Science Discovery
2017; 5(5): 331-339
http://www.sciencepublishinggroup.com/j/sd
doi: 10.11648/j.sd.20170505.16
ISSN: 2331-0642 (Print); ISSN: 2331-0650 (Online)

Research on Universal Processing Framework of Multi-source Heterogeneous Data Integration in GIS Domain Based on Software Technology

Xu Qiang*, Fan Hong, Dang XiaoHu

College of Computer Science, Shaanxi Normal University, Xi'an, China

Email address:
609145649@qq.com (Xu Qiang)
*Corresponding author

To cite this article:
Xu Qiang, Fan Hong, Dang XiaoHu. Research on Universal Processing Framework of Multi-source Heterogeneous Data Integration in GIS Domain Based on Software Technology. *Science Discovery.* Vol. 5, No. 5, 2017, pp. 331-339. doi: 10.11648/j.sd.20170505.16

Received: June 19, 2017; **Accepted**: July 18, 2017; **Published**: August 7, 2017

Abstract: It is difficult to make fine-grained random call in GIS system development or data processing owing to the multi-source heterogeneity of spatial data. Moreover, to convert the data into a unified format easily can lead to the loss of the core attributes. In this paper, we discuss the translation methods among several different data formats, including attribute and business data, network data, CAD data, heterogeneous GIS spatial data, and domestic and foreign GIS data, etc, based on software technology like FME, MyFME, Arcgis, MessagePack package, crawling tools and so on. Then, we propose a universal processing framework for multi-source heterogeneous data integration in the GIS field, and verify the performance of the proposed framework through integrating several kinds of data.

Keywords: Framework, Business Data, Multi-source, Universality

基于软件技术探索GIS领域多源异构数据集成的普适性处理框架

徐强*，范虹，党小虎

计算机科学学院，陕西师范大学，西安市，中国

邮箱
609145649@qq.com（徐强）

摘要：空间数据的多源异构性，导致GIS系统开发或数据处理时更细粒度的随机调用比较困难，而简单转换为统一格式又容易造成数据核心属性的丢失。鉴于这一现状，本文采用FME、MyFME、Arcgis、MessagePack程序包和爬取工具等软件技术，解决了属性业务数据空间化、网络数据空间化、CAD数据转换到GIS格式、异构GIS空间数据统一最优化传输、以及各国GIS数据统一共享利用一些难题，提出了GIS领域多源异构数据集成的普适性处理框架，并通过各类数据的集成入库实例验证了所提处理框架的性能。

关键词：框架，业务数据，多源，普适性

1．引言

过去的20年里，各领域的数据量在大规模增加。根据IDC的报告显示，2011年全球产生的数据量达1.8 ZB（≈10^{21}B），是近五年内的9倍，而且这个数据还会在将来的两年再翻一倍[1，2]。毋庸置疑，大数据时代已经来临，大数据所蕴含的巨大价值也已引起了很多行业的重视。然而，由于这些数据来源不同，且数据格式、数据标准、数据管理平台和管理方式各不相同，使得这些数据之间无法互联共享，形成了一个个散乱的“信息孤岛”，导致“数据爆炸但知识贫乏”的现象，造成了极大的资源浪费。因此，如何集成这些量纲不一、形式多样、既有定量数据又有定性文字描述的数据，为信息资源共享和综合利用提供统一平台，成为亟待解决的重要问题。

早期的数据集成技术出现于二十世纪八十年代，以多数据库和联邦数据库这两种传统的研究为主。最早的多数据库法是由惠普实验室数据库技术部开发的Pegasus系统[3]，而对联邦数据库法的研究以美国密歇根-迪尔伯恩大学为首的几所北美大学取得了较大进展[4]。随后空间数据集成得到了国内外科研人员的重视，并取得了很多的研究成果，如：Li等[5]开发了对空间敏感的数据集成的原型系统、Belussi等[6]提出了基于几何对象统计表示的空间数据库的多精度表示框架、程海军[7]提出了利用Oracle Spatial技术进行数据集成和转换的方法、张文江[8]应用GML作为空间数据集成的中间数据交换格式，实现基于不同数据模型的空间数据的相互转换，Bansal [9]提出了基于语义的ETL框架进行多源数据的集成等。当前为了满足实际应用的需求，GIS领域数据与非GIS领域数据的集成成为研究主体，诸如业务数据、网络数据、CAD数据等与GIS数据的集成等，相应也有一些研究成果[10-12]，但研究过程依然存在诸如集成效率较低、因各种数据存储相对独立导致的查询效率不高、爬取数据框架实现难度偏大等问题。从GIS领域自身来说，还存在中国国产数据无法与国外数据共享的难题，目前基于GML的统一格式集成方式存在数据体积过大，网络传输效率低问题，比如于一男[13]等人就注意到了XML文档的传输速度问题，从压缩文档算法传输方面做了研究，并没有从数据本身入手解决。因此探索研究面向实际应用的具有普适性且较为简洁的多源异构数据集成方法成为必要。

本文利用FME在抽取数据方面的优势，研究业务数据以及网络爬取得来的网络数据抽取并且空间化，利用Arcgis处理CAD数据到GIS领域的转换问题，利用MyFME打通了中国国产数据到国外GIS格式数据的通道，利用一种比XML、JSON更加轻量级的格式来进行GIS集成后数据的极小化存储。最终根据空间数据集成领域的情况提出了多源异构数据集成的普适性处理框架，并通过各类数据的集成入库对框架进行验证。

