# 遥感影像样本大数据建库与应用方法①

程 滔

(国家基础地理信息中心, 北京 100830)

**摘 要**: 研究一种基于关系型数据库与分布式文件系统融合的遥感影像解译样本大数据建库方法. 解析了数据库建设过程, 讨论了建库关键技术问题与解决方法; 在建成样本数据库基础上, 研究了样本数据应用方法, 分析了几种应用实例, 探索了样本数据应用模式; 以选取的研究区域内大规模样本数据为对象, 验证了数据库建设中数据处理过程及效率, 同时, 以地理国情普查分类体系中地表覆盖 10 个一级类为例, 对研究区域各类别样本空间分布与密度等特征进行了分析. 结果表明: 利用关系型数据库与分布式文件系统融合的方法对样本大数据进行建库与管理, 对样本数据的检索、分析及推广应用, 具有很好的效能和适用性.

**关键词**: 遥感影像样本; 大数据; 数据库; 分布式文件系统; 应用

## Database Construction and Its Application of Sample Big Data Collected in Remote Sensing Image

CHENG Tao

(National Geomatics Center of China, Beijing 100830, China)

**Abstract**: A database construction method which is based on the integration of relational database and distributed file system is researched for a large sample data base for interpretation of remote sensing images. It analyzes the database's construction process, and discusses the key technical problems and solution method. Based on the sample database, it studies the application method of sample data, analyzes some kinds of using cases and explores the application mode of sample data. Individual region's massive sample data are selected for verifying the method and its efficiency. At the same time, it takes 10 first-level classes which are defined in the land cover classification system for example, to analyze the spatial distribution and density characteristics of all kinds of sample data. The results show that the method of database construction and management which is based on the integration of relational database and distributed file system is very effective and applicative for sample data's searching, analyzing and promoted application.

**Key words**: remote sensing image sample; big data; database; distributed file system; application

第一次全国地理国情普查样本数据库建设是地理国情普查成果数据库建设的任务之一, 可为从事遥感影像解译的研究人员与工程技术人员提供丰富的解译标志信息, 提高地表覆盖分类、土地利用分类等精度, 从而提高研究成果质量[1]. 第一次全国地理国情普查在全国范围内采集的遥感影像解译样本点数量达到 300 多万个, 数据文件量达到 1250 多万个, 并将在后续地理国情监测中不断积累递增.

为了提高样本数据检索、分析及推广应用效率, 促进应用服务, 需要对这些数据进行科学存储和管理[2,3]. 利用数据库对样本数据进行管理, 是一种可靠的方法, 数据存储的逻辑性强, 能够提高数据检索效率. 成熟的关系型数据库技术采用结构化的语言(Structured Query Language, 缩写 SQL), 用二维表结构分行、列对数据进行存储, 调用数据时遵循固定的请求格式, 甲骨文(Oracle)在 20 世纪 70 年代率先推出这项技术, 该技术也是目前应用最为广泛的数据库技术[4].

然而, 随着云计算、互联网等技术的发展, 文档、

① 基金项目:国家自然科学基金(41301464);国家基础地理信息中心科技创新发展基金课题（2017-KJ-G01）
收稿时间:2016-08-15;收到修改稿时间:2016-09-18 [doi:10.15888/j.cnki.csa.005723]

图片、图像、视频、文本、XML 等非结构化、半结构化数据增长迅速, 关系型数据库虽支持二进制大对象(BLOB), 能将数据直接入库存储, 但未提供对这类复杂数据类型的快速存储、访问方法[5]; 所以这类数据的存储, 已不方便用关系型数据库二维逻辑表来表现, 需要增大数据库的开发工作量才能满足应用需求. 因此, 大数据管理方法与计算处理能力在极大提升的同时, 也面临一些挑战[6,7].

