第 20 卷第 2 期
2001 年 6 月

地 理 科 学 进 展
PROGRESS IN GEOGRAPHY

Vol 20, No. 2
June, 2001

文章编号: 1007-6301 (2001) 02-0137-09

地学数据集成的理论基础与集成体系

李 军[1,2], 庄大方[1]

(1. 中科院地理科学与资源研究所, 北京 100101; 2. 中科院遥感应用研究所, 北京 100101)

摘要: 地球空间数据 (简称地学数据) 来源的拓宽、更新手段的发展和应用领域的扩大使数据集成或集成使用的研究和实用化成为必需。简单地理解, 地学数据集成是指不同来源、不同性状数据在相同环境下的使用。地学数据是对地理现象和过程及过程时空特征认知基础上的表达, 地学数据集成的基础主要表现在: 地理现象和过程的空间和时间统一性、地学过程时空过程的连续性、地学现象和过程的层次性、地学数据认知的一致性、依赖于元数据的地学数据的透明性、数据内容和形式的相对独立性等; 在此基础上, 作者在论文中描述了基于地学知识和地理信息系统功能的地学数据集成概念模型和过程, 并对地学数据集成过程中涉及到的问题进行了说明。

关 键 词: 地学数据集成; 集成理论依据; 数据认知

中图分类号: N 941 **文献标识码**: A

1 引言

集成 (Integration) 是指通过结合分散的部分形成一个整体[1]。关于地学数据集成, 目前仍没有公认的定义。根据其侧重点归结为以下几类: ① GIS 功能观点。认为数据集成是地理信息系统的基本功能, 是原数据层经过缓冲、叠加、获取、添加等操作获得新数据集的过程[2]; ② 简单组织转化观点。认为数据集成是数据层的简单再组织, 即在同一软件环境中栅格和矢量数据之间的内部转化或在同一简单系统中把不同来源的地理数据(如地图、摄影测量数据等) 组织到一起; ③ 过程观点。认为数据集成是在一致的拓扑空间框架中地表描述的建立或使同一个地理信息系统中的不同数据集彼此之间兼容的过程; ④关联观点。认为数据集成是属性数据和空间数据的关联, 如 ESRI[3]认为数据集成是在数据表达或模型中, 空间和属性数据的内部关联; 这些观点从不同角度揭示出地学数据集成的多样性和综合性。应该说数据集成不是简单地把不同来源的地学数据合并到一起, 还应该包括普通数据集的重建模过程[4], 以提高集成的理论价值。

从逻辑上分析, 数据集成指不同来源、格式、特征的地学数据逻辑上或物理上的有机集中。有机是指数据集成时充分考虑了数据的空间、时间和属性特征, 以及数据自身及其表达的地理特征和过程的准确性[5]。

收稿日期: 2001-02; **修订日期**: 2001-04

基金项目: "九五"重中之重科技攻关项目重大自然灾害评估系统研究项目资助 (96-B02-02-02)

作者简介: 李军 (1968-), 男, 河北大名人, 博士后。主要从事地学数据基础研究及地理信息系统应用基础研究, 发表有关论文多篇。

2 地学数据集成的理论基础

从表面上看，地学数据集成只是一种技术和手段，但具体分析数据集成的内容不难发现，它涉及许多学科和多种技术。首先，数据集成是面向应用项目的，需要用到各类专业知识；其次，数据集成中空间、属性和时间等数据特征的处理与计算机、数据库、网络、数据政策等技术和学科有关系；最后，集成结果的表达与具体需求和已形成的数据规范和标准分不开。

2.1 数据集成依赖的学科与技术

地学数据集成有着深厚的理论和技术基础支撑，数据集成首先是一种数据处理技术，它与许多技术分不开，同时数据集成面向应用与数据的内涵分不开，它的表达必然要靠许多理论的支持（图 1）：

