范化和标准化。所以地球科学数据库不但要提供基本的空间和属性数据,还应该包括大量的引导信息以及由纯数据得到的推理、分析和总结等。这些都要由地学数据的元数据系统来实现。

1 元数据(Metadata)的内涵

到目前为止,科学界仍没有关于元数据的确切公认的定义。元数据的最简短的定义是:“元数据是关于数据的数据”,但这种定义不能清楚描述元数据是什么。许多专家学者从不同的侧重点出发给元数据以不同的描述。Ashrafi kuilboer^[1]认为元数据是数据库管理领域的概念,是关于数据组织的数据;Bretherton 和 T. Singley^[2]认为元数据是对数据的描述以及对数据集中数据项的解释,它能提高数据的利用价值;Hans-J. Lenz^[3]认为:在统计型数据库中元数据是关于宏数据(Macrodata)和基本数据(Microdata)的数据,元数据没有标准属性,即没有固定格式;而 Kapetanios^[4]等人则认为元数据是数据与信息之间的某种东西,它可以沟通数据和信息。根据以上定义,参考其他学者的观点,关于元数据有以下说明:

(1) 元数据的目标:元数据的根本目标是使数据库更易于使用;另一个目标是为计算机辅助软件工程(CASE)服务。

(2) 元数据的内容: 元数据包括对数据集的描述; 对数据集中各数据项; 对数据来源、数据所有者及数据序代(数据生产历史)等的说明; 数据质量的描述, 如: 数据精度、分辨率、源数据的比例尺等; 数据处理信息, 如量纲的转换等; 数据转换方法; 数据库更新、集成的方法等。

(3) 元数据的性质: 元数据是数据的描述性数据; 对不同领域的数据库, 元数据的内容有很大差异; 元数据应尽可能多反映数据的特征及规律。

(4) 元数据的作用: 通过元数据可以检索、访问数据库, 可以有效利用计算机的系统资源, 可以对数据进行加工处理和二次开发等。

在此基础上, 我们认为元数据是以数据高效利用和交换为目的数据集说明性数据, 它主要包括对数据集、与数据集相关信息、数据集各数据项说明以及数据用户访问、检索、更新数据库的方法, 同时元数据也包括基于不同数据领域, 如何尽可能全面反映基本数据的信息。

2 与元数据有关的概念

元数据的内容及实现方法, 本是数据库管理领域的内容, 但数据应用领域的拓宽, 数据交换量的增加, 特别是不同学科领域中数据的交叉利用, 要求数据用户对元数据有一定了解, 才能较好地利用地学数据库及其它数据库; 然而元数据这一概念并未被广泛接受, 有关其内容亦未达到共识, 元数据的内容可能或多或少存在于其它有关数据的说明集中, 以下简要列出一些元数据有关的概念。

(1) 数据字典(Data dictionary) 数据字典是易接受和易理解的概念。它是关于数据集图解描述、数据集图解、数据和产品数据字典形式的文件集^[5], 是管理元数据的基础。鉴于当前数据字典不再单是数据结构的简单描述文档, 它已正超出了普通字典的特征, 所以一些学者和数据生产者建议取消该概念。

(2) 数据百科全书(Data Encyclopedia) 数据百科全书是在数据字典的基础上形成的, 其内容比数据字典更丰富, 它不仅包括物理层的数据结构, 还包括结合计算机辅助设计工具的数据发展过程等。

(3) 数据仓库(Repository) 数据仓库是数据百科全书的雏形, 它不包括数据生产者的说明, 数据仓库用于存贮和控制元数据库。

(4) 元数据库(Metadatabase) 元数据库是信息和数据的集成, 关于数据过程模型、数据生产规则及数据模型的信息系统。

(5) 元数据管理信息系统(Metadata Information Management System) 是用于元数据及其相关过程, 集成化、标准规范化的管理系统, 得益于该系统, 数据用户可以便捷地了解数据库的内容, 获取自己所需数据或是集成示同的数据库。

3 元数据的类型

不同领域数据的元数据, 不论从内容上, 还是从结构和形式上都有较大的差异。因而, 元数据的分类体系也不尽相同。

3.1 根据全球信息源字典 (Global Information Resources Dictionary. GRID) 模型^[6]分类

(1) 整体功能 是各子系统基本组织的最高层描述, 由它可使用户在可操作水平上理解各子系统及各子系统之间的相互作用。

(2) 功能模型 其内容更详细, 它描述主要模型 (关于数据的和关于数据处理的) 及各子系统所包含的决策逻辑。主要模型包括: 数据模型、数据流模型和高级计算机辅助软件工程 (CASE) 模型。