2．多源异构数据集成入库核心问题的处理

2.1．专题报表以及网络业务数据处理方案

2.1.1．专题数据批量空间化并集成入库

GIS系统中所使用的数据有相当一部分来自于报表调查数据，如随后实例中19个丝绸之路经济带中包含能源净出口、人均GDP等数据的报表数据。这些统计数据量大且包含有丝绸之路经济带以外的国家数据，此类数据需要解决的技术难点在于，第一，在WebGIS系统开发的时候往往需要一些依赖于集合体的属性数据，而这些属性数据并非天生依附于几何体，往往存储在报表中，如何将其空间化并附加于GIS几何体成为问题；第二，如何过滤消除无用数据成为GIS要素属性成为技术难点；第三，如何解决多报表附着于同意几何体，成为难点。

实现报表的空间化，基于FME提供了最简单的ETL空间化的方式，第一步，针对多报表分别在工作空间指定抽取读取源和一个抽取写源头，根据识别的属性字段，连线待用字段源头与目的地实现空间化，自动剔除对所用目的无关的无效数据；第二步，利用FeatureMerger转换器设定Requestor和Supplier，并设置两表关联字段，类似于数据库中的主外键，两两合并，以此类推，完成多表合并操作，值得注意的是，对于关心的目标数据，可以手动拉区线段，实现目标字段的抽取转化，即可解决无效数据过滤问题。

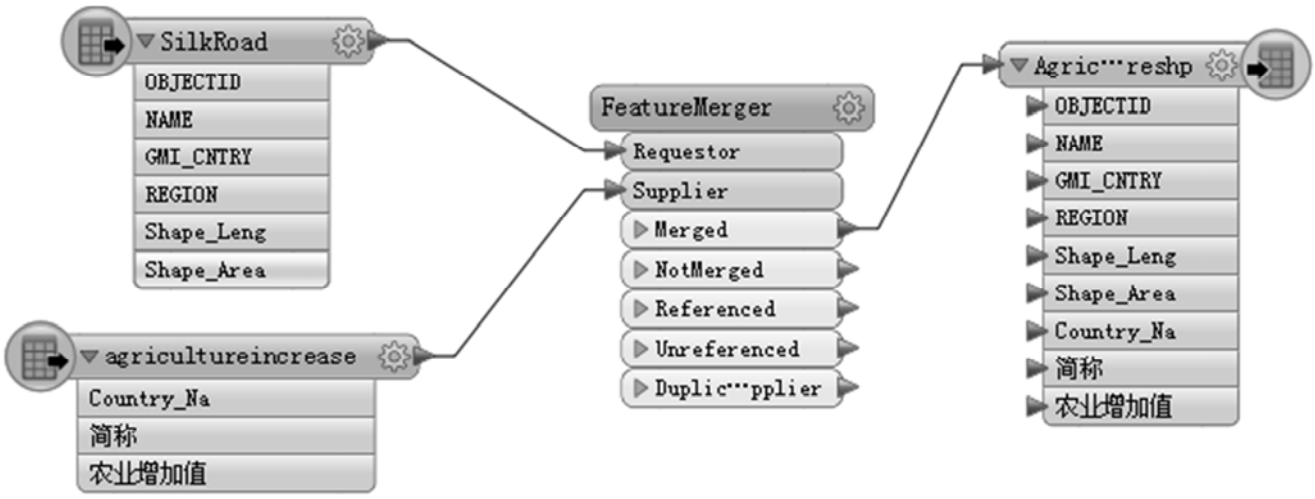

图1 农业增加值FeatureMerger转换器工作空间。

2.1.2. 专题数据批量空间化并集成入库实例验证

本文以丝绸之路经济带的原油和天然气数据空间化为例，抽取读取源为报表文件，抽取写源头设定为shp格式文件，并制定与要合并的GIS要素一样的地理坐标系（比如WGS1984坐标系），抽取转换即可，手动指定暴露的属性，实现空间化过程。实现空间化后的业务数据虽然有了空间格式，但因为没有空间位置信息和图形信息，还没有空间特性。所以接下来要使空间化后业务数据依附于各自归属的空间位置或者图形，在此指的是国家名称，为后续可视化分析提供方便。空间化后的要素类需要关联相关国家原有的GIS要素类文件（如shp文件）实现统一存储，借助 FME的FeatureMerger转换器，定制关联Requestor和Supplier设置关联字段以过滤剔除无用数据，手动指定所需目标属性字段，实现要素合并，最后通过ArcCatalog中Database Connection项指定数据库平台、授权方式、实例名、数据库名称建立数据库连接，import要素类到SilkRoad数据库，单从农业增加值专题数据。FeatureMerger转换器的工作空间如图1所示。

2.1.3. 库中报表数据细粒度随机调用

业务数据集成入库后，在系统开发时就可以对每一个要素类进行随机化的细粒度调用，例如可以小至一条河流的随机调用。对于多源数据服务系统开发后的初步效果，借助空间数据引擎ArcSDE来实现。首先借用PropertySetClass类的SetProperty设置主机名、实例名、用户名、密码、数据库版本；然后利用SdeWorkspaceFactoryClass类的Open方法伴随设置好的PropertySetClass对象打开工作空间；进而利用返回的工作空间对象get_Datasets方法获得枚举集合；最后遍历集合，遍历每一要素类构造要素图层，加载到MapControl进行显示即可。图2给出了初步显示效果，这一过程在工程项目实践中往往被抽取成工具类，进而使数据的操作变得简单。