(1) 相关学科的理论与规则是地学数据集成的理论依据 地学数据集成涉及的多个学科在数据数据集成中所起的作用有一定差异。

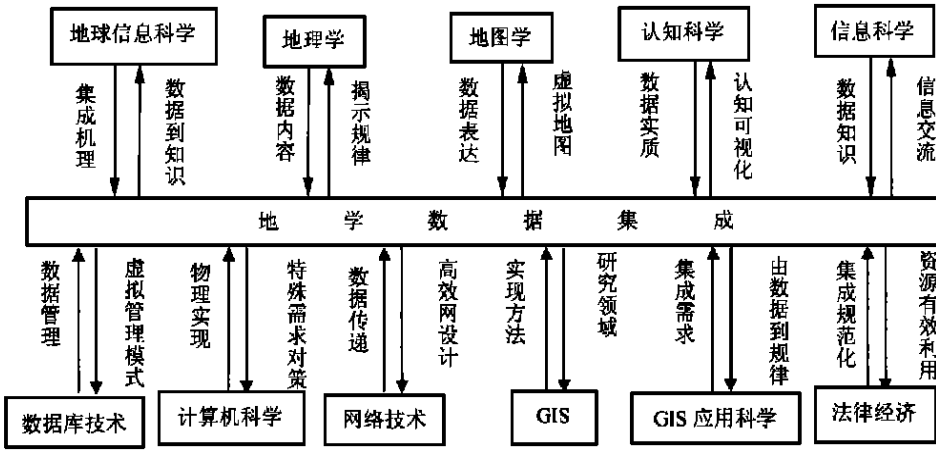

图 1 地学数据集成与相关学科和技术的关系图

Fig. 1 Relationship among geo-spatial data integration and relating subjects and technique

地球信息科学系以地球表层为研究对象域（电离层-莫霍面），以人-地相互作用关系为主题，以服务全球变化和区域可持续发展为目标，将卫星应用等多项技术为主体的高速全息数据化集成的科学体系，形成能对人流、物流、能流进行时空分析与宏观调控的战略技术系统[6]。地球信息科学是地学数据、地理信息获取、加工处理、再表达的理论依据。客观世界的地理过程经过加工形成的数据是信息的载体，通过对地理信息的认识可以形成关于地理世界知识形式的规律，进而形成可以表达的地学数据。

地理学是以研究具有空间展布特征对象和过程的经典学科[7]，从地理学的角度出发，地学数据的内容是具有空间展布规则的地学过程，这为数据表达、获取和处理提供知识规则；地图学是将地球表面具体的或抽象的过程、特征进行可视化处理表达的学科，地学数据表达仍然继承了许多地图学的特征和处理方法，虽然地图与地学数据有一定的区别[8])。地图学为集成中数据的处理和表达提供了参考方式，数据集即是一种虚拟地图[9]。

认知心理学揭示了作为主体的人对客观世界的概念化定义描述的过程[10]，所以它有助于说明地学数据的实质，从而为数据集成的各类模糊性处理提供认知角度的理论依据。

信息科学与计算机科学是地学数据物理表达处理的基础，数据集成过程即是数据、信息、知识相互转换处理的过程；信息科学为数据集成中数据信息流的处理提供了范式或原型。数据集成的最终实现是靠软件支持的计算机来完成的，数据集成中所有的概念和逻辑模型的实施离不开计算机物理基础，在计算机要处理的各种问题中时空数据的处理是有特殊困难的一种应用[11]，因而需要计算机实现处理地学数据的特殊需求对策。

(2) 数据集成需要许多技术支持，各种技术的协同应用才能保证数据集成的实现。

计算机及其它方面的新技术在数据库中已得到了广泛的应用[12]，数据库技术为地学数据集成中对数据组织、检索、更新等操作功能有独特要求提供了可行的方法。地理信息系统是以地学数据为处理对象的专业系统[13]，其功能是数据集成中必然要用到的，所以说地理信息系统是地学集成实现方法和工具。分布式地学数据集成的实现与网络研究分不开，网络技术为地学数据传递提供了可靠的模式和方法[14]。

(3) 集成应用与数据政策法规分别是集成的目标和保障，地学数据集成是为地学数据应用项目服务的，所以集成的具体需求来自于数据应用项目，但需求是千差万别的，地学数据集成即是在集成需求普遍性的基础上实现对特殊需求的满足，如在重大自然灾害集成项目中即用到了多种地学数据甚至是非地学数据的集成[15]；数据在集成处理中可以实现由数据到规律的转化，即由具体应用中数据的特征和具体数据处理中需求抽象总结成规律性的知识，以完善集成的发展。