(3) 结构模型 主要描述数据库的框架结构及其基本知识, 内容有: 数据结构、规则结构、规则内容等。

(4) 用户与数据源 指对数据集中实体的描述, 内容包括硬件、文件和文档等。

3.2 根据数据的内容分类

造成元数据内容差异的主要原因有两个: 其一, 不同性质、不同领域的的数据 (如: 物理数据和空间数据) 所需要的元数据内容有差异; 其二, 为不同目的而建设的数据库其元数据也有很大的差异。在此基础上, 元数据可分为三种类型,

(1) 科研型元数据 其主要目标是帮助用户获取各种来源的数据及其相关信息, 它不仅包括诸如: 数据源名称、作者、主体内容等传统的、图书管理式的元数据, 还包括数据拓扑关系、数阵关系等。这类元数据的任务是帮助科研工作者高效获取所需数据。

(2) 评估型元数据 主要服务于数据利用的评价, 内容包括数据最初收集情况、收集数据所用的仪器、数据获取的方法和依据、数据处理过程和算法、数据质量控制、采样方法、数据精度、数据的可信度、数据潜在的应用领域等。

(3) 模型元数据 用于数据模型的元数据与数据的元数据在结构上大致相同, 其内容包括: 模型名称、模型类型、建模过程、模型参数、边界条件、作者、引用模型描述、建模使用软件、模型输出等。

3.3 根据元数据描述对象的不同分类

(1) 数据层元数据 指描述数据集中每个数据的元数据。内容包括: 日期戳 (指最近更新日期)、位置戳 (指示实体的物理地址)、量纲、注释 (如: ‘关于某项的说明见附录’)、误差标识 (可通过计算机消除)、缩略标识、存在问题标识 (如: 数据缺失原因)、数据处理过程等。提取是关于每个具体数据项、每个数据的元数据。

(2) 属性元数据 是关于属性数据的元数据, 内容包括为表达数据及其含义所建的数据字典、数据处理规则 (协议), 如: 采样说明、数据传输线路及代数编码等, 协议的确定包括: 列举手册、相应命令、相应类的含义、修改算法的信息等。

(3) 实体元数据 是描述整个数据集的元数据, 内容包括: 数据集区域采样原则 (指区域性数据库)、数据库的有效期、数据时间跨度等。

3.4 根据地学元数据的作用分类

按地学元数据的作用可以把元数据分为两种类型, 即说明元数据和控制元数据。说明元数据是专为用户直接使用的元数据, 它一般用自然语言表达, 如: 源数据覆盖的空间范围、源空间数据图的投影方式及比例尺的大小、数据集说明文件等, 这类元数据多为描述性信息。控制元数据是指用于人为间接作用于数据库或用于计算机操作流程控制的元数据, 这类元数据由一定的关键词和特定的句法来实现, 其内容包括: 数据存贮和检索文件、检索中与目标匹配方法、目标的检索和显示、分析查询及查询结果排列显示、根据用户要求

修改数据库中原有的内部顺序、数据转换方法、空间数据和属性数据的集成、根据索引项把数据绘制成图、数据模型的建设和利用等。两种元数据没有严格的界限,前一类侧重于数据库的说明;后一类元数据主要是与数据库操作有关的方法等。

4 元数据的获取与管理

4.1 元数据的获取

地学元数据的获取是个较复杂的过程。相对于基础数据 (Primary Data) 的形成时间,元数据的获取可分为三个阶段,即:数据收集前、数据收集中和数据收集后。对于模型元数据,则这三个阶段分别是模型形成前、模型形成中和模型形成后三个阶段。第一阶段收集的元数据主要指根据要建设的数据库的内容而设计的元数据,内容包括:(1) 普通元数据,如:数据类型、数据覆盖范围、使用仪器描述、数据变量表达的内容、数据收集方法等。(2) 专指性元数据即针对要收集的特定数据(如中国 1950—1980 年 30 年间的逐旬降水数据)的元数据,内容包括数据采样方法、数据覆盖的区域范围、数据表达的内容、数据时间、数据时间间隔、空间上数据的高度(或深度),使用的仪器、数据潜在的利用等。第二阶段中,元数据随数据的形成同步产生。如:在测量海洋要素时,一般情况下,海洋测点的水平和垂直位置、深度、温度、盐度、流速、海流流向、表面风速、仪器设置等是同时得到的。第三阶段指数据收集到后,根据需要收集的元数据,这些元数据只有在数据收集到以后才能获取的元数据包括:数据处理过程描述、数据的利用情况、数据质量评估、浏览文件的形成、拓扑关系、影像数据的指示体及指标、数据集大小、数据存放路径等。