图2 专题数据可视化(以丝绸之路农业增加值专题数据查询为例)。

2.1.4. 网络批量爬取数据并集成入库

GIS系统开发经常需求一些官方的数据，这些数据通常公布在其如国家统计局，地方测绘局等官方网站，技术难点为：第一，网页数据如何自动化爬取，甚至有时候需要定时爬取；第二，网页数据爬取后是存储在数据库，如何将库中数据导入FME工作空间，并实现之前的空间化，进而使用成为难点。

解决两大难点，第一步，八爪鱼作为按照流程采集的工具，帮助任何需要从网页获取信息的客户实现数据自动化采集，编辑，规范化[14]，精选开放工具八爪鱼，鉴于其定时化与自动化，将要采集的网址输入到软件中的浏览器页面。建立表格循环，这个地方需要注意的是，我们把鼠标挪动，浅蓝色的背景表示的都是可以选中的选项，表格抓取的话需要先选中的是一行，通过鼠标无法直接实现，需要我们先选择第一行中的第一格内容，系统会弹出一个对话框来，选择[TR]选项实现整行的选择。如图所示，然后将每行的元素全部显示到列表中，点击[循环]，建立循环框。提取字段的时候需要特别注意，先要选中循环列表中的一行，然后在下方的浏览器上找到你刚选中的那行，点击你要选取的字段，提取相应的内容。之后，就是设置采集时间（单击采集可以不用设置直接跳过），选择单机或云采集的方式开始提取数据，提取完毕之后导出想要的格式即可，目前可以导出成TXT、EXCEL、HTML、网站发布、数据库等格式，只支持单列爬取的网站需要合并多表。第二步，借助FME提供的半开放SQLCreator转换器，指定连接字符串、查询语句，特别要指定暴露属性，实现

从爬取的数据库中抽取数据，并空间化为空间要素类，反倒回数据库，进而开发利用。

2.1.5. 网络批量爬取数据并集成入库实例验证

本文以八爪鱼爬取工具爬取国家测绘局公布的分地区按年GDP数值统计数据为例。具体操作：打开网页操作，将光标定位到要采集列的第一行单元格点击操作，接着创建一个元素列表用以处理一组元素，添加并编辑列表，继续进行点击同列第二行元素，添加编辑列表，循环操作即可完成单列循环采集，接着抓取此列任意一个行要素编辑列名，从而映射为数据库的表即可完成爬取过程。由于国家统计局网站只支持单列采集，就存在数据库中多表合并的问题，在此利用select into语句完成合并为单表存储，爬取的数据也存在空间化并附加为其他源地图属性的问题，涉及到FME的数据库操作。借助FME的半开放SQLCreator转换器，指定连接字符串、查询语句，特别要指定暴露属性，实现从爬取的数据库中抽取数据，并空间化为空间要素类，转换工作空间如图3所示。最后按照业务数据所描述的属性合并依附方式进行要素合并，并实现最终入库，完成数据的处理，为后续的开发做好准备。数据集成入库后，做与专题数据同样的处理，得到如图4所示的效果。