地理国情普查样本数据文件数量庞大, 且包含 ACCESS、JPG、TIFF、TFW、XML 等多种数据格式, 从数据模型角度划分, ACCESS 属于结构化数据, JPG、TIFF、TFW 属于非结构化数据, XML 属于半结构化数据.

针对地理国情普查样本数据特点, 本文研究一种基于关系型数据库与分布式文件系统融合的样本大数据建库方法, 将各类模型的数据分别存储在不同的物理位置, 并对结构化数据进行空间化处理, 增强数据的检索性能与可视化体验, 以满足大数据建库与后续应用的需求. 首先分析建库过程与关键技术, 解决大数据、批量处理过程中的技术问题; 然后在完成样本数据库建设的基础上, 研究探索样本数据的应用方法与模式; 最后通过选取大规模样本数据集, 结合空间分析, 对研究方法进行验证.

## 1　研究方法

### 1.1 数据分析

地理国情普查样本数据的原始数据由地面照片、遥感影像实例以及样本信息描述数据库三部分组成. 其中, 地面照片采用 JPG 格式; 遥感影像实例采用 TIFF 格式; 影像坐标信息采用 TFW 文档格式; 影像投影信息采用 XML 格式; 样本信息描述数据库采用 ACCESS 数据库, 由记录地面照片属性信息的 PHOTO 数据表(包括照片的标识符、照片文件名、拍摄时间、拍摄点经度、拍摄点纬度等 19 项属性)、记录遥感影像实例属性信息的 SMPIMG 数据表(包括遥感影像实例标识符、遥感影像实例文件名、影像类型、影像分辨率、影像拍摄时间等 14 项属性)、以及反映地面照片和遥感影像实例对应关系的 PHOTO_IMG 关系表(包括地面照片的标识符、遥感影像实例标识符等 5 项属性)三个表格构成, 表格数据类型包括 Text、Date、Double、Float、Short Integer[8].

为了便于地理国情普查样本数据的展示、检索、分析, 在原始数据经过入库质量检查合格的基础上, 需要利用原始数据记录的空间位置信息(地面照片拍摄点经度、拍摄点纬度, 或者根据对应遥感影像实例四个角点经纬度计算出的中心点坐标), 生成样本点位矢量数据[9], 该衍生数据为点状图形数据, 其属性信息包括地面照片所有属性信息, 并添加了要素唯一标识码属性.

这种结构化、非结构化、半结构化数据在入库前均以文件形式存储, 并组成了地理国情普查样本数据的完整数据模型.

### 1.2 数据库建设方法

地理国情普查样本大数据建库过程是数据库建设与管理的核心, 原始数据在经过数据整理、入库检查、问题解决、重新整理等处理过程后, 需要进行属性结构调整、表格数据空间化等处理, 经入库质量检查合格后, 进行数据入库操作.

在数据入库过程中, 地理国情普查项目采用 Oracle 数据库技术, 在数据库设计时, 分别按照表格数据、矢量数据、文档数据这几种形式作为数据存储结构. 结构化数据直接存储在 Oracle 数据库表中; 空间化后的矢量数据存储在 Oracle Spatial 中, 具体采用 SDO_Geometry 字段进行物理存储, 属性信息存储在相应的属性字段中; 非结构化、半结构化的文档数据存储在分布式文件系统中[10].

这种基于关系型数据库与分布式文件系统融合的样本大数据建库方法的处理流程如图 1 所示.