同其它数据元数据的获取方法一样,元数据的获取方法主要有五种,即:键盘输入、关联表、测量法、计算法和推理法。键盘输入一般工作量大且易出错,如有可能应尽量避免,但对某些元数据而言(如数据变量表达的内容)只能由键盘输入;关联表方法是通过公共项(字段)从已存在的元数据或数据中获取有关的元数据,例如:通过区域的名称从数据库中得得到区域的空间位置坐标等;测量方法容易使用且出错较少,如用全球定位系统(GPS)测量数据空间点的位置等;计算方法指由其它元数据或数据计算得到的元数据,如:水平位置可由仪器设置及时间计算得到,区域的面积可由多边形拓扑关系计算出来。该方法一般用于获取数量较大的元数据;推理方法指根据数据的特征获取元数据。

在元数据获取的不同阶段,使用的方法也有差异,在第一阶段主要是键入方法和关联表方法;第二阶段主要采样测量方法;第三阶段主要是计算和参考方法。

4.2 元数据的管理

地学元数据的管理的理论和方法涉及到数据库和元数据两个方面。由于元数据的内容、形式的差异,元数据的管理与数据涉及的领域有关,它是通过建立在不同数据领域基础上的元数据信息系统实现的。国外在元数据方面已有许多成功的经验,下面列举了两种方法供参考。

IGBP 为满足世界各地用户对数据的需求建立了 IGBP 元数据信息系统,它由三个主模块组成(图 1),即一个关系数据库、地理信息系统和用户界面核心软件。元数据在数据中心形成后提供给 IGBP,然后经过加工处理进入数据库,并根据元数据性质分类,分别存放,对关系数控的操作是通过建立在地理信息系统(GIS)基础上的中心软件来实现的,该软件

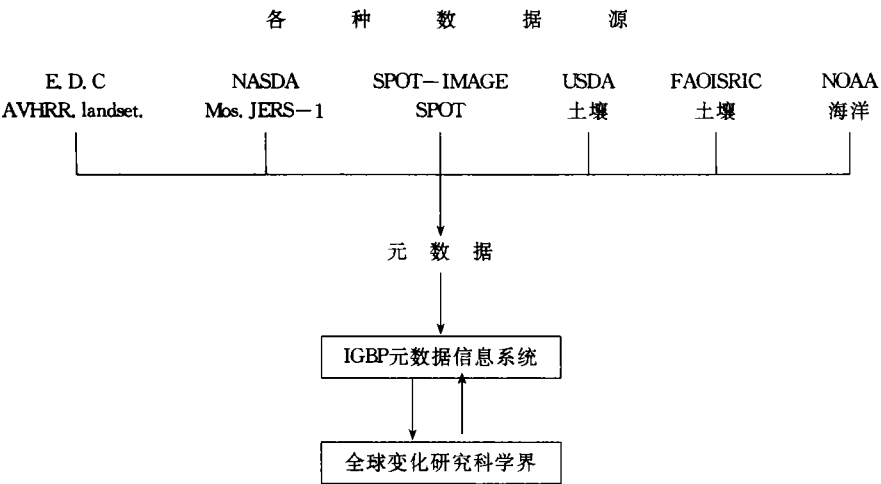


图1 IGBP元数据信息系统示意图
Fig. 1 Diagram of IGBP metadata information system

提供了用户利用元数据信息系统的交互式界面,通过 Intnet 网络,世界各地的数据用户可与分布在 IGBP 核心工程中心及各个合作实验室的 IGBP 元数据信息系统联系。

Francis. P. Bretherton 等^[2]认为元数据管理可通过元数据库 (Metadatabase) 实现 (图 2),在该系统中,物理层存放数据与元数据,该层由一些软件 (如 WAIS) 通过一定的逻辑关系与逻辑层关联起来,如由颜色 (属性)、字符串型数据 (数据类型) 可以同绿色 (值) 关联起来。在概念层中用描述语言及模型定义了许多概念,如:实体名称、别名、允许属性值的类型、缺省值、允许输出及输入的内容、临时实体的作用、元数据的变化、操作模型等。通过这些概念及其限制特征,经过与逻辑层关联可获取、更新物理层的元数据及数据。另外,全球信息源字典 (GRID) 采用两步实体关系模型 (Two-stages Entity Relationship Model) 来管理元数据。