图3 专题gdp数据SQLCreator工作空间。

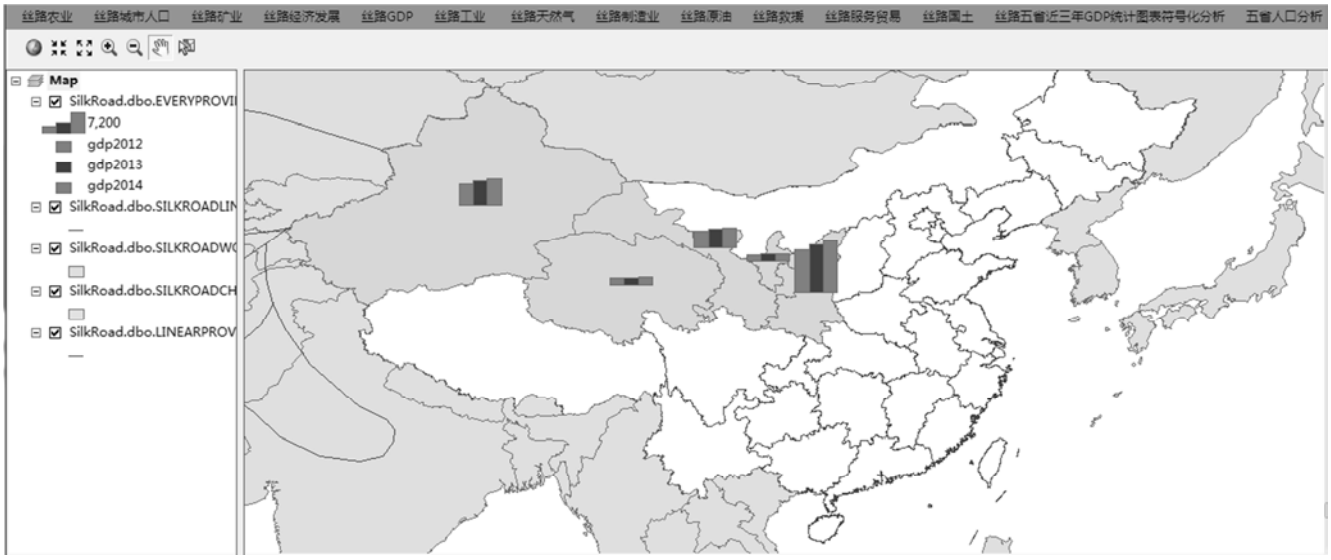

图4 网络数据爬取空间可视化展示(丝绸之路五省三年gdp情况)。

2.2. CAD数据集成入库问题解决方案

2.2.1. CAD数据集成入库问题解决方案

GIS系统开发中，很多时候都包含着CAD数据的入库研究，然而这其中也存在着一些技术难点，如：冯文娟[15]利用FME初步实现了CAD数据到shp煤矿数据的转换，但数据较为简单，对扩展属性等问题没有着重研究；郑江等人[16]尝试结合DWGDirect.NET与Arcgis Engine编程方式实现CAD数据到GIS数据的转换，编程难度较高。上述方法实现起来都较为复杂，且属性容易丢失，不具备普遍适用性。另外，CAD数据往往是野外考察、土地考察、测绘等方面的调查结果形式，因为比较重视清晰的视觉、标准，而在数据结构和模型、数据属性和域、分析功能等方面不太重视；而且，相比SHP格式数据，DWG格式数据无法处理海量数据，并且无法有效管理结构化数据[17]，使得研究CAD集成到GIS数据成为必要。

解决此类难点，本框架内，利用ArcMap来对CAD数据进行处理。具体过程为：第一步，利用select by Attribute工具对Layer属性对数据进行分离，并导出为一个个要素类；第二步，使用多条件查询得到界址线所分割的多边形，并检查接边处理效果，如有问题，需要面状手动修复，若无，进行以下步骤，并导出，然后基于空间位置，最后利用join工具进行注记作为属性到多边形的套结。

2.2.2. CAD数据集成入库实例验证

本文以第二次土地调查广东省从化市联群村数据为例给出具体分离过程，文中数据既包含DWG格式数据固有的属性数据，也包含门牌号、姓名、地类号等扩展属性。对于该数据的处理，首先利用ArcMap导入dwg格式勘测图，然后针对Annotation、point 和 polyline三种形式通过select by Attribute工具伴随layer区别来分离数据，但Annotation形式需要采用feature to point工具来实现导出为GIS要素类的过程，而point和polyline形式直接右键导出即可。

对于CAD数据的有些注记，需要转换为点来处理。但是从语义方面来讲，它附属于一定的几何体，是这些集合体固有的属性，比如调查数据中地类号、队名、姓名预编号注记是属于JMD图层的，需要进行属性的附加。具体操作是首先通过上述的select by Attribute分离出JMD polyline图层，为了处理界址线，还需要在使用select by Attribute时候条件语句增加界址线，即："Layer" = 'JMD' OR "Layer" = '界址线'，然后利用feature to polygon实现到多边形几何体的构建，进而导出要素类，最后利用join工具将JMD导出要素与多个注记按照空间位置的方式附加，完成注记附加。另外，对每一要素类的生成都要在入库之前进行检查，杜绝可能对建库造成影响的因素（例如接边，拓扑错误等）[18]。根据检查结果做相应的处理，即：对道路、河道以及没有构面前的居民区线等要素进行接边处理，对已构成的面状要素进行面状修复。最终按照前两种数据的入库方式将批量的shp文件导入到数据库，完成整个CAD数据的继承入库，整个联群村的处理流程如图5所示。数据入库后处理效果如图6所示。