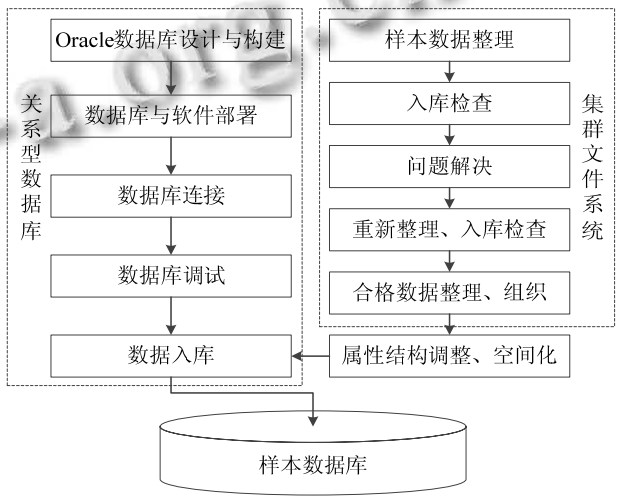



图 1　样本数据库建设流程图

### 1.3 关键技术分析

根据本文数据库建设方法, 在样本数据库建设的

整个流程中, 关键技术主要表现在以下两个方面:

(1) 大数据整理与存储

全国地理国情普查样本数据具有文件数量庞大、数据总量大、各模型数据格式各异的特点, 而数据库建设对大数据整理的要求是存储结构规范、逻辑关系严密、结构化整理. 在这种形势下, 为了利于大规模数据的更新与维护, 在数据整理与存储过程中, 可按照全国行政区划或测区(一般为县级或地市州级行政单位), 逐级整理清晰.

在分布式文件系统中, 对于一个行政区划或测区内的所有样本数据, 保持固定的耦合存储结构(如表 1 所示); 各行政区划或测区样本数据集之间并行排列; 采用县级或地市州级、省级、国家级逐级往上集中存储. 这样的存储方式有利于样本数据的快速检索、修改、移动、删除等操作.

在关系型数据库中, 利用各样本数据的照片文件名字段、遥感影像实例文件名字段, 存储其行政区划或测区级的相对路径, 从而, 在数据库管理系统调用样本数据时, 便可读取此记录寻址到样本数据在分布式文件系统中的物理位置.

表 1 样本数据目录组织方式

| 目录结构 | 说明 | 示例 |
|---|---|---|
| 行政区划或测区 | 行政区划或测区 | 420102 |
| SMPDATA.mdb | 数据库文件 | SMPDATA.mdb |
| PHOTO | 地面照片目录 | PHOTO |
| XXXXX.jpg | 地面照片文件 1 | PH20130518154133113230631320030 1.jpg |
| *****.jpg | 地面照片文件 n | ********** .jpg |
| SMPIMG | 遥感影像实例目录 | SMPIMG |
| XXXXX.tif | 遥感影像实例文件 1 | RS20130518154133113230631320030 1.tif |
| XXXXX.tfw | 遥感影像实例文件 1 | RS20130518154133113230631320030 1.tfw |
| XXXXX.xml | 遥感影像实例文件 1 | RS20130518154133113230631320030 1.xml |
| *****.tif | 遥感影像实例文件 n | **********.tif |
| *****.tfw | 遥感影像实例文件 n | **********.tfw |
| *****.xml | 遥感影像实例文件 n | **********.xml |

(2) 大数据入库检查方法与问题解决

第一次全国地理国情普查项目是国家重大专项项目, 成果数据库的建设必须符合工程项目的标准和要求, 因此, 对数据库的质量要求较高. 为保证数据库质量, 满足推广应用需求, 样本数据在整理规范的基础上, 需进行入库检查, 并对检查发现的问题进行有效解决, 合格后才能入库.

面对样本文件数量庞大的现实状况, 本研究利用一种从宏观到微观的综合质量检查方法[11], 采用大数据量批处理的模式, 结合全国行政区划地图, 全面实现样本数据各项内容的入库质量检查. 从宏观整体角度, 检查样本数据组织正确性与完整性、遥感影像实例与地面照片的匹配性与冗余性、数据表定义与属性项定义正确性等内容. 从微观具体角度, 检查各样本点数据的完整性与有效性、数学基础与空间位置正确性、文件命名及格式正确性、属性数据正确性等内容. 每一个样本点数据检查均保存一条检查结果记录.

