DOI:10.3876/j.issn.1000-1980.2009.03.022

基于多维组合的水利科学数据分类体系及其编码结构

耿庆斋[1,2],张行南[1],朱星明[2]

(1.河海大学水文水资源与水利工程科学国家重点实验室,江苏 南京 210098;2.中国水利水电科学研究院,北京 100044)

摘要:为满足水利科学数据管理和共享的需要,在分析我国水利信息分类与编码研究现状以及现有分类体系面临的主要问题的基础上,依据水利科学数据所具有的特性,提出了多维组合的水利科学数据分类体系结构,构建了由科学属性、获取方法、数据载体和时空定位组成的多维水利科学数据分类体系,并对其进行了规范化的编码设计,形成了 3 段 18 位的多维码编码结构.目前,该分类与编码体系已在水利科学数据共享网中得到实际应用.

关键词:水利科学数据;数据分类;分类体系;数据编码;多维组合

中图分类号:G202;TV214 **文献标识码**:A **文章编号**:1000-1980(2009)03-0346-05

长期以来,我国水利行业、部门和单位积累了大量的水利科学数据资源,特别是 1949 年后,随着科学技术的迅速发展,水利科学数据更是成倍增长,已具备了数据积累的基础.国内主要的水利行政部门、数据管理部门、高等院校、科研机构以及很多涉及水利的企业都已建立了相应的水利信息数据库和数据库管理系统,但已建数据库由于当时目标所限,数据的分类和编码自成体系,无法满足数据管理、共享和信息化发展的需要.因此,建立一个完善、统一的水利科学数据分类与编码体系,对水利科学数据采集、加工、利用、共享、管理以及各类应用系统建立都至关重要.本文在充分分析水利科学数据的本质属性和特征的基础上,提出了基于多维组合的水利科学数据分类与编码体系.

1 信息分类编码研究现状

发达国家非常重视信息分类编码工作,美国从 1945 年起就开始研究信息分类编码标准化问题,并经过 6 年的时间完成了国家物资分类编码[1],而且陆续研制了一系列其他方面的分类编码,如《美国制造业产品分类编码》、《美国标准行业分类目录》及《美国和加勒比海边远地区水文体系的标识代码》等.前苏联、罗马尼亚、日本等国家也从 20 世纪 60 年代开始投入大量人力、物力进行信息分类编码研究工作[2-3].

一些国际组织也很重视信息分类编码标准化研究工作,以实现世界范围内信息资源的共享.国际劳工组织在 20 世纪 60 年代末发布了 ISCO—88《职业分类国际标准》,联合国从 70 年代开始陆续颁布了一些标准,如:ISIC Rev.3《行业分类国际标准》、SITC Rev.4《国际贸易分类标准》、CPC—1998《主要产品分类》.1999 年联合国开发署又正式推出 UN/SPSC—1998《联合国产品和服务分类编码标准》.国际标准化组织(ISO)也相继制订了一系列有关信息分类编码方面的标准,如 ISO 3166:1993《国家与地区名称代码》、ISO 4217:2001《货币与资金代码》等.欧洲共同体、关税合作委员会等国际组织在近 30 年来也编制发布了相当数量的分类编码标准[4].

我国于 1979 年起开展信息分类编码标准化工作,在研究一些发达国家和组织的分类编码标准基础上,取得了较快的发展,很多研究成果已作为国家标准或行业标准正式颁布[5-11].“十五”期间,在国家“十五”科技攻关项目“中国可持续发展信息共享系统的开发研究”中,通过分析研究面向人口、资源、环境、灾害、经济与社会等可持续发展领域的信息资源,提出了中国可持续发展信息分类和编码标准[12].在科学数据共享工程中,制订了 SDS/T 2122—2004《科学数据分类编码方案》和 SDS/T 2121—2004《数据分类与编码的基本原则

收稿日期:2008-06-02

基金项目:国家科技基础条件平台工作重点项目(2004DKA20220)

作者简介:耿庆斋(1979—),男,山东泰安人,博士研究生,主要从事水文水资源与水信息学等研究.

与方法》,科学数据共享工程中的多数试点项目在以上 2 个标准的基础上研制了各自的科学数据分类编码,如《气象资料分类编码和命名规范》、《地震科学数据分类与分级方案》、《地球系统科学数据分类体系》、《医药卫生科学数据分类与编码》等[13-15].以上研究成果已应用于科学数据共享工程中,为科学数据共享工作的顺利开展和标准化建设打下了良好的基础.