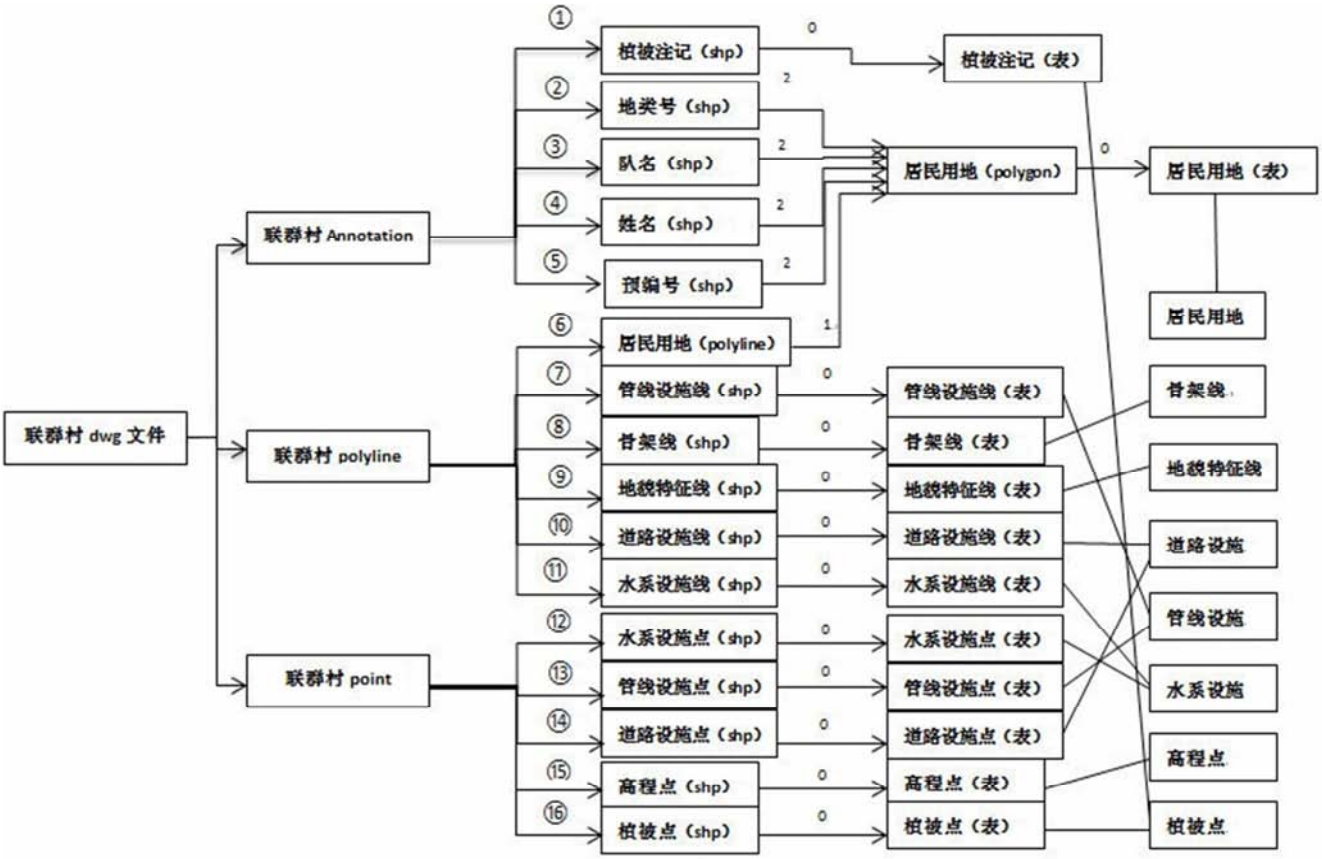

图5 联群村dwg数据集成入库流程图。

注：①:①：Select by Attribute（Layer=ZBTZ），并利用Feature to point导出

②:②-⑤操作与①类似，只在图层不同，图层分别为：地类号，队名，姓名，预编号

③:⑥：Select by Attribute（Layer=JMD），并普通另存为导出

④:⑦-最终，操作与⑥类似，只在图层不同，图层分别为：GXYZ, assist, DMTZ, DLSS, SXSS, SXSS, GXYZ, DLSS, GCD, ZBTZ

⑤:0：检查并入库1: feature to polygon并进行构面检查，出现错误进行手工构面 2: join based on spatial location

⑥：最后的无箭头实线表示属于

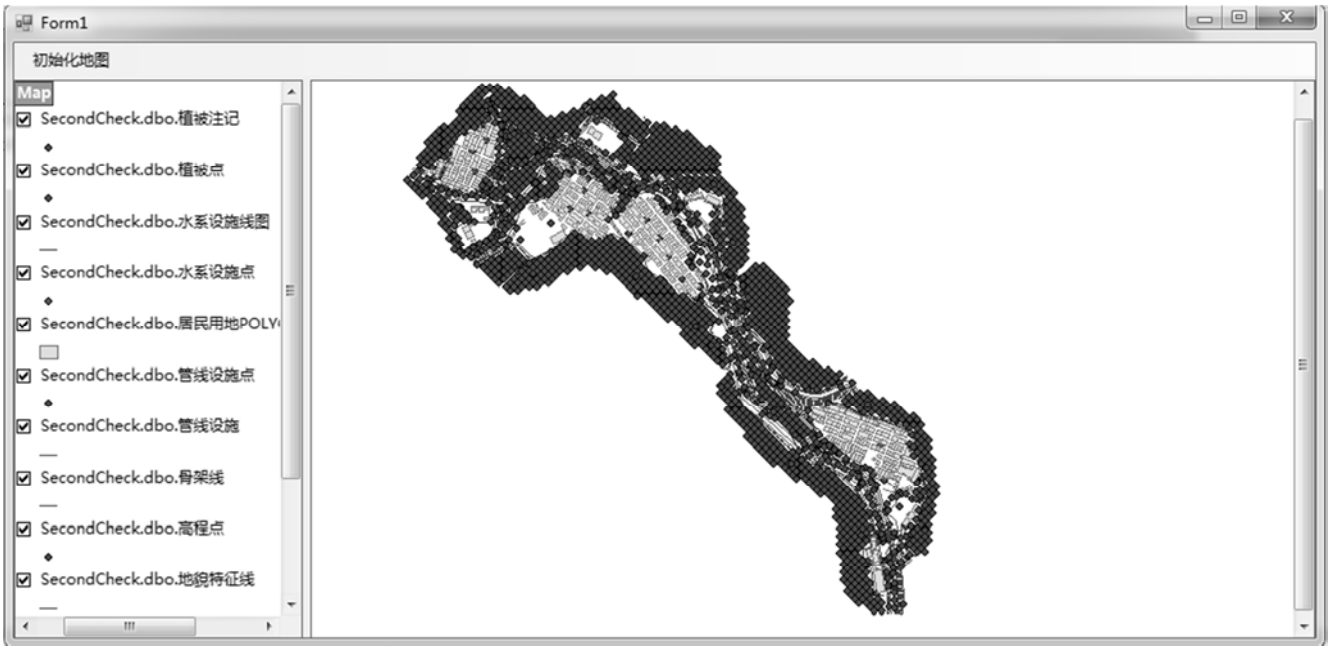

图6 CAD数据集成到GIS展示图（第二次土地调查联群村调查图）。

2.3．国产与国外空间数据集成问题解决方案

2.3.1．国产与国外空间数据集成问题解决方案

GIS开发或者数据分析过程中，有时候需要一些异构平台数据集的集成，这成为难点。比如国产数据集与国外数据集的互相转换共享就是一个重要研究领域。本文以SuperMap的udb数据集与Arcgis的shp数据集集成为例实现集成。MyFME采用FME插件技术，提供对各国GIS软件的数据格式提供转换支持。能够将MapGIS、SuperMap等数据格式和其他国内外GIS数据格式进行方便、快捷、自由的转换[19]。实现较为简单，只需指定各国国内与国外集合要素的映射，即可完成转换。

