doi:10. 3969/j.issn. 1672-4623.2023.01.014

面向时空地理数据库的在线异构整合研究

王　慧[1]，蔡志刚[1]，唐　伟[1]，任　易[1]，吴承红[2]
（1. 浙江省测绘科学技术研究院，浙江 杭州 310012；2. 浙江省水利河口研究院，浙江 杭州 310020）

摘　要：时空地理数据库的异构整合是保障新型测绘数据服务能力的关键技术问题。在分析数据库异构整合理论和总结现有技术方法不足的基础上，提出了基于语义映射和ETL理论的科学、标准化异构整合技术流程；并基于FME，利用数据库框架XML文件、图层语义映射表等外部文件方式，开发了逻辑严谨、扩展灵活、自适应强的自动化处理工具，形成了B/S模式的在线异构整合解决方案。经项目广泛应用验证，该成果对实现时空地理数据库整合需求的快速响应以及提升数据服务的效率与质量具有重要支撑作用。

关键词：异构整合；FME；在线；ETL；时空地理数据库
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2023）01-0069-05

Study on Online Heterogeneous Integration of Spatio-temporal Geodatabase

WANG Hui[1], CAI Zhigang[1], TANG Wei[1], REN Yi[1], WU Chenghong[2]
(1. Surveying and Mapping Science Institute of Zhejiang, Hangzhou 310012, China;
2. Zhejiang Institute of Water Resources and Estuary, Hangzhou 310020, China)

Abstract: As the key technology, heterogeneous integration of spatio-temporal geodatabase guarantees the ability of data service. Based on analyzing the database heterogeneous integration theory and summing up the deficiencies of existing technology and methods, we proposed a scientific and standardized heterogeneous integration technology process based on semantic mapping and ETL theory. Then, we used external files such as XML, semantic mapping tables, to develop the automatic processing tools with rigorous logic, flexible expansion and strong adaptability based on FME, and formed B/S mode of online heterogeneous integration solution. Finally, we verified this solution by wide range of applications. The study has important supporting significance for realizing quick response to spatio-temporal geodatabase integration requirements and improving the efficiency and quality of data services.

Key words: heterogeneous integration, FME, online, ETL, spatio-temporal geodatabase

传统测绘地理信息正向追求高精、高效和智能化的新型测绘快速转型升级。新型测绘需要高效智能化构建统一、实用的结构化时空地理数据库成果，而在已积聚的海量多源地理信息数据库中，异构性成为空间数据信息共享的瓶颈[1]。在实现时空地理数据高度集成的过程中，数据格式转换、数据库结构整合、数据要素融合是紧密衔接的主要环节。已有研究在CAD图形和地理空间数据库的数据格式互转[2-4]、异构专题数据库的结构转换[5-7]、基于栅格和矢量的多源地理信息空间融合[8-10]等方面进行了较系统的剖析。为实现将地理空间数据库异构整合的智能化处理代替繁重且易出错的人工处理，已有学者基于MapGIS、ArcGIS Spatial ETL、ArcGIS Object开发设计了多源异构空间数据转换、重构、合并、集成工具，提高了数据集成效率[11-13]。然而，随着时空地理数据库数据体量的不断增长以及异构整合场景的不断丰富，现有技术常用的桌面端模式已不能满足快速、自适应转换的需求，因而亟需对异构整合关键技术展开深入研究。

鉴于时空地理数据库异构整合转换本质上是异构数据库的抽取、转换和重新统一组装的问题，即ETL过程[14]，本文重点基于ETL理论和语义映射研究了异构整合整体技术方案和关键技术，并依托FME平台进行异构整合功能开发，形成B/S模式、基于云计算的面向时空地理数据库的科学、标准、自动化的在线整合平台，以实现新型测绘时期对海量时空数据库整合的快速响应。

1　时空地理数据库概述

1.1　时空地理数据库

地理信息数据仓库涵盖了数量庞大且种类丰富的

收稿日期：2021-10-26。
项目来源：浙江省自然资源厅重点科技资助项目（2020-53）；浙江省测绘科学技术研究院院长科研资金资助项目（DFP2021D0504）。