2　我国水利信息分类与编码研究现状

2.1　我国水利信息分类与编码现状分析

水利信息分类编码是进行水利信息交换和实现信息资源共享的重要前提,是水利科学数据共享标准化的一项最为重要的工作[16].水利科学数据分类与编码标准研究是水利信息化标准研究的组成部分之一,因此,在进行水利科学数据分类与编码体系研究之前,着重分析我国在水利信息分类与编码标准方面取得的研究成果.

水利作为我国古老的行业之一,其信息已按照多种方式进行了分类,并以不同的形式记录在不同的载体之上.为促进水利信息的科学管理和标准化,我国在水利信息的分类编码标准化方面进行了大量的研究,形成了一系列行业标准[10-11,17-24],这些标准改变了以往不同部门间使用各自信息分类编码的现状,解决了许多基础信息重复整编、互不统一、无法共享的局面,对推动我国水利信息化建设和信息共享起了重要作用.但是,由于这些标准多数是针对某一个专题或一项研究制订的,对于水利信息分类与编码的全局性考虑不充分.因此,为了适应水利信息化发展的需要,水利部已组织编制水利信息分类标准,以便建立一个完善的水利信息分类体系,这对水利科学数据分类编码的研究也具有一定的借鉴作用.

2001 年水利部颁布的《水利技术标准体系表》[25]针对水利信息化标准的现状和特点,将水利信息按照专业门类划分为水文、水资源、水环境、水利水电、防洪抗旱、供水节水、灌溉排水、水土保持、小水电及农村电气化、综合利用等类别,此分类为水利行业技术标准的管理和使用提供了便利,这也是对水利信息分类的一次有益尝试,具有重要的理论和实践价值.但是,该分类仅是针对我国水利技术标准成果的管理这一特殊用途的一种分类体系,随着水利信息化发展和应用需求,其分类层次和内容体系还有待进一步完善.

在国家层面上,我国于 1992 年颁布的国家标准 GB/T 13745—92《学科分类与代码》,将水利工程学科划分为水利工程基础学科、水利工程测量、水工材料、水工结构、水力机械、水利工程施工、水处理、河流泥沙工程学、海洋工程、环境水利、水利管理、防洪工程、水利经济学等[26].可以看出,该分类是一种学科分类,而且存在学科分类过粗的现象,与水利行业有关的学科未得到充分体现,不能满足水利行业的科学数据管理和科技创新工作的实际需要.尤其是在实际工作中产生的水利科学数据,国家学科分类与代码标准并不能完全适用.

2.2　现有分类体系面临的主要问题

综上所述,根据不同的业务需求和管理要求,从不同的角度出发,形成了不同的水利信息分类体系,综观现有各分类体系,主要面临如下 2 个问题:

a. 现有的水利信息分类不能完全满足水利科学数据共享分类的要求.现有水利信息分类主要偏重于分类的全面性和系统性,属于学科意义上的分类,是根据对象的属性或特征加以区分和类聚,并将区分的结果按照一定次序进行归类,然后依此制定了大量的国家标准和多种行业标准,如 GB/T 13745—92《学科分类与代码》和《水利技术标准体系表》中的水利信息分类体系.水利科学数据共享分类的目的在于数据资源的管理、利用和共享,是把分类方法应用于科学数据管理,从利用的角度,考虑到数据使用者而对科学数据所作的分类,是面向水利科学数据集的分类.学科意义上的分类不是科学数据共享分类,不能满足科学数据面向管理的分布式、实用性的网络共享的要求[27].

b. 原有分类体系如何与共享分类体系对应.由于数据来自不同的部门,无论是数据质量、数据格式,还是数据组织结构和字段命名规则都存在差异.如果对现有数据进行标准化改造,生成新的符合一定标准的数据集,必然要破坏原有的分类.各个部门如果按照新的分类标准为现有数据分类,新的分类标准很有可能无法与基础数据完全匹配,破坏了数据集的封装,使得分类无法进行.因此,一方面为保证数据的完整性,原有分类体系不能轻易改变,另一方面又必须让数据进入共享体系,满足水利科学数据共享的需要,这也是一对亟待解决的矛盾.