2.3.2．supermap数据到arcgis数据的集成实例验证

udb数据具有point、line和region的结构，以某校园的udb数据为例，相应的point直接转换为shp固有的point, line准换为polyline, region相应的polygon类别，借助myfme对中国国产数据的支持，根据上述转换规则，进行工作空间的构建，抽取路线、照片兴趣点、建筑物三个类别部分的转换如图7所示，转换前后的校园地图对比如图8所示。

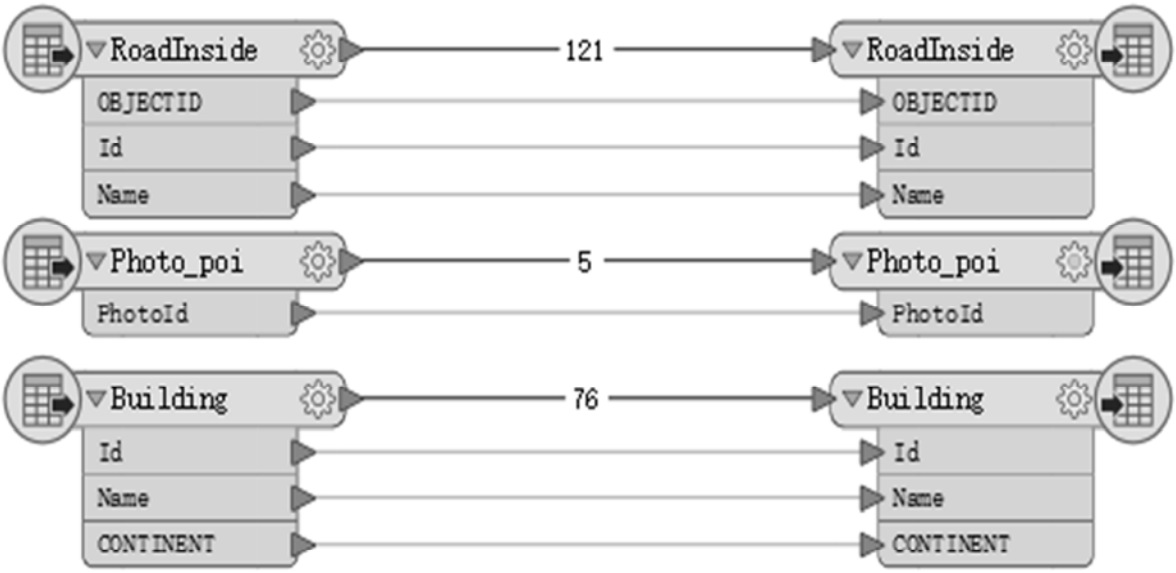

图7 路线，照片兴趣点，建筑物转换图示。

图8 (左)udb数据图，(右)shp数据图。

2.4．基于MessagePack最优化空间数据集成问题解决方案

2.4.1．基于MessagePack最优化空间数据集成问题解决方案

在GIS系统开发过程中，尤其设计网络编程的领域，空间数据传输速度成为重要考量，进而基于GML语言的研究成为热点。但是GML文档数据不够精简，影响网络传输速度。本文提出一种比JSON更加轻量级的数据格式MessagePack来实现数据集成进而更加轻量级的加快数据传输，极速完成GIS系统的升级，这也解决了GML数据冗余的问题。

MessagePack是一种高效率的二进制序列化格式，一些小整数被编码为单字节，有些典型的小字符串仅仅需要额外的一个字节，而不是字符串本身，如JSON串{“compact”:true，“schema”:0}，存储需要27字节，转换为MessagePack：82 A7 compact C3 A6 schema 00，仅仅需要18字节，大大压缩了存储，尤其在大数据量或者典型字符串多的数据中，典型字符串如true, false等。

实现基于MessagePack空间数据集成的过程为：第一步，使用gdal编程模型实现空间数据到gml文档的转换，样例如图9所示；第二步，利用MessagePack-cli命令实现gml到MessagePack数据格式的压缩存储；第三部，利用socket或者Netty多线程非阻塞传输，最后反向破解。

```
ogr.RegisterAll（）;
gdal.SetConfigOption
（"GDAL_FILENAME_IS_UTF8"，"YES"）;
gdal.SetConfigOption（"SHAPE_ENCODING"，""）;
gdal.SetConfigOption （ "OGR_FORCE_ASCII" ，
"NO"）;
StringstrVectorFile="C:\\cities.shp";
DataSource ds = ogr.Open（strVectorFile，0）;
Driver dv = ogr.GetDriverByName（"GML"）;
dv.CopyDataSource（ds，"C:\\cities.xml"）;
```