依据检查结果记录, 对影响入库、应用的问题进行有效解决, 主要包括: 遥感影像实例数学基础错误(包括坐标系统错误、中央经线错误等)、样本信息描述数据库表内容为空、遥感影像实例四角点坐标错误、影像投影信息文件 XML 记录的内容错误(为规定之外的内容)、影像坐标信息文件 TFW 记录的内容错误、地面照片无对应遥感影像实例、遥感影像实例无对应地面照片、个别行政区划或测区内数据缺漏等问题.

经过入库检查与问题解决, 形成最终的符合数据库建设要求的样本数据.

## 2 应用方法探索

地理国情普查样本大数据建库的目的是提供应用服务, 利用大数据计算与分析, 可以挖掘大量有价值的信息[12]. 本文对样本数据应用方法的研究探索, 分为直接应用与衍生应用两个层次. 直接应用是从样本数据库直接检索、获取样本基本信息, 为遥感影像解

译提供解译标志信息; 衍生应用是在基本信息的基础上, 利用空间分析方法, 得出一些规律性的特征信息.

对样本数据库的检索, 检索条件可以是多样性的, 可以根据地表覆盖类型(一级类、二级类、三级类)、空间范围(经纬度范围、行政区划范围、大区划范围(如华东、华南、华中、华北、西北、西南、东北)、主题功能区范围等)、时间段(地面照片的拍摄时间、遥感影像的拍摄时间)等, 以及这些检索条件的多条件检索.

(1) 反映研究区域地表覆盖类型及地面实地地物特征的应用

在一些遥感影像解译工作中, 会存在通过内业解译无法准确判读地表覆盖类型的情况, 在没有外业工作环节的情况下, 可以利用样本数据库, 检索研究区域空间范围内的样本数据, 通过区域内分布的样本点基本信息, 辅助遥感影像解译工作.

(2) 反映相似地理环境区域的地表覆盖类型特征的应用

利用样本数据, 可在邻近区域或相似地理环境区域(这些研究区域外业工作难以到达或限制到达, 或未计划开展外业工作), 通过同类地物光谱、纹理比对以及地理相关分析等方法, 开展遥感影像解译.

并且, 可以利用检索出的样本点对应的遥感影像实例的光谱、纹理、形状等特征, 作为地表覆盖监督分类的先验知识.

(3) 反映样本数据空间分布与密度特征的应用

全国行政区划单位分为省级、地级、县级、乡级等, 地理国情普查样本数据一般按照县级或地市州级行政单位进行整理与存储.

因此, 利用数据库中的样本点位矢量数据, 以及全国行政区划范围矢量数据, 通过空间叠置分析与统计计算, 可获取到各级行政区划范围内、各地表覆盖类型样本数据的空间分布与密度特征. 这一特征也可以反映研究区域内的地物多样性特征, 并在一定程度上间接反映研究区域内的交通通达情况.

(4) 反映同一地表覆盖类型在全国不同区域、同一季节形态特征的应用

我国地域广阔, 同一地表覆盖类型在不同的区域, 可能会表现出不同的特征, 利用样本数据库, 检索某一类地表覆盖类型(例如阔叶乔木林), 与全国典型区域矢量数据进行空间叠置分析, 便可获取同一地表覆盖类型在全国不同区域、同一季节形态特征.

(5) 反映同一地表覆盖类型在相同区域、不同季节形态特征以及影像特征的应用

地理国情普查使用的遥感影像数据的获取季节和时间不尽相同, 样本数据在采集过程中, 地面照片的拍摄季节和时间也不尽相同, 从样本数据库中检索出这些信息, 便可获取同一地表覆盖类型(例如阔叶乔木林)在相同区域、不同季节的形态特征以及影像特征.

(6) 与地形、地貌等特征相关的专题分析应用

我国地形、地貌特征丰富, 利用地形、地貌矢量数据, 与样本数据库中某一类地表覆盖类型(例如针叶乔木林、针叶灌木林)进行空间叠置分析, 可以获取该地表覆盖类型在不同地形、地貌区域的表现特征. 同样地, 通过不同地域样本记录的地表覆盖类型的种类分析, 也可在一定程度上反映不同地域地物多样性特征.