3 水利科学数据分类体系及其编码

3.1 水利科学数据的多维属性

水利科学数据的分类就是根据其不同的学科、观测手段、自身属性等特征加以区分和类聚,并将区分的结果按照一定的次序进行归类[28].因此,要研究水利科学数据的分类体系,首先要分析水利科学数据所具有的特性.

a. 科学性.科学数据伴随着科技活动的开展而产生,因此水利科学数据具有科学性.

b. 多源性.由于水利科学数据可能来自不同的机构或个人,存在数据获取手段多源性、存储格式多源性等特点.

c. 多样性.水利科学数据的多源性决定了其多样性的特点,主要表现在:存储介质多样性、数据表达方式多样性等.

d. 多时空尺度.水利科学数据还具有时空特性,因此水利科学数据的一个特点就是多时间尺度和多时空尺度.

水利科学数据的科学性、多源性、多样性和多时空尺度等特性决定了水利科学数据具有多维属性的特点.

a. 科学属性维.科学属性维主要反映水利科学数据的科学属性,根据相对独立、相互间又有相关联系的知识体系,可以将数据划分成不同的从属关系和并列次序,组成一个有序的分类体系.水利科学数据的科学属性分类可参考 GB/T 13745—92《学科分类与代码》标准和水利行业的现行分类体系,划分为水文、水资源、水环境、水旱灾害、节水灌溉、水土保持、水利工程、水利经济等一级类目,它们各自还可细分为二级类目,原则上划分的层级不超过 3 级.

b. 获取方法维.获取方法维主要依据数据的获取方法进行分类,可分为观测数据、实验数据、调查数据、考察数据、统计数据、遥感数据、研究数据、图形数据、实物数据等.

c. 数据载体维.数据是客观存在的,又是虚无的,必须借助数字或符号才能表现出来,还必须将这些数字或符号寄载于某种物体上.数据载体维按数据的存储介质可分为纸质数据、薄膜数据、磁带数据、磁盘数据、光盘数据、网络数据等,其中网络数据主要指通过网络发布或流通的数据(集),是未来实现数据共享、交换的主要发展方向.

d. 时空定位维.设置水利科学数据的时空定位维主要反映水利科学数据的时间和空间特征.时间信息分类的依据主要是与数据直接或间接相关联的各种现象发生、发展的过程与周期,时间信息分为短时间尺度(秒、分、时、日)、中时间尺度(旬、月、季、年)、长时间尺度(十年、百年)和地质时间尺度(千年、万年).空间信息分类的依据主要是通过记录数据相关联的空间位置实现空间定位,空间信息分为小尺度(县乡镇或小流域级)、中尺度(省市或大流域级)和大尺度(国家级).

3.2 水利科学数据分类体系研究

3.2.1 多维水利科学数据分类体系的提出

科学高效的分类体系可以准确描述水利科学数据的内容,对水利科学数据的收集、检索、存储、共享、交换以及利用具有重要作用.在进行水利科学数据分类过程中,很难直接采用单一的特性对水利科学数据进行具体划分,针对现有水利信息分类体系所面临的 2 个主要问题,并结合水利科学数据具有多维属性的特点,构建了以科学属性、获取方法和数据载体 3 个维度(属性)为基础的坐标体系模型,并结合水利科学数据所具有的时空特性,加上时空定位维,形成一个多维组合的水利科学数据分类体系(图 1).

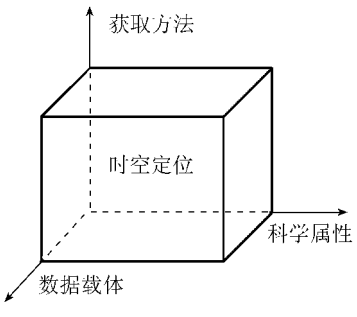

图 1 多维水利科学数据分类体系

Fig. 1 Multi-dimensional water science data classification system

3.2.2 水利科学数据分类体系结构

在图 1 中,每一维表示某一类数据,数据维中的数据都限制在某一个特定分类子系统之内.在水利科学数据分类体系中,每一个分类子系统映射为分类空间的一个数据维,每一个数据维包含一颗以维度为根级的分类树.根据分类情况的复杂程度,根级分类树包含若干子

树,子树由语义完整、相互独立的域构成,域是叶级分类.树、子树和域构成了该数据维下的层级分类结构,如图2所示.对于构成复杂的维度,可将其进一步映射为分类子体系,在子体系下进行维度划分,形成嵌套分类模型.