时空地理数据库（图1），各类数据库的建设依据为数据字典。数据字典规定了要素分类、数据库分层、数据库图层属性结构和各要素的图形表示等具体内容，是数据库组织异构性的主要体现。不同数据库中水系面要素的图层结构如图2所示，其中数据库中图层名称、字段名称、字段类型均存在一定差异。



a 天地图数据库　b 927海岛数据库　c 三调数据库　d 1∶2 000数据库　e 新版1∶2 000数据库

图1 不同结构的时空地理数据库图层示例

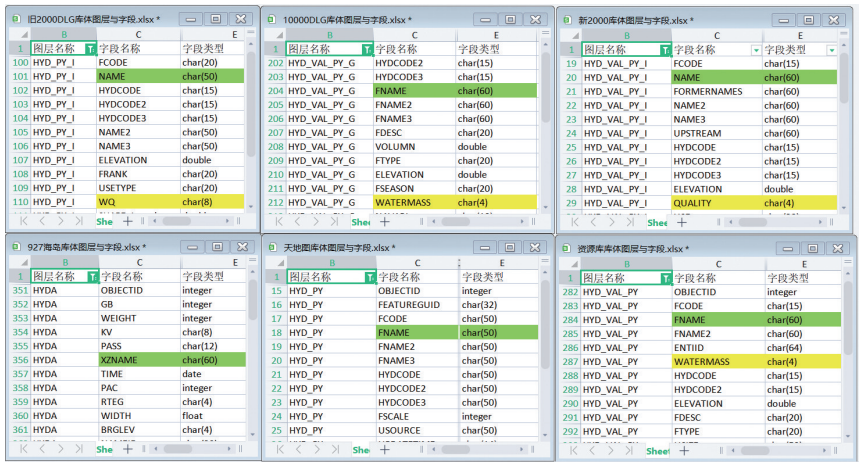


图2 不同时空地理数据库中水系面层的字段结构示例

时空地理数据库异构整合的目标是将依据不同数据字典创建的数据库整合转换至统一框架结构的目标数据库，通常涉及要素代码转换、要素图层转换、属性结构转换等步骤。相较于数据库要素级的处理，数据库的异构整合转换在数据库建设工作中的占比更多、难度更大、技术门槛更高，且整合转换的质量好坏、效率高低将直接影响时空地理数据库成果的正确性以及地理信息数据服务的有效性和敏捷性。

1.2 现有技术分析

面向时空地理数据库异构整合的研究总体围绕3个方面：数据库整合集成的关键过程分析、异构整合技术流程的制定和多源异构空间数据自动集成工具的开发。现有研究成果初步形成了异构整合的技术流程，在一定程度上提高了异构整合的效率，但存在以下问题亟待解决：①异构整合过程中异常问题的处理不够深入、全面，整体工作流有待细致梳理，形成严谨、科学、标准化的技术方案；②异构整合的自动化实现常限于某一种GIS平台，桌面模式无法稳定支撑海量数据库的规模化转换，在云计算资源的支撑下开展在线异构整合技术研究十分迫切；③在异构整合功能开发的过程中，需要充分考虑扩展性、灵活性和自适应性，设计一个通用的转换方案或平台以满足不同异构转换场景。

2 基于ETL的异构整合关键技术分析

2.1 语义映射关系的自动创建

源数据库和目标数据库的映射关系是实现数据转换的基础，其创建依据是对应的数据标准或数据字典，主要面向要素代码、图层名称和属性名称3个层面，如表1所示。通过源描述与目标描述进行语义匹配可自动创建映射关系，再以数据字典为辅助完善映射关系，该方式能大量减少人工作业，形成的语义映射文件能重复应用于大批量数据库异构整合中。语义映射的关键是确保要素、属性、图层都能找到各自的映射，通过映射关系在数据转换后实现数据结构逻辑的重新组织，并解决形式表达层面上的冲突。

表1 映射关系示例

映射关系	源名称示例	源描述示例	目标名称示例	目标描述示例
要素代码	4201023351	建筑中国家高速公路（边线）	4201040351	国家高速公路—建筑中（边线）
要素图层	HYD_PT_I	水系（点）	HYD_VAL_PT_I	水系点要素
属性字段	ORIENTATION	方向	ANGLE	旋转角度

2.2　异构整合的ETL关键环节

从ETL理论技术角度分析异构整合，即以海量、复杂、多样的地理空间数据库为数据源，从中按需提取，再经过数据清洗、加工、整理和转换等步骤，最后装载至目标框架结构数据库中。