## 3 结果与分析

以湖北省地理国情普查采集的样本数据为研究实例, 验证数据库建设中关键数据处理过程及效率, 并对部分应用成果进行分析讨论.

### 3.1 研究区概况

根据《中华人民共和国行政区划简册 2015》[13], 湖北省面积约 19 万 km2, 范围内县级行政区划 103 个, 人口合计 6165 万人.

湖北省位于中国中部偏南、长江中游, 空间位置介于北纬 29°05′至 33°20′, 东经 108°21′至 116°07′, 地形地势大致为东、西、北三面环山, 中间低平, 略呈向南敞开的不完整盆地, 在全省总面积中, 山地占 56%, 丘陵占 24%, 平原湖区占 20%. 全省水资源、土地资源、生物资源、矿产资源丰富, 地表覆盖类型呈多样性特征.

### 3.2 数据处理主要过程

湖北省采集的样本点数量为 11.3 万个, 在分布式文件系统中, 按照 100 个测区对数据进行整理与存储(部分县级行政区划数据进行了合并).

样本数据入库检查时, 为了进一步提高检查效率, 将数据按照测区分为 5 组, 并发进行检查; 样本点位矢量数据以及样本信息描述数据库的表格数据, 经质量检查后, 同时录入至 Oracle 数据库. 这两项处理过程的效率如表 2 所示, 计算机配置为 64 位 Window 7 操作系统、8GB 内存.

表 2 样本数据入库检查与入库的效率

| | |
|---|---|
| 样本点数量/万个 | 11.3 |
| 数据文件量/万个 | 45.3 |
| 检查所需时间/分钟 | 98 |
| 检查占用内存/MB | 60～800 |
| 矢量与表格数据入库所需时间/分钟 | 30 |

从表 2 可以看出, 样本数据入库检查与入库的效率能够满足数据库建设流程中对大数据检查与入库的进度要求.

### 3.3 应用成果分析

地理国情普查内容体系中, 地表覆盖定义了 10 个一级类, 分别为耕地、园地、林地、草地、房屋建筑(区)、道路、构筑物、人工堆掘地、荒漠与裸露地表、水域, 并定义了 87 个三级类[14]. 基于湖北省地理国情普查采集的样本数据, 利用本文提出的应用方法, 得出了一些应用成果, 这里对部分应用成果进行分析.

(1) 样本数据空间分布与密度特征

研究区样本数据的空间分布如图 2 所示, 经统计分析可知: 湖北省范围内的地表覆盖类型涵盖了 10 个一级类, 81 个三级类.

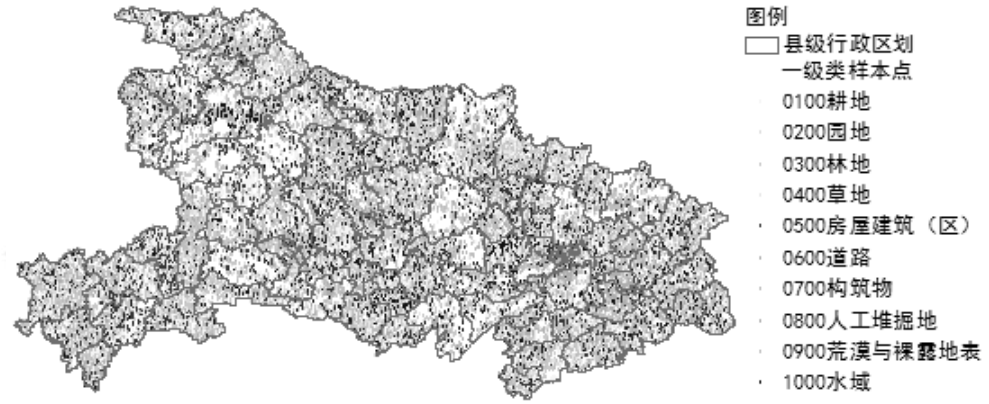



图 2 样本数据空间分布特征

以 10 个一级类为例, 各地表覆盖类型样本点数量统计直方图如图 3 所示.