数据法规与政策指涉及数据制作、传播、共享使用有关的法规和政策，数据共享的立法，可以减少数据集成的工作量。从客观上讲，法规政策和法律的根本目的是在某些社会集团利益得到充分保证的基础上使某些社会活动的有序化，但从实际意义上某些数据信息方面的法规政策可能阻碍了数据共享。如目前一些数据供应组织的收费标准不同类型[16]。收费政策本身对某些数据用户来说即是一种数据共享的限制。地学信息数据方面的政策和法规是地学数据集成具体实现中必然考虑的因素。

2.2　数据集成的理论依据

地学数据集成与多个学科、多种技术有关系，其理论依据也是多方面的。这里主要就地学数据自身存在、内部和外部特征来分析地学数据集成的可行性。

2.2.1　统一的空间场

地学数据是关于地球表层各类地学过程、现象及其它有空间位置需求现象过程的数据，地学数据存在的空间场是统一的，即连续地表空间。对空间的表达可以有多种坐标体系[17]，如经纬度表示的球面坐标、平面坐标等，但不论以何种形式表达，存在于地表空间的地物之间的拓扑关系是可以量度的，地物存在依赖的空间基础是相对不变的 (图 2)。空间的连续性为地学过程在地球表面的连续展布提供了基础。

2.2.2　地学过程的空间连续性

地学过程的连续性表现在空间和时间上。空间连续性表现为独立地学过程在空间上分布的非间断性和同类地学过程个体的连接特征，如：河流发育过程中对应的一条河流 (即使是一条很小的支流) 也有属于自己的流域区，并且其流域区在空间上是靠河流的河道连接起来的连续体；多个河流之间在空间上又是邻接的，如我国的黄河流域与长江流域在空

间上是邻接的。城市化的过程也具有这样的特点（图 3），城市A、B、C 在城市化过程中，在理想状态下它们对周围区的吸引力呈近圆性向外逐渐减弱，在更大范围上各城市的吸引区是相互重叠的。时间连续性表现在任意时段的地学过程状况都是整个地学进程中的一个片段，不论时间的计量单位是什么，它们之间都是连续的，如：1994 年后必然是 1995 年，1 月之后必然是 2 月等。

图 2 地球表面各大洲的分布图

Fig. 2 Continents distribution on earth surface

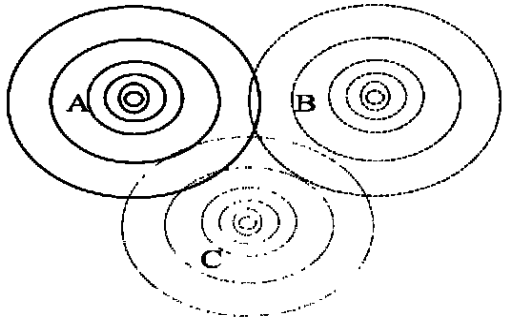

图 3 城市的空间吸引及其间的关联性

Fig. 3 Urban spatial attraction relationship

时间和空间上的连续性为地学数据特征空间上和时间上的拼接、合并、提取、拆分提供了依据。这种连续性的另一种结果是数据的时间和空间特征在一定程度上可以转换，这在遥感数据处理中已得到了应用[18]。

2.2.3 地学过程的层次等级性

地学过程的层次等级性表现为空间域和时间域上的等级性。空间等级层次性最明显的表现是地学过程在空间上的可分解性，如干流由很多二级河流组成，二级河流又由许多三级河流构成等；国家级行政区由省级行政区构成，省级行政区又由诸多地市级单元构成，然后依次有县、乡、村、组等各级行政单元组成。由此可以将要描述的地学现象按类别层次的组织形式表达，表 1 给出的我国 1∶50 万精度土地利用数据中利用类型的二级分级模式[19]。