1）数据清洗。时空地理数据库作为数据转换的数据源难免存在不合规范的数据，数据清洗的目标是分析找出数据缺失（如水系面丢失常水位岸线、丢失水系交汇线、水系对象线等）、数据重复（如要素自重复、要素唯一编码重复等）、要素代码异常（如超出要素代码表、要素图层错误等）和属性不规范（如小数取位不正确、属性赋空不正确等）等问题数据，并进行处理。在数据清洗环节设计自动化检查和处理算法，开发工具集实现数据补缺、数据自动化去重、属性规范化，对于保证数据转换的正确性具有重要意义。

2）数据转换。数据转换的前提是已有基于数据字典创建的映射关系表，但不排除数据库与数据字典存在属性字段命名或类型冲突，因此考虑先进行数据库结构框架的自动提取，并与数据字典进行对比，在确保数据库符合对应数据字典要求的条件下再开展数据库的异构转换。为适应不同的异构转换场景，本文采取调用要素代码映射表、图层映射表、属性字段映射表等外部配置表的方式实现源数据库向目标数据库的异构转换。

3）数据重新装载。由于可扩展标记语言（XML）能对文档结构和数据进行约定，在实现数据信息标准化、统一化和共享上具有独特优势[15]，因此在异构整合前先创建或在ArcGIS平台下导出目标框架结构XML文件，再通过外部调用该文件，从而保证将源数据库精准地装载至目标框架结构的数据库中。

2.3　数据库异构整合的异常问题

在地理空间数据库的ETL过程中，必不可少地会面临数据源异常和转换过程异常等问题，如数据源中要素代码所指要素放错层、数据源中要素代码异常、数据转换过程中要素与目标图层的几何类型不一致（面转线、面转点、线转面、线转点等如表2所示）、要素几何类型特征码与实际要素几何类型不一致（表3）等问题。

表2　属性类型冲突示例

目标图层名称	目标属性名称	目标属性描述	目标属性类型	原始图层名称	原始属性名称	原始属性描述	原始属性类型
HYD_VAL_PY_I	RANK	航道等级	Short Integer	HYD_PY_I	FRANK	航道等级	Text(20)
TRA_NET_LN_I	LANECOUNT	车道数	Short Integer	TRA_LN_I	FNUMBER	车道数	Text(3)
TRA_AUX_PY_I	STOREY	层数	Short Integer	TRA_PY_I	OGLAYER	层数	Text(3)

表3　几何类型冲突示例

新要素代码	新要素名称	新图层名称	新几何类型	旧要素代码	旧要素名称	旧图层名称	旧几何类型
3804090500	挑廊（面）	RES_BUD_PY_I	面	3804090351	挑廊（线）	RES_LN_I	线
3102010106	居住小区（名称对象点）	ANO_HGN_PT_I	点	3102010330	住宅区（名称对象线）	RES_LN_I	线

在异构数据库的整合转换中，应考虑数据库的完整性和安全性以及转换过程的稳定性和抗干扰性。鉴于异构数据库整合转换中错误常发生在最后的数据写入阶段，为减少耗时本文采用前置异常冲突检查、对异常数据标记分流输出的处理方式，在ETL完整执行后再对异常数据库进行修改处理。

3　在线异构整合的设计与实现

3.1　异构整合功能模块的开发

本文通过对异构整合关键环节的分析梳理，构建了异构整合技术流程，如图3所示。FME Workbench中内置了上百个转换器，可支持数据库的快速ETL实现，其中PythonCaller转换器能调用ArcPy、FME Objects和sys等函数类库，从而为FME+Python+ArcPy的集成和拓展提供无限可能，Schema Maper等转换器能快速实现属性映射。

基于FME进行异构数据库整合转换功能开发时，只对源数据库进行读取操作，不会因异常而回滚全部操作，能避免破坏源数据库。此外，本文采用外部配置表作为映射关系约束，以目标数据库XML框架文件为输出结构约束，从而提高转换的自适应性和正确性。其主要子功能模块包括数据库框架结构提取、要素属性预检查、水位岸线与交汇线自动生成、映射关系表自动创建、异常数据检查输出、匹配数据自动转换等，最终形成全流程异构整合工具集发布至FME Server。