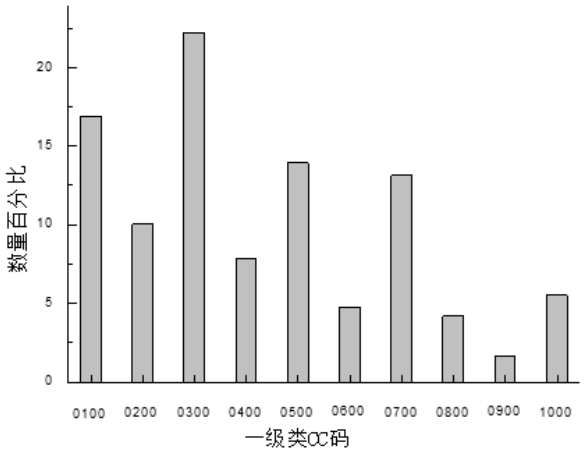



图 3 各地表覆盖类型样本点数量统计直方图

从图 3 可以看出, 湖北省范围内, 林地类型的样本数据采集是最多的, 占样本数据总量的 22.20%, 空间分布也较广, 所有县级行政区划均有分布; 其次是耕地, 占 16.87%.

自然地表覆盖类型的三级类, 是遥感影像分类的难点, 也是反映生态环境及气候变化特征的主要类型, 湖北省地理国情普查采集的这些类型的样本数据比较丰富, 为遥感影像分类积累了宝贵的资源.

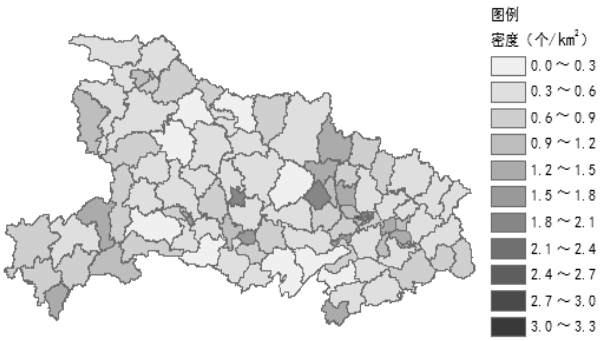



图 4 各县级行政区划样本点密度特征

利用样本数据的空间分布数据与各县级行政区划面积数据, 可以统计得出各县级行政区划样本点密度特征, 如图 4 所示.

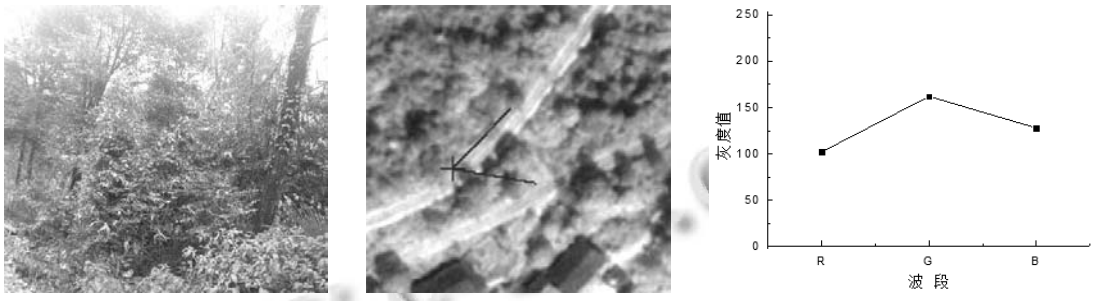
县级行政区划样本点密度区间值为[0.14～3.09], 密度特征在一定程度上也反映了各县级行政区划内地物多样性特征.