表 1 土地利用数据分级类型

（比例尺 1∶50 万）

Tab. 1 Hierarchy system of landuse data

（scale 1∶500 000）

一级分类	二级分类	一级分类	二级分类
耕地	水田	交通用地	铁路
	旱地		公路
	其它		机场
园地	果园		港口
	桑园		其它
	茶园	水域	河流
	其它		湖泊
林地	森林		冰雪
	灌木林		水库
	其它		其它
……			

时间上的层次等级性表现为时间在度量上的可分解性，Clifford J 和 Rao A 给出了一种时间全域的描述[20]，其中，时间全域中的每个时间单元称为时间域，不同级别的时间域之间存在继承和组成关系，如：年由月组成和继承；月由日组成和继承等。

对空间和时间等级性的综合认识可以形成对地学过程的整体的级别性认知，如对地球表面的地理认知表现之一的地理意象可以分为 5 种类型[21]：综合体类、景观类、区域类、地理系统类和区域地理系统。地学过程在时间和空间上的层次等级性为地学数据综合提供了理论依据，地学数据综合的过程是主动的，它与制图概括有一定的区别和联系。

2.2.4 认知过程的一致性

地学数据的获取过程是以主体的人对客观世界认知结果的表达 (图 4)。数据生产者根据自己的经验、知识、数据要求、满足的条件等借助于数据位置和属性获取工具对客观的地理世界进行模型表达、模拟、描述、定义、解释等以获得数据的基础材料, 然后对数据材料进行规范化、标准化处理则形成地学数据。

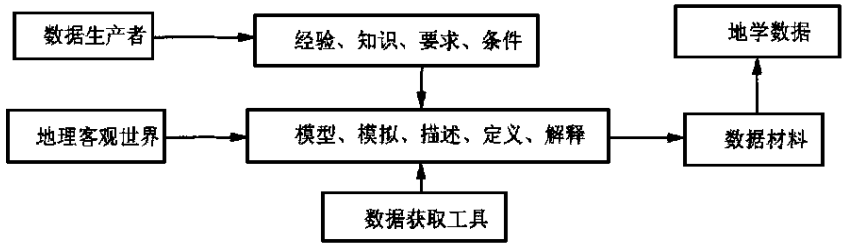

图 4　地学数据获取的认知过程图

Fig. 4　Cognition process of Geo-spatial data Capturing

影响地学过程空间认知和表达的要素包括: 内部因素, 如个人的认识能力, 主体的知识背景, 感觉限制和态度等; 外部因素, 如地学过程信息获取工具、表达的媒体等。认知过程的统一性主要表现在:

(1) 数据表达的一致性。当把地学客体加工形成有抽象意义的地学数据时, 在数据表达上表现出了很多共性。矢量数据中, 数据的空间部分可表现为点、线、面、体等; 属性部分则表现为与空间部分可以通过一定形式关联起来的数字、字符等内容。并且数据的空间部分和属性是完全统一的, 即离开了空间形状属性便成了没有空间容器的纯属性, 离开了属性的空间形状也将成为没有地学过程意义的纯几何图形。因为这种一致性, 地学数据才可以被处理。

(2) 数据体系的一致性。基于地学过程的客观性和地学认知过程的科学性, 地学数据表现出内容体系的一致性。具体的分类、分级的数量值可以有差异, 但这种分类分级的体系是共同的, 如我国公路级别划分中有国道、省道、县道、乡村道路等分级体系, 也有一级公路、二级公路等分级体系, 虽然具体分类的名称和量级可能不同, 但各类各级之间有一定的可比性, 这对数据集成中属性的一致化、语义识别有很大意义。

(3) 相同层次上内容的一致性。出于对数据的客观需要和地学过程认知的级别层次性, 数据在相同或近似层次上 (如数据精度) 内容有一致性。表现在数据形式上, 在低精度 (小比例尺) 数据中城市都以点表示其空间位置, 而在高精度 (大比例尺) 数据中城市是以面状要素表示其空间位置, 同样在低精度城市数据中表示城市的整体性质, 如城市的总人口、生产总值等, 而在高精度城市数据中, 则要表现城市中各功能区的属性, 如某个城区的人口、文化素养等。这对数据集成中多比例尺数据的处理有一定帮助意义。