图9. gdal实现shp到GML转换

2.4.2．基于MessagePack最优化空间数据集成实例验证

文章设定WEBGIS开发环境，利用socket编程测试基于GML与MessagePack方式传输的效率差异，选取的GML文档数据量级分别为2.43MB、15.8MB、50MB。实验过程中首先利用python库xmltodict转换XML到JSON，该转换流程能针对任意结构的XML以及JSON结构，转换的流程代码如流程1所示；进而利用MessagePack-cli客户端压缩序列化JSON为MessagePack格式，命令格式为：msgpack-cli encode <input-file> [--out=<output-file>]；然后进行网络传输比较，使用java socket传输环境，统计前后时间消耗。表1列出二者传输效率对比，结果表明，基于MessagePack的空间数据传输效率提高幅度很大。

```
//引入三个库
import xmltodict
import json
import sys
//XML到JSON的转换
def pythonConversionXml2Json（pa）:
//打开XML
f = open（pa）
//读取XML到字符串
xml_str = f.read（）
//加载sys库
reload（sys）
//设置utf-8编码
sys.setdefaultencoding（'utf-8'）
//利用xmltodict库转换XML文档到字典
converted_dict = xmltodict.parse （ xml_str,
encoding='utf-8'）
//用json库导出字典为JSON字符串，indent设置输出格式
json_str = json.dumps（converted_dict, indent=1）
//往文件写json字符串
out = open（pa+".json"， 'w+'）
print >>out, json_str.decode（'utf-8'）
//JSON到XML的转换
def python_conversion_json_2_xml（pa）:
//打开JSON
f = open（pa）
//读取JSON到字符串
json_str = f.read（）
//json库加载json字符串
decode_json = json.loads（json_str）
```

```
//加载sys库
reload（sys）
//设置utf-8编码
sys.setdefaultencoding（'utf-8'）
//调用xmltodict库的unparse方法转换预解码JSON到XML字符串
converted_xml = xmltodict.unparse（decode_json）
//写出XML到文档
out = open（pa+".xml"， 'w+'）
print >>out, converted_xml
//调用转换
pythonConversionXml2Json（'Frankfurt_Street_Setting_LOD3.xml'）
```

流程1．通用XMLJSON互转python代码流程

表1 GML文档与MessagePack文档传输效率对比。

数据	Castle_Herten.xml（2.43MB）（GML）	Castle_Herten.xml.json.bin（MessagePack）	Frankfurt_Street_Setting_LOD3.xml（15.8MB）（GML）
历经时间（毫秒）	178	58	740

表1 继续。

数据	Frankfurt_Street_Setting_LOD3.xml.json.bin（MessagePack）	Berlin_Alexanderplatz.xml（50MB）（GML）	Berlin_Alexanderplatz.xml.json.bin（MessagePack）
历经时间（毫秒）	373	1194	594

3．多源数据集成入库的普适性模型框架

目前多源数据处理过程中往往都是针对特定数据的集成展开研究，没有统一的模型。因而，本文在前期研究基础上提出了一种具有普适性的框架模型。

业务数据分为两类：一类是多源的报表数据，特别针对大量的excel报表数据。首先利用FME指定抽取源文件与目的格式文件，可以默认全部属性设置，也可手动设置所需属性，剥离对研究无用属性，实现抽取转换功能实现空间化；然后利用FeatureMerger转换器进行基于关联字段的合并，实现GIS地图与专题报表数据的合并，过滤出所需数据进行准确性检查；接着基于ArcCatalog进行数据库的ArcSDE连接，进而要素类的入库。另一类是网络数据。对源头数据中的网络报表数据，分析网站爬取特点。对只支持单列爬取的网站，需要多次利用爬取工具，打开网页创建列表循环以处理一组元素来实现循环数据采集，并且自定义字段名称，进而导出数据到SqlServer数据库中；对可支持多列爬取的数据进行多列同时爬取情况，只需一次创建循环列表来爬取入库。不过在进行循环采集的时候需要设置循环采集的基本单位为一整行，才能够实现正确多列爬取；同时需在数据库利用select into语句实现多个单列采集表的合并。然而，无论单列还是多列采集都需要利用FME 的SQLCreator半开放转换器实现数据导出，并抽取转换空间化，进而利用上述方式借助FeatureMerger转换器进行要素合并，借助ArcCatalog要素入库。

对需要处理的非GIS测绘类CAD数据，首先将原始数据导入ArcMap，着重对Annotation、point、polyline三种类型进行处理。针对三种类型基于Layer属性，利用select by Attribute工具实现数据分离，无用数据或者研究中不涉及数据可忽略不处理。分离之后要进行数据完整性准确性检查，无需构面的polyline类型以及点类型，可直接右键分别导出为临时Polyline和Point型shp要素类格式，注记类型则须利用Feature to point 工具实现导出。利用feature to polygon工具实现某些polyline 到polygon的线构面，并且检查构面完整性，如出现遗漏进行手动接边处理，再重复以上构面过程。利用sql语句“select LayerName from C where A and B”and条件连接的方式实现界址线的处理，最后对某些polygon需要利用join工具基于空间位置实现注记扩展属性在构面研究对象上的附加利用。经过以上过程所有对象被转换成了要素类，利用ArcCatalog按照上述入库步骤进行shp要素类相应的入库操作，实现数据空间属性数据统一入库管理。通过以上模板，实现了最终的多源数据集成入库。