(2) 地表覆盖类型在地面照片与同季相遥感影像

上的形态及光谱特征

样本数据在采集过程中, 地面照片按照外业工作规划, 有计划地拍摄, 其拍摄季相与遥感影像实例的拍摄季相一般不同. 而经过长期的数据积累, 拍摄季相会不断丰富, 样本数据库中将能积累出大量的两者季相相同的样本数据, 利用这些数据, 可以对比得出地物光谱的区域、季相特征, 为遥感影像的自动分类提供有力的辅助信息.

图 5 为阔叶乔木林样本数据, 地面照片拍摄时间为 2014 年 10 月 28 日, 遥感影像实例的拍摄时间为 2013 年 10 月 14 日, 两者季相一致, 在正射纠正后的 8bit Pléiade 卫星遥感影像(R、G、B 三波段)上的波谱特征曲线如图 5(c).




(a)地面照片　(b)遥感影像实例　(c)波谱特征曲线

图 5　阔叶乔木林样本数据

可以看出: 研究区内的阔叶乔木在 10 月份呈现生长茂盛的形态, 连片生长的阔叶林在遥感影像上纹理比较均匀、平滑, 在红波段上表现出了强吸收特征.

## 4　结语

(1) 大数据存储、管理、分析与信息挖掘是当下众多领域研究的热点, 能够产生巨大的经济价值和社会影响力, 而高质量的数据和有效的数据管理是大数据产生服务价值的重要前提. 本文研究的基于关系型数据库与分布式文件系统融合的样本大数据建库方法, 解决了不同模型数据的存储、管理与数据调用问题, 通过实例验证与分析, 表明该方法能够保证入库数据的有效性、可用性以及数据库的质量, 有利于样本数据的合理科学管理与推广应用.

(2) 本文探索性研究的样本数据应用方法是样本数据应用范围的一部分, 样本数据在应用服务过程中, 数据时相与数据量还将不断丰富和积累, 应用服务的范围也会不断拓展和丰富, 能够产生更大的服务价值.

## 参考文献

1 国务院第一次全国地理国情普查领导小组办公室.地理国情普查数据采集技术方法.北京:测绘出版社,2013.

2 刘露.全球海量遥感影像数据的分布式管理技术研究[硕士学位论文].长沙:国防科学技术大学,2007.

3 韩晶.大数据服务若干关键技术研究[博士学位论文].北京:北京邮电大学,2013.

4 Price J.精通 Oracle Database 12c SQL & PL/SQL 编程(第 3 版).北京:清华大学出版社,2014.

5 黄飞鹏.海量遥感影像管理系统的设计与实现[硕士学位论文].上海:华东师范大学,2011.

6 孟小峰,慈祥.大数据管理:概念、技术与挑战.计算机研究与发展,2013,50(1):146–169.

7 刘智慧,张泉灵.大数据技术研究综述.浙江大学学报(工学版),2014,48(6):957–972.

8 程滔,袁如金,高志宏,高崟,史晓明.遥感影像解译样本数据一体化整理方法.地理信息世界,2014,21(5):96–100.

9 国务院第一次全国地理国情普查领导小组办公室.地理国情普查数据库建设技术方法.北京:测绘出版社,2015.

10 周江,王伟平,孟丹,马灿,古晓艳,蒋杰.面向大数据分析的分布式文件系统关键技术.计算机研究与发展,2014,51(2):382–394.

11 程滔.地理国情普查样本数据入库质量检查方法研究.测绘通报,2015,10(10):103–106.

12 李清泉,李德仁.大数据 GIS.武汉大学学报(信息科学版),2014,39(6):641–644,646.

13 中华人民共和国民政部.中华人民共和国行政区划简册 2015.北京:中国地图出版社,2015.

14 国务院第一次全国地理国情普查领导小组办公室.地理国情普查内容与指标.北京:测绘出版社,2013.