3.2　在线异构整合平台开发与应用

FME平台通过引入在线空间数据ETL技术，可建立基于SOA的空间ETL服务，其多节点的并行处理能力为海量数据的快速整合提供了支撑。因此，联合

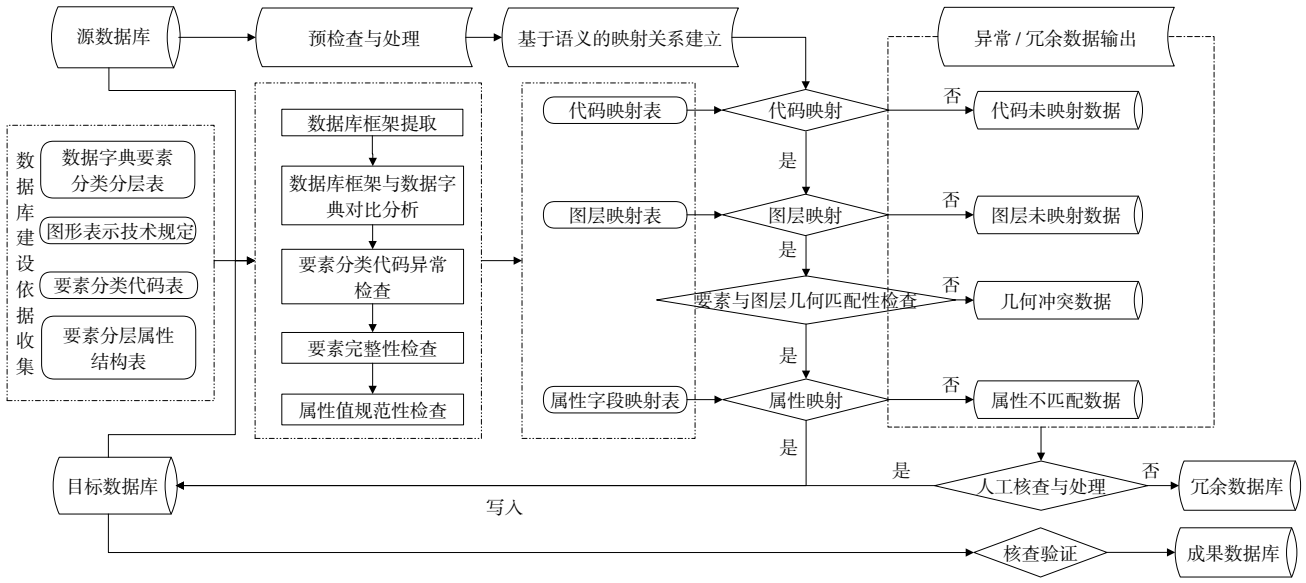


图3　时空地理数据库异构整合技术流程图

FME平台的Server和Integration的API接口进行在线异构整合功能的二次开发。其关键技术路线如图4所示。

管理员在后台进行目标数据库模板XML、代码映射表、图层映射表、属性字段映射表和输出结果路径等统一参数的预配置，用户即可访问前端主页，实现本地时空地理数据库的一键上传和在线异构整合结果的一键下载，如图5所示。通过浙江省新型测绘体系数据库整合建设项目的大量应用，验证了本文研究成果在时空地理数据库异构整合中的正确性、完整性、高效性和灵活性。

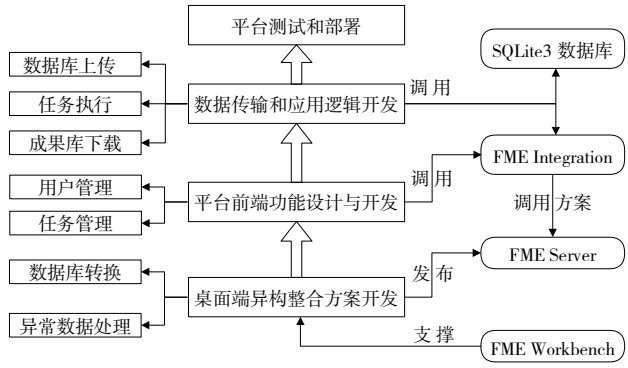


图4　在线异构整合开发关键技术



a 在线异构整合方案的后台配置



b 本地源数据库上传



c 整合转换后结果下载

图5　时空地理数据库在线异构整合的应用

4　结　语

本文借助ETL理论，提出了一套科学、标准化的数据库异构整合处理流程，并基于FME提供的ETL技术和ETL服务进行二次开发；采用读取数据库框架XML文件、图层、属性、代码映射表等外部文件的方式，实现了海量多源异构时空地理数据库的在线高效精准整合，为构建全省框架结构统一的地理空间数据