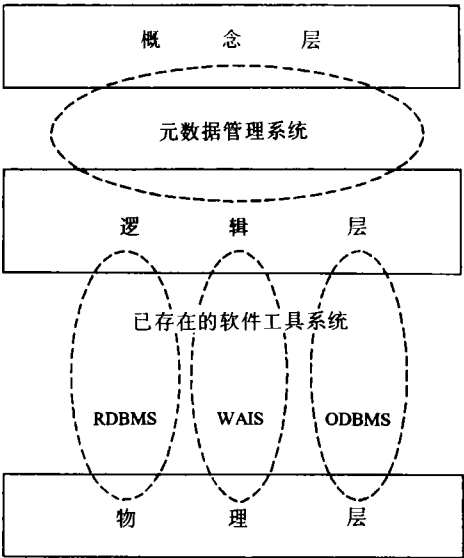


图 2 元数据管理信息系统
Fig. 2 Metadata management information system

5 地球科学数据中元数据的用途

5.1 帮助用户获取数据

通过元数据,用户可对空间数据库进行浏览、检索和研究等。一个完整的地球科学数据库除应提供空间数据和属性数据外,还应提供丰富的引导信息,以及由纯数据得到的分

析、综述和索引等。通过这些信息用户可以弄明白诸如：“这些数据是什么数据？”；“这个数据库对我有用吗？”；“这是我需要的数据吗？”；“怎样得到这些数据？”等一系列问题。地球科学数据涉及许多方面，如：大气、海洋、地质、气象气候、土壤、植被、水系、灾害、生态脆弱带等，地球科学数据由于其空间特征，与其它数据存在着较大的差异，即使在地学内部，不同领域的研究者也难于完全掌握其它地学分支领域的的数据及其特征，用户可通过数据库提供的元数据信息，迅速了解数据库的内容，并对数据进行浏览和检索，以获取自己需要的数据部分。如：中国（1：400 万）县区数据库除有空间数据和属性数据（如：县区编码、县区名称、县区的大小、相邻的县区是什么等），通过元数据，用户还能得到如：数据覆盖范围、源空间数据地图的投影方式、数据生产日期、数据的变动、数据的精度和数据量的大小等，在此基础上，用户便可确定是否需要该数据。

5.2 数据质量控制

不论是统计数据还是空间数据都存在数据精度问题，影响空间数据精度的原因主要有两个方面：一是源数据的精度；一是数据加工处理工程中精度质量的控制情况。地球科学数据应该：

（1）有准确定义的数据字典，以说明数据的组成，各部分的名称，表征的内容等。

（2）保证数据逻辑科学地集成，如：植被数据库中不同亚类的区域组合成大类区，这要求数据按一定逻辑关系有效的组合。

（3）有足够的说明数据来源、数据的加工处理工程、数据释译的信息。

这些要求可通过元数据来实现，这类元数据的获取往往是地学、计算机领域的工作者来完成的。数据逻辑关系在数据中的表达要由地学工作者来设计；地学数据库的编码要求一定的地学基础。数据质量的控制和提高要有数据输入、数据查错、数据处理专业背景知识的工作人员；而数据再生产要由计算机基础较好的人员来实现。所有这方面的元数据按一定的组织结构集成到数据库中构成数据库的元数据信息系统来实现上述功能。

5.3 在数据转换中的应用

处于不同平台的、不同数据库的内部结构和管理方法往往有一定的差异，如：面向对象的数据库和关系数据库的管理体制方法有较大的差异，因而不同数据库之间数据的转换是必要的，并且这一转换过程应尽可能在无人干预的条件下实现，元数据具有这些功能，并且在转换工程中，通过交互式界面可以增加、删除或是更新数据库的内容，这要求在不同水平层次上的交互过程，不但要有一个友好的界面和数据转换标准，而且要有一个可控制这些转换标准的用户选择功能，由此用户得到想要的数据库。空间数据有多种数据格式，如：SDTS、UNGEN、GRIDASCII、DLG 等，不同用户的软硬件平台可能需要某种格式的数据，或某种投影方式的数据，这就要求数据格式的转换，再者数据库提供的可能是矢量数据，而用户需要栅格数据，且对栅格数据中网格的大小有一定的限制，这要求地学数据元数据系统不但能实现上述转换，还应该允许用户介入自行确定如网格大小等参数。

5.4 数据存贮和功能实现

元数据用于数据库的管理，可以有效降低数据存贮的空间，减少数据用户查寻数据库及获取数据的时间，从而降低数据库的费用。数据库的建设和管理费用是数据库整体性能的放映，通过元数据可以实现数据库的设计和系统资源的利用方面开支的合理分配，数据库许多功能（如：数据库检索、数据转换、数据分析等）的实现是靠系统资源的开发来实