对需要共享的各国国内外地理数据，优先选择为FME安装插件MyFME，因其对中国国产数据的完美支持，只需正确将中国国内数据的数据结构手动映射到国外的数据结构，即可简单完成工作空间的创建，比如SuperMap中的point映射为Arcgis的point, line准换为polyline, region相应的对应polygon类别，同样可以集成比如mapgis的数据。这样就为各国内外数据共享打通了通路，后期只是专注于数据的配准，使用。

WEBGIS系统的开发需要考虑用户体验的问题，而用户体验体现在数据的请求速度上，前期研究得出了基于GML统一格式进行多源数据的集成，进而进行网络传输，这确实为web开发提供了一种方式。但是空间大数据的爆炸让这种方式在效率上受到了一定的影响，本文提出一种GML格式的优化-利用MessagePack来实现GML格式的压缩序列化存储，进而网络传输，使用时仅需要反序列化解压即可。面对大数据量GML文档的时候，利用python库xmltodict转换XML到JSON，进而利用MessagePack-cli客户端压缩序列化JSON为MessagePack格式，进而网络编程传输，编程方式比如socket, netty，多线程均可，目的系统收到后，同样利用python库xmltodict unparse进而恢复GML，进而进行后续的svg或者其他方式可视化，编辑与分析，构建可用的系统。

4．结论

文章利用FME、arcgis系列工具和爬取等工具结合具体应用研究了专题数据空间化以及依附地图问题、网络数据爬取集成入库问题、CAD数据无损集成到GIS问题，给

出了详细的解决方案，实现多源异构数据的集成；利用MyFME插件研究了国产数据与国外异构数据转化问题；并提出利用MessagePack替代XML亦或者JSON，成为更轻量级的统一集成的格式，更加方便地应用于web的数据快速传输。文章最后给出了具有通用性的模型处理方法，具有实际可操作性，为后续开发提供了方便。

致谢

基金项目：国家自然科学基金项目：退耕驱动的黄土丘陵区人-地响应机制及模型研究（41271518）。

参考文献

[1] Gantz J, Reinsel D (2011) Extracting value from chaos. IDC iView, pp 1–12.

[2] Min Chen, Shiwen Mao, Yunhao Liu. Big Data: A Survey [J]. Mobile Netw Appl, 2014, 19: 171–209.

[3] Ahmed R, DeSmedt P, Kent W, et al. Pegasus: A system for seamless integration of heterogeneous information sources [C]. Compcon Spring'91. Digest of Papers. IEEE, 1991: 128-136.

[4] Attaluri G K, Bradshaw D P, Coburn N, et al. The CORDS multidatabase project [J]. IBM Systems Journal, 1995, 34(1): 39-62.

[5] Li L, kumar Nalluri A, Ai L. Space-Aware Data Integration for Ocean Observing Systems [J]. Procedia Environmental Sciences, 2011, 11: 285-290.

[6] Belussi A, Migliorini S. A framework for integrating multi-accuracy spatial data in geographical applications [J]. GeoInformatica, 2012, 16(3): 523-561.

[7] 程海军，胡圣武，张子平．GIS数据格式集成方法的探讨[J]. 河南理工大学学报：自然科学版， 2006, 25(1): 37-41。

[8] 张文江．地质灾害数据集成关键技术研究[D]．成都：成都理工大学，2013。

[9] Bansal S K. Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration [C]. Big Data (BigData Congress), 2014 IEEE International Congress on. IEEE, 2014: 522-529.

[10] 卢一枝，陈军华．Deep Web 数据库集成技术的研究[J]．上海师范大学学报(自然科学版)，2016, 45(4): 422-427。

[11] 陈惠敏，胡飞虎，耿泽飞等．基于GIS的灾害应急管理系统业务数据和空间数据的集成[J]. 自然灾害学报，2011, 20(1): 163-167。

[12] 董浩然，谢欢，陈鹏等．基于GIS主题爬虫的在线房产估价系统与优化[J]．地理信息世界．2016, 23(2), 107-112。

[13] 于一男，关佶红，周水庚等．GSPress：一个GML流压缩器[J]．小型微型计算机系统．2011, 32(3): 397-401。

[14] 崔玉洁，廖坤．借助八爪鱼采集器实现过刊网刊元数据的自动提取[J]．编辑学报．2016, 28(5), 485-487。

[15] 冯文娟．基于FME实现CAD到GIS数据格式转换研究及实例[J]．煤矿现代化．2015(3), 98-100。

[16] 郑江，闫世浩，陈宝枝等．基于DWGdirect.NET与Arcgis Engine的CAD地形图转换[J]．测绘与空间地理信息．2015, 38(2), 176-178。

[17] 李伙友，吴善和．DWG与SHP数据格式互转换方法研究[J]．集美大学学报（自然科学版). 2015, 20(1), 76-80。

[18] 于艳超，许捍卫，杜婵娟．基于DWGDirect的CAD到GIS数据转换研究[J]．地理空间信息．2015, 13(1), 84-86。

[19] myfme的博客．用MyFME扩展你的ArcGIS应用[EB/OL]. http://blog.sina.com.cn/s/blog_7cfdb4de0100uyw1.html. 2011-10-13。

作者简介

徐强，男，山东省泰安市新泰市人，硕士研究生，主要研究方向为空间数据集成，移动gis开发。

范虹，女，副教授，硕士生导师，主要从事空间数据挖掘研究。