库提供了一种集约型的解决方案，对优化地理空间数据库的生产组织模式以及进一步提升自然资源数据支撑效率和数据服务质量具有一定的参考意义。

参考文献

[1] 刘占伟,刘厚泉. 基于GML的多源异构空间数据集成系统的设计[J]. 计算机工程与设计,2007,28(8):1 962-1 965

[2] 李逵. FME在基础地理信息数据库建设中的应用研究[J]. 测绘通报,2016(3):115-117

[3] 陈年松. 基于FME的CAD与GIS数据共享研究[D]. 南京:南京师范大学,2008

[4] 俞志强,许大璐,赵志强,等. 一种基于AutoCAD的异构数据对象化智能识别方法[J]. 测绘通报,2014(5):103-106

[5] 阳成. 地形地籍异构数据整合建库方法研究与应用:以广州吕地居委会地形地籍数据整合入库为例[J]. 测绘与空间地理信息,2020,43(9):146-151

[6] 宋晓红,张立朝,禄丰年,等. 地理国情普查中多源异构数据整合研究[J]. 测绘通报,2014(9):104-107

[7] 武雪丽,董健全,李智昕. 异构数据库的组合转换及其现实应用[J]. 计算机工程与应用,2004,40(11):188-190

[8] 陈换新,孙群,肖强,等. 空间数据融合技术在空间数据生产及更新中的应用[J]. 武汉大学学报(信息科学版),2014,39(1):117-122

[9] 杨灿,汪齐松. 多源地理矢量空间数据融合研究[J]. 测绘通报,2019(12):112-115

[10] 孙群. 多源矢量空间数据融合处理技术研究进展[J]. 测绘学报,2017,46(10):1 627-1 636

[11] 张怡鹏,高德政. 多源异构空间数据集成工具系统的设计与研究[J]. 测绘通报,2009(2):21-24

[12] 林璐,王永红,苟建雄. ArcGIS Spatial ETL工具在基础地理信息数据库建设中的应用[J]. 北京测绘,2017,31(5):127-129

[13] 王育红,刘康晨. 面向新型测绘的时空数据属性连接及扩展方法研究[J]. 测绘通报,2021(5):5-9

[14] 徐俊刚,裴莹. 数据ETL研究综述[J]. 计算机科学,2011,38(4):15-20

[15] 秦英,马永起,蒙立荣. 一种基于XML的数据转换方法的设计与实现[J]. 微型机与应用,2017,36(20):30-33

第一作者简介：王慧（1990—），硕士，注册测绘师，主要研究方向为遥感地理信息数据智能处理算法研究与开发。

常用参考文献著录要求

期　　刊 [序号] 主要责任者. 题名：其他题名信息 [文献类型标识/文献载体标识]. 刊名，出版年，卷(期):起止页码

专　　著 [序号] 主要责任者. 题名：其他题名信息 [文献类型标识/文献载体标识]. 出版地：出版者，出版年：引文页码

析出文献 [序号] 析出文献主要责任者. 析出文献题名：其他题名信息 [文献类型标识/文献载体标识]// 原文献主要责任者. 原文献题名：其他题名信息. 出版地：出版者，出版年：引文页码

专　　利 [序号] 专利申请者或所有者. 专利题名：专利号 [文献类型标识/文献载体标识]. 公告日期或公开日期

电子公告 [序号] 主要责任者. 题名：其他题名信息 [文献类型标识/文献载体标识]. (更新或修改期) [引用日期]. 获取和访问路径

报纸文章 [序号] 主要责任者. 题名：其他题名信息 [文献类型标识/文献载体标识]. 报纸名，出版日期(版次)

附表1　文献类型和标识代码

参考文献类型	文献类型标识代码
普通图书	M
会议录	C
汇编	G
报纸	N
期刊	J
学位论文	D
报告	R
标准	S
专利	P
数据库	DB
计算机程序	CP
电子公告	EB
档案	A
舆图	CM
数据集	DS
其他	Z

附表2　电子资源载体和标识代码

电子资源的载体类型	载体类型标识代码
磁带（magnetic tape）	MT
磁盘（disk）	DK
光盘（CD-ROM）	CD
联机网络（online）	OL

注：1、英文和中文参考文献的要求、格式一样。

2、不能少于7条。

3、作者人数在3个以上时写出3个加“等”字，只有3个或不足3个则全部写出。

4、凡可标注数字对象唯一标注符（doi码）的，一律标注于该条文献最后，用“.”与之前内容间隔。

5、参考文献在正文中要有与之相对应的标示。参考文献的排序为在正文中首次出现的顺序，对于不能在正文中明确标注的文献，一律放在所有能标注文献的后面。

6、希望本刊文献能得到您的关注。