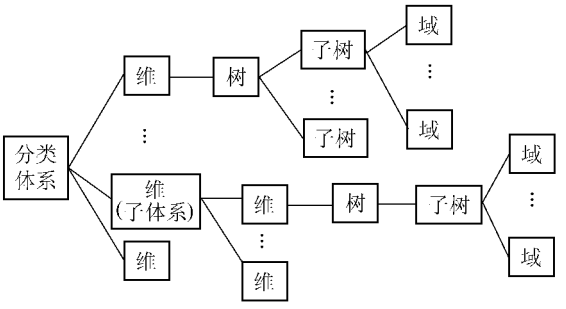

图 2　水利科学数据分类体系层次结构

Fig. 2　Hierarchy of water science data classification system

3.2.3　基于多维组合的水利科学数据分类体系的特点

a. 在对现有水利信息分类体系和水利科学数据的特点进行详细分析的基础上,实现了水利科学数据的多维信息的描述,该体系包括的类别信息比较完整,可满足水利科学数据共享的需要,具有科学性和实用性.

b. 多维分类方法遵循一般分类学的标准,其开放的体系结构,可较好地适应水利科学数据共享不断发展的需要.

c. 增加了水利科学数据的时空定位维,充分反映了水利科学数据所具有的时空特性.

3.3　数据编码结构

数据编码是将数据赋予具有一定规律、易于计算机和人识别和处理的符号,并形成对应的代码表的过程.根据编码对象的特征或根据所拟订的分类方法,应采用不同的编码方法.分类体系是编码的基础,依据本文提出的多维组合分类体系,水利科学数据分类采用分段编码设计和组配技术,对组配类目内容进行编码,实现多维分类体系的组配化,形成水利科学数据分类的多维码.

本文提出的水利科学数据分类编码由 3 段 18 位数字或字母组成,为提高分类编码的可读性,各段间用分隔符‘.’隔开,如图 3 所示.其中:第 1 段 8 位数字表示数据的科学属性、获取方法和数据载体;第 2 段为区域码,由 8 位数字和字母组成;第 3 段为时间码,由 2 位数字组成.在进行编码时,若编码位数不足各段要求位数,则在已知码位后用 0 补足.

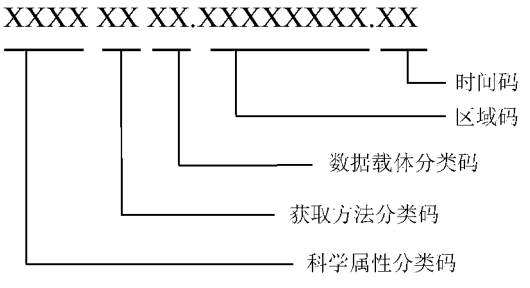

图 3　多维水利科学数据分类编码结构

Fig. 3　Code structure of multi-dimensional water science data classification system

为了进一步说明多维码的应用,现举 2 个水利科学数据集的编码予以说明.如‘海河流域永定河日降水量数据集’的编码为 11010106.ACC00006.14,其中,11010106 表示该数据为降水量数据,1101 为降水量分类码,依次,01 表示该数据为观测数据,06 表示数据以网络形式提供,ACC00006 为河流代码,14 表示日数据.又如‘北京市年水资源总量数据集”,其编码为 12030405.11000000.24,1203 为水资源总量分类码,04 表示统计数据,05 表示光盘数据,11000000 为北京市代码,24 表示年数据.

在本分类编码中,力求直接采纳已颁布的分类编码标准,如在区域码中,流域分区的编码采用 SL249—1999《中国河流名称代码》标准的河流代码表,行政分区则使用 GB/T2260—2002《中华人民共和国行政区划代码》,由于河流代码为 8 位码,而我国的行政区划代码为 6 位码,为了使编码统一,行政区划代码后用 0 补足使其也为 8 位码.