现的,因而这类元数据的开发和利用将大大增强数据库的功能并降低数据库的建设费用。

6 我国地学数据库元数据的设计

地学数据元数据管理信息系统的建设和元数据功能的实现与数据库管理技术及地理信息系统的发展有直接的联系,从我国地学数据分布和内容出发,现阶段我国地学数据元数据的建设应集中在以下几个方面:

(1) 数据源信息 主要包括:数据的研制单位、数据的存贮界质、比例尺、数据的发行单位、数据的变化等。

(2) 数据库的说明 指对地学数据产品的说明信息,内容包括:数据的表达方式、数据的精度、数据质量的控制、数据存贮格式、空间和属性数据的组织方式、数据的存贮界质和数据量、数据二次开发的接口等。

(3) 数据格式的转换 不同的数据应用平台要求的数据格式也有很大的差别,元数据应实现数据格式的转换,因而元数据应提供关于数据格式、数据存贮、数据投影方式的说明以及数据格式转换和投影变化的方法。

(4) 数据共享方法 数据共享的前提是数据的标准化和规范化,地学元数据应提供地学数据生产的规范流程和标准,并就当前条件下以交换和低价供应方式实现数据共享。

7 结论

在国内,数据产业刚刚起步,对于数据的研究不多,关于元数据的规范、标准、理论和实现方法手段的研究则更少。地球科学数据多分散在各部门或个人手中,但数据共享是社会发展的需求,它可以避免许多地学数据库建设中低级的重复劳动,提高现有地学数据的价值。地学元数据信息系统的建设是地学数据共享的基础,地学元数据可实现的功能有:

- (1) 通过地学元数据可获得关于数据集及数据集中各项的说明。
- (2) 通过地学元数据可得到数据要素的说明、数据库编码说明等。
- (3) 通过地学元数据可对数据库进行检索、查询、浏览和获取数据。
- (4) 地学元数据有数据格式转换、数据投影方式转换等功能。
- (5) 通过地学元数据可对地学数据库进行有效管理、更新、集成等。
- (6) 地学元数据可以较好的跟踪和控制数据质量。
- (7) 地学元数据信息系统为 GIS 软件高效利用地学数据提供了条件等。

参 考 文 献

- 1 Noushin Ashrafi. The Information Repository: A Tool for Metadata Management Journal of Database Management Vol. 2 No 2, 1995 spring.
- 2 Francis. P. Bretherton, Paul T. Singley, Metadata a Users' View IEEE 1994 PP166—176.
- 3 Hans—J. Lenz. The conceptual Schema and External Schemata of Metadatabases IEEE 1994 pp160—172.
- 4 Epaminondas Kapetanios . A Knowledge —Based System Approach for Scientific Data Analysis And the Notion of Metadata.

- 5 Huzzah A Data Dictionaries: Path to Standard Database Programming & Design, 1989, 2(8) 26—35.
6 Cheng Hsu. etc. A Metadata System Information Modeling and Integration IEEE 1990 PP616—624.

A STUDY ON THE METADATA OF EARTH SCIENCE DATA (GEO-METADTA)

Li Jun

(*School of Tourism and Environmental Sciences Shaanxi Normal University, Xi'an 710062*)

Chen Chongcheng

(*Dept. of Environmental and Resources Engineering Fuzhou University, Fuzhou 350002*)

Abstract

Metadata is an important concept in the field of database management, and the metadata of earth science data which is called geo-metadata is valuable for earth-scientists to use and explore earth science data. Metadata is not a new concept, but the application of metadata of earth science data is extended only by the utility of computer technology and the development of geographical information system (GIS). Geo-metadata are used to prompt management and utility of earth science database which is different from ordinary database because of its spatial characteristics. Up to now, an universally accepted definition of metadata have not been created yet, but general views about the utility, contents, characteristics, function of metadata have been universally accepted. Metadata is a concept which has relations with Data dictionary, Data Encyclopedia, Data Repository, Meta-database, and metadata information management system through which metadata can be manipulated and managed. The classification system of metadata varies due to contents, utility field, and function of metadata. The general methods for metadata collection are key-in (through computer keyboard), look-up tables (relates), inferring (from existing metadata or primary data), measurement (by some experiments), and computation (by existing data items). The function of geo-metadata includes: (1) to help earth science data users to get specified data easier, (2) to control data quality, (3) to convert data formats and transform data between different projection system, and (4) to descend the cost of earth science data management. Being at its infant stage, the development of GIS and data industry is eagerly necessary in China, so the study on the standards and the normalization of earth science data and its metadata is urgent. At the end of the paper, the author lists the purposes for the construction of geo-metadata in China.

Key words metadata, geo-metadta, metadata classification, metadata management, data sharing