2.2.5 依赖于元数据的地学数据透明性

地学数据的透明性是在数据集成前用户可以对要集成的对象数据有逐级 (可以是数据集层次, 也可以是数据特征层次) 了解, 即数据从形式到内容对用户来说都是透明的。这种透明性主要靠地学元数据实现[22]。地学数据集成是对作为空间位置、属性和时间整体的地学过程或地学过程片段的综合处理, 数据的透明性为数据集成的预处理和实际的内容集

成奠定了基础。

2.2.6 数据形式和内容的相对独立性

数据形式指诸如数据存贮格式、存在介质、表达方式等一类的外部特征，内容指地学数据的空间位置、属性、时间、精度等一类的特征。相对独立性表现在形式的变动时内容保持原来的特征或者只有可控制、可描述的微小变动，而当数据内容发生变化时其形式可以保持完全不变。相对独立的根本原因在于数据形式是数据内容的一种载体，一种外在表现，跟数据内容没有必然的因果关系。这就保证了可以对数据进行诸如：格式转换、投影变换、网络传输、提取、多数据集合并等集成操作而不改变以数据内涵；也可以对数据记录进行删除、添加、合并、属性归一化等内容处理而保持数据外部形式的原有一致性。

3 数据集成系统结构

数据用户对数据要求的多样性决定了地学数据集成目标的复杂性，这里从地学数据集成的作用机理和集成中数据流的运行状况及实际的地学数据集成应用角度，给出了地学数据集成系统结构（图 5）。

地学数据集成系统包括：网络支撑的集成系统界面、数据检索查询功能块、数据集成块、地学规则功能块、数据预处理模块、元数据功能块和数据质量控制功能块等部分组成。各部分在系统中有自己独特的作用。

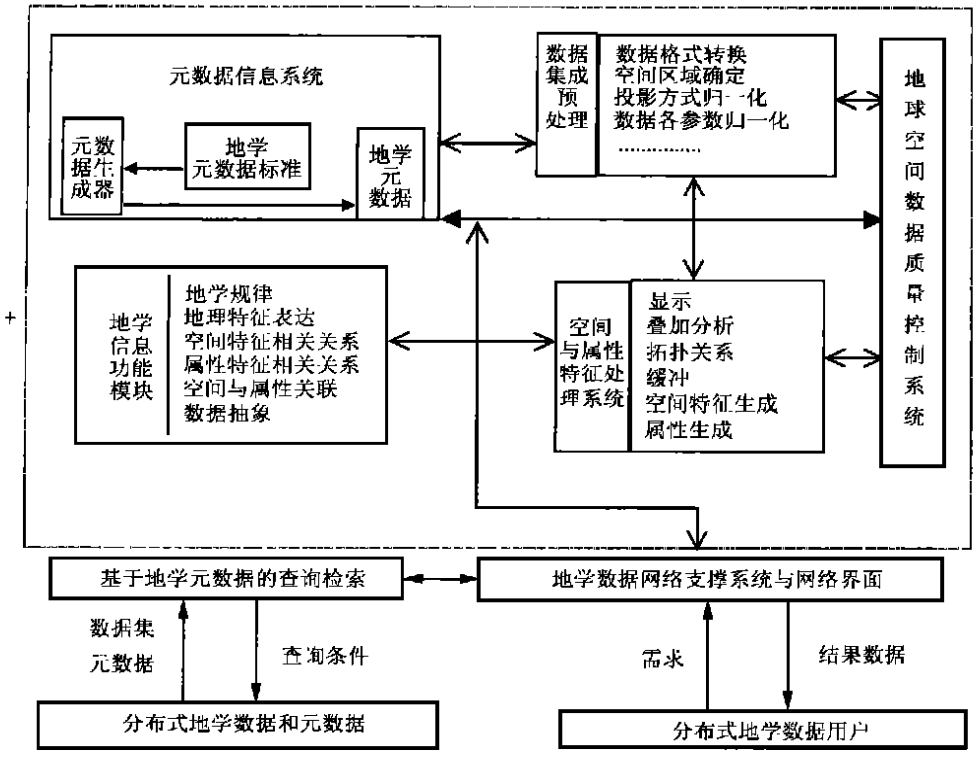

图 5 面向应用目标的地学数据集成系统结构图

Fig. 5 Diagram of application objected geo-spatial data integration systems

(1) 网络支撑的集成系统界面。该部分的作用是: 将集成系统中的各模块贯穿起来; 以服务器形式处理并对用户提出的问题给予反映; 控制系统功能流等。

(2) 数据检索查询功能块。根据用户需求通过元数据在分布式地学数据库中寻找满足条件的数据集, 并将符合条件的数据集的元数据内容反馈给系统, 为系统的下一步操作提供依据。