4　结　　语

水利科学数据分类编码的目的是为了有效地组织和管理各类水利科学数据资源,因此,为满足水利科学数据管理、查询、检索和共享的需要,结合水利科学数据资源多维属性的特点,提出了多维组合的水利科学数据分类体系,并制定了相应的数据编码结构.多维分类体系和编码方法的研究,对水利科学数据管理和科学研究具有促进作用,一方面,建立在统一分类标准基础上的水利科学数据分类体系,能够方便研究人员和数据使用者之间的相互交流,便于水利科学数据的交换和共享,另一方面,多维分类体系及其编码设计,提供了一种新的研究思路和应用方式,解决了水利科学数据的多维属性特点,并充分考虑和采纳现有分类体系.本文研究的水利科学数据分类与编码体系已应用于水利科学数据共享网并取得良好的效果,基本满足了水利科学数据发布、管理和共享的需要.

参考文献:

[1] 周建勤,张铎.物流信息分类编码标准体系探讨[J].物流技术与应用,2000,5(2):37-39.
[2] 孙玉荣.物料分类编码系统的研究[J].上海标准化,2002(12):35-37.
[3] 邓均华.数字图书馆与数字化分类法[J].中国图书馆学报,2001(4):76-77.
[4] 刘植婷.信息分类编码标准化综述[J].世界标准化与质量管理,2004(4):50-52.
[5] GB/T 7026—1986,标准化工作导则 信息分类编码标准的编写规定[S].
[6] GB/T 7027—2002,信息分类和编码的基本原则与方法[S].
[7] GB/T 10113—1988,分类编码通用术语[S].
[8] GB/T 13923—1992,国土基础信息数据分类与代码[S].
[9] GB/T 19317—2001,专题地图信息分类与代码[S].
[10] GB 15218—1994,地下水资源分类分级标准[S].
[11] SL 249—1999,中国河流代码[S].
[12] 刘若梅,蒋景瞳,贾云鹏,等.可持续发展信息共享标准化研究和相关标准制订[J].资源科学,2001,23(1):23-28.
[13] 路鹏,刘瑞丰,李志雄,等.地震科学数据的分级分类探讨[J].西北地震学报,2007,29(3):248-251,255.
[14] 廖顺宝,蒋林.地球系统科学数据分类体系研究[J].地理科学进展,2005,24(6):93-98.
[15] 李集明,熊安元.气象科学数据共享系统研究综述[J].应用气象学报,2004,15(增刊):1-9.
[16] 朱星明,耿庆斋.略论水利技术标准中信息共享类标准存在之问题[J].水利技术监督,2006,14(3):6-9,16.
[17] SL213—98,水利工程基础信息代码编制规定[S].
[18] SL/T200—97,水利系统政务信息编码规则与代码[S].
[19] SL261—98,中国湖泊名称代码[S].
[20] SL259—2000,中国水库名称代码[S].
[21] SL262—2000,中国水闸名称代码[S].
[22] SL263—2000,中国蓄滞洪区代码[S].
[23] SL190—96,土壤侵蚀分类分级标准[S].
[24] SL385—2007,水文数据 GIS 分类编码标准[S].
[25] 水利部国际合作与科技司.水利技术标准体系表[M].北京:中国水利水电出版社,2001.
[26] GB/T 13745—92,学科分类与代码[S].
[27] 陈骥君,曹彦荣.科学数据共享分类方法初探//[C].第四届海峡两岸 GIS 发展研讨会暨中国 GIS 协会第十届年会论文集.北京:中国地理信息系统协会,2006:419-424.
[28] 耿庆斋,朱星明.水利科学数据共享标准体系研究与构建[J].水利学报,2007,38(2):233-238.

Classification system and code structure of water science data based on multi-dimensional assembly

GENG Qing-zhai[1,2], ZHANG Xing-nan[1], ZHU Xing-ming[2]

(1. *State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing* 210098, *China*;
2. *China Institute of Water Resources and Hydropower Research, Beijing* 100044, *China*)

Abstract: In order to meet the requirements of management and sharing of water science data, the status quo of research on the classification and code of water information as well as the main problems of the existing classification system were analyzed. A multi-dimensional water science data classification system was put forward based on the characteristics of the water science data. The multi-dimensional water science data classification system consisted of scientific attributes, acquisition methods, data carriers and space-time orientation. The standard codes were designed according to the classification system, and a multi-dimensional code structure with 18 digits was divided into three sections. At present, the classification and code system has been applied in the water science data network.

Key words: water science data; data classification; classification system; data code; multi-dimensional assembly