Institute for Advanced Sustainability Studies (IASS)

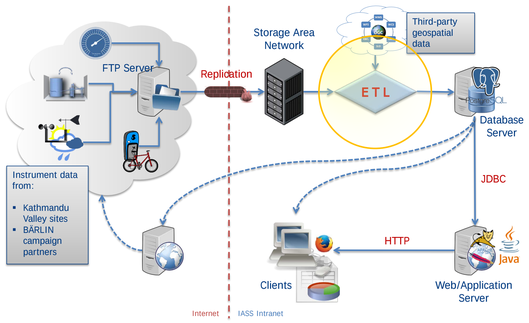
Investigation and evaluation of the open source ETL tools GeoKettle and Talend Open Studio in terms of their ability to process spatial data

Kristin Kuhnert[1,2] and Jörn Quedenau[1]

[1] Institute for Advanced Sustainability Studies (IASS), Potsdam, Germany, [2] University of Potsdam

Background

Raw instrument data from field campaigns on air quality (BAERLIN2014[1], SusKat[2]) are typically provided in a wide range of file formats and at different levels of quality. To ensure a consistent data basis for data mining in a spatial context, web-based reports etc., we have integrated all experimental data as well as additional third-party georeferenced data in a relational database using open source ETL tools.

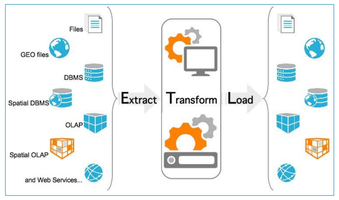


Data sources	Format
Bicycle, mobile van and stationary measurements	
GRIMM and DiSCmini particle counter (PM1, PM2.5, PM10), rel. humidity, temperature	CSV
GPS tracks	GPX
Video camera	MPEG-4
Different gases (CO_2, CO, NO_x, O_3 CH_4), particle number of different sizes, temperature, rel. humidity	NetCDF
Automatic weather stations	Excel
Third-party spatial data	
Senatsverwaltung Berlin (road categories, vegetation, building heights etc.)	Shape file, WFS
CORINE land cover	GeoTIFF

The ETL Process

Extract

- Connect to various data sources
- Extract data from these data sources



Source: DB Best Chronicles[3]

Transform

- Any function applied to the extracted data between the extraction from sources and loading into targets
- Operations are for instance: movement of data, validation, modification, integration, calculation

Load

- All processing required to load the data into a target system

ETL Tools

- Type of software used to construct and execute ETL processes
- Automation of complex and repetitive data processing without producing any specific code
- Clear graphical presentation of the transformation steps
- Provide tools for identifying errors and analyzing performance

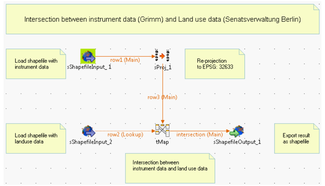
Open source ETL tools for processing spatial data

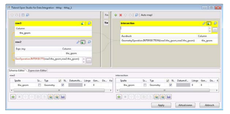
Talend Open Studio (TOS-DI)[4] with the **Spatial Data Integrator** extension (SDI)[5]	**GeoKettle** (spatially-enabled version of the Pentaho Data Integration Suite)[6]

Use Cases

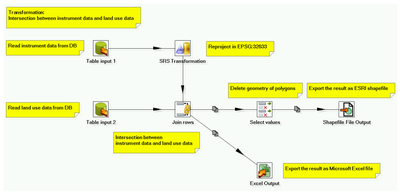
1. Intersection between two spatial datasets

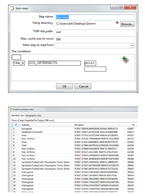
a) TOS-DI





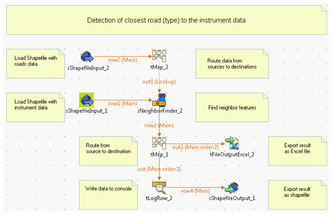
b) GeoKettle

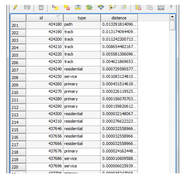




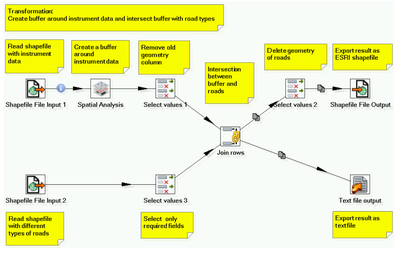
2. Neighbor finder

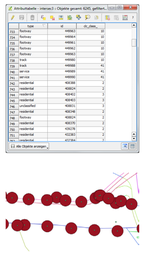
a) TOS-DI





b) GeoKettle





Results

	TOS-DI (with SDI)[4,5]	GeoKettle[6]
License	Apache License 2.0	LGPL
Software architecture	Code generator[a], Eclipse-based	Metadata driven interpreter
Avail. components	~750	~200
Processing spatial data		
Vector data	⊕⊕	⊕⊕
Raster data	⊖[b] (support discontinued)	⊖[b]
Extensibility & customization		
Custom components (Java API)	✓	✓
Script components	Java, Groovy	Java, JavaScript, R
View and modify generated code	✓	×
Metadata		
User-defined metadata	⊕⊕	⊕
Shared repository	×	✓[c]
Version control	✓	×
Miscellaneous		
Non-spatial ETL	⊕⊕	⊕
ELT support	⊕	⊖
Scheduler	✓	✓
Other pros	Integrated debugger, slightly better performance	Per-component data preview, flat learning curve

a) Java (Perl support discontinued) b) with Sextante module, c) relational database

As there is no clear favorite, we decided to use both tools for our in-house data management, depending on the specific use case. While TOS-DI provides a wider range of features, GeoKettle's shared repository is of great value for collaborative work. Both tools still lack sufficient raster data support.

References

[1] Bonn, B., von Schneidemesser, E., Andrich, D., Quedenau, J., Gerwig, H., Lüdecke, A., Kura, J., Pietsch, A., Ehlers, C., Klemp, D., Kofahl, C., Nothard, R., Kerschbaumer, A., Junkermann, W., Grote, R., Pohl, T., Weber, K., Lode, B., Schönberger, P., Churkina, G., Butler, T. M., and Lawrence, M. G.: BAERLIN2014 – The influence of land surface types on and the horizontal heterogeneity of air pollutant levels in Berlin, Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2016-57, in review, 2016.

[2] Rupakheti, M., Panday, A.K., Lawrence, M.G., Kim, S.W., Sinha, V., Kang, S.C., Naja, M., Park, J. S., Hoor, P., Holben, B., Sharma, R.K., Mues, A., Mahata, K S., Bhardwaj, P., Sarkar, C., Rupakheti, D., Regmi, R.P., and Gustafsson, Ö.: Air pollution in the Himalayan foothills:
overview of the SusKat-ABC international air pollution measurement campaign in Nepal, Atmos. Chem. Phys. Discuss., in preparation, 2016.

[3] DB Best Chronicles: Talks on Big Data, Mobile Apps, Web and Software Development (https://www.dbbest.com/blog/)
[4] https://sourceforge.net/projects/talend-studio/
[5] https://talend-spatial.github.io/
[6] http://www.spatialytics.org/projects/geokettle/

The IASS is sponsored by

SPONSORED BY THE Federal Ministry of Education and Research
FONA Research for Sustainable Development BMBF
LAND BRANDENBURG Ministerium für Wissenschaft, Forschung und Kultur