(3) 数据集成模块。是集成系统的核心模块, 它以地理信息系统的功能为基础并增加了一些特殊功能。它执行对数据集空间属性及其相关关系的具体处理, 以形成符合条件逻辑或物理数据集 (库)。

(4) 地学规则功能块。该功能块相当于一个地学专家知识系统, 提供一系列地学数据相关规则。它服务于数据质量检测、评价和控制、集成中具体数据实体特征的处理和新形成数据集中某些特征的处理。

(5) 数据预处理模块。根据系统提出的要求将所需要集成的数据逻辑或物理的集中到一起; 对集成数据的外部特征 (数据格式、投影形式等) 一致性和数据自身特征 (不同数据集中对应特征空间位置、属性、数据的时间等) 的一致性进行检查并做相应的处理[23]; 完成对数据的分割操作等。

(6) 元数据功能块。提供数据集元数据模式和生成功能; 记录系统处理过程中关于系统和数据的动态信息以辅助系统实现其它操作。

(7) 数据质量控制功能块。该功能块的作用是评价、检测数据质量; 通过系统控制集成处理中各种影响数据质量的各类参数的设定; 利用数据质量标准对集成结果进行评价等。

地学数据集成系统的各功能块是针对系统中可能出现的各种问题设置的, 在某一具体集成应用项目中可能只用到系统的某些功能。系统各模块是相互关联的整体, 网络支撑是系统的整体平台, 数据检索查询、数据集成、地学规则、数据预处理等是集成中具体处理问题的依据与实现模块, 元数据是数据正常处理及处理后保证数据质量的基础, 数据质量控制贯穿于整个集成系统。

4　结论

地学数据集成是信息共享的基础, 对地学数据集成理论依据和模式的研究在于为数据集成模式建立、涉及技术研究奠定基础。地学数据集成与地学数据表达地学内容的特征分不开, 所以在地学数据集成中必须考虑数据的统一空间场、等级性、认知的一致性等特征, 在此基础上建立起基于数据内容的数据集成模式。利用这种模式, 可以更有效处理数据集成处理过程中的相关问题。根据这些集成依据, 作者在参加一个卫星影象地面应用处理系统设计中提出了综合数据库系统模型, 并在实践中得到了很好的应用[24]。

参考文献:

[1] 李军, 费川云. 地球空间数据集成研究概况[J]. 地理科学进展, 2000, **19**(3): 203-211.

[2] Shepherd I D H. Information integration and GIS in Geographical Information Systems: Principles and Application (Vol 2)[A]. In: Maguire D J, Goodchild M F, Rhind D W (ed). Longman[C]. London, 1991. 337-360.

[3] ESRI GIS Data management[C]. ESRI, 1990.

[4] David Martin and Ian Bracken. The integration and socioeconomic and physical resource data for applied land

management information systems[J]. *Applied Geography*, 1993: 45-53.
[5] 李军, 周成虎. 地学数据集成多尺度问题基础研究[J]. 地球科学进展, **15**(1): 48-52.
[6] 陈述彭, 何建邦, 承继成. 地理信息系统的基础研究——地球信息科学[J]. 地球信息- 科学·技术·产业, 1997 (3): 11-20.
[7] 牛文元. 理论地理学[M]. 北京: 商务印书馆, 1992. 7-33.
[8] Ruas A, Lagrange J P. Data and knowledge modelling for generalization[A]. In: Jean-Claude muller, Jean - Philippe lagrange, rebert Weibel (ED) GIS and Generalization methodology and practice: GisData 1[C]. Taylor & Francis, 1995. 73-90.
[9] Kraak M F, Ormeling F J. Cartography Visulization of Spatial Data[M]. Longman, 1996. 47-50.
[10] 王苏, 汪安圣. 认知心理学[M]. 北京: 北京大学出版社, 1992. 240-305.
[11] 林众, 冯瑞琴. 计算机与智力心理学[M]. 杭州: 浙江人民出版社, 1996. 11, 116-172.
[12] 李昭原. 数据库技术新进展[M]. 北京: 清华大学出版社, 1997. 1-27.
[13] 周成虎. 地理信息系统透视[J]. 地理学报, 1995, **50**(增刊): 27-35.
[14] 张健挺. 地理信息网络共享的研究和应用进展[J]. 地理科学进展, 1998, **17**(4): 73-78.
[15] 池天河. 重大自然灾害遥感监测与评估集成系统[M]. 北京: 中国科学技术出版社, 1995. 67-76.
[16] IDPFEO. Toward an Integrated International Data Policy Framework for Earth Observations[a]. AWorkshop Report, The International Space University, Strasbourg, France, January, 1997.
[17] 吴忠性. 在电子计算机辅助制图情况下地图投影变换的研究[A]. 见: 地图投影与地图学论文集[C]. 北京: 中国地图出版社, 1993. 20-41.
[18] 陈崇成. 基于空间信息集成技术的城市环境时空分析研究[A]. 中国科学院地理研究所博士学位论文[C], 2000.
[19] 马霭乃. 地学编码模型[A]. 见: 全国性资源与环境信息系统研究[C]. 北京: 测绘出版社, 1991. 122-126.
[20] Clifford J, Rao A. A Simple, General Structure for Temporal Domain[M]. In: Temporal Aspects in Information Sysmem, Elservier Science Publisher, 1988.
[21] 鲁学军. 地理学认知理论与地理专家决策模型研究[A]. 见: 北京大学博士研究生学位论文[C]. 1996. 26-44.
[22] 李军, 周成虎. 地球空间数据元数据标准初探[J]. 地理科学进展, 1998, **17**(4): 55-63.
[23] 宋关福, 钟耳顺, 刘纪远 等. 多源空间数据无缝集成研究[J]. 地理科学进展, 2000, **19**(2).
[24] 李军, 刘高焕, 迟耀斌 等. 大型遥感图像处理系统中综合数据库设计及应用[J]. 遥感学报, 2001(1): 41-45.

Theories and Systems of Geo-spatial Data Integration

LI Jun[1,2], ZHUANG Da-fang[1]

(1. Institute of Geography Sciences and Resources Research, CAS, Beijing 100101 China;
2. Institute of Remote Sensing Application, CAS, Beijing 100101 China)

Abstract: Following the extension of capturing method for geo-spatial data, development of database updating technology, and application of geo-spatial, Geo-spatial data integration is found more and more necessary than ever before. Simply saying, geo-spatial data integration is the processes in which geo-spatial data captured from different sources and created with internal and external characteristic can be used within one uniform platform. The paper describes why geo-spatial data can be integrated and how to integrate.

In the first part of the paper, the authors focus on the theory foundations of geo-

spatial data integration, such as: (1) spatial and temporal uniform of geo-processes and geo-phenomena that are considered as the content of geo-spatial represented. (2) Temporal and spatial continuity of geo-processes and phenomena, accordingly, coming to the result that different part of one geo-spatial feature can be joined together in an integrated dataset. (3) Hierarchy of geo-processes and phenomena make possible the data generalization in data integration. (4) The similar cognition processes in different geo-spatial data capturing processes make possible the comparison of same spatial feature in different dataset. (5) Based on geo-spatial metadata, how to make the content and format of geo-spatial knowable to data users. (6) The independence of contents of geo-spatial data on it's format.

Based on the theory foundation, at the final part of the paper, the authors describe a general geo-spatial data integration system which includes geo-spatial metadata information system, geo-information processing model, network based data user interface, distributed geo-spatial database, and geo-export knowledge system.

Key words: Geo-spatial data integration; Data cognition; Theory of data integration