The Plan4business Approach to Transfer Open Data into Real Estate Businesses

Jan Ježek, Tomáš Mildorf, Karel Charvát Jr., Karel Charvát

The Plan4business Approach to Transfer Open Data Into Real Estate Businesses

Jan Ježek[1], Tomáš Mildorf[1], Karel Charvát Jr.[2], and Karel Charvát[3]

[1] University of West Bohemia, Pilsen, Czech Republic
{jezekjan, mildorf}@kma.zcu.cz
[2] University of Economics, Prague, Czech Republic
xchak07@vse.cz
[3] Help Service - Remote Sensing, Benešov, Czech Republic
kch@hsrs.cz

Abstract. Spatial planning data including urban, regional, spatial or zoning plans are not aggregated so far. Creating time series or comparative analyses on these data sets is not yet possible. The EU funded project Plan4business develops a service platform that can serve to users as a full catalogue of spatial planning data linked with other data sources such as statistics, OpenStreetMap, Urban Atlas and Corine Land Cover that are published as Open Data. The Plan4business platform will offer to clients not only the data itself in an integrated, harmonised and thus ready-to-use form, but also rich analysis and visualisation services via an API and an interactive web frontend. The users include mainly citizens, local authorities and real estate agencies. This paper introduces the problems of data integration and selected technical components of the Plan4business platform supporting data reuse and analysis.

Keywords: plan4business, open data, spatial planning data, land use, integration, harmonisation, analyses, real estate businesses

1 Introduction

In the last decade, many changes have happened with regard to data provision and data management in both public and commercial sectors. The PSI Directive (Directive 2003/98/EC on the re-use of public sector information) adopted in 2003 triggered a lot of debates about how public information should be handled, shared and priced.

Information, as a meaningful interpretation of data and relations between them [7], plays a crucial role in decision making mechanisms in different domains, such as environment or spatial planning, and on various levels, from local to global. The underlying data for decision making processes represent the most expensive part of information systems. It is therefore important to ensure that:

- data are collected once and maintained at the level where this can be done most effectively;

adfa, p. 1, 2011.

- it is possible to combine data from different sources and share them between many users and applications;
- data are collected at one level of government and shared between all levels;
- data needed for good governance are available on conditions that are not restricting their extensive use;
- it is easy to discover which data are available, to evaluate their fitness for purpose and to know which conditions apply for their use. [2]

The above mentioned rules are the key principles of the INSPIRE (Infrastructure for Spatial Information in the European Community) initiative aiming to create a European Spatial Data Infrastructure (ESDI). This initiative is supported by the legal framework of the INSPIRE Directive that came into force in May 2007.

Because of different historical developments in countries, heterogeneous data coming from various providers are not easy to combine, share and analyse. The reuse of data is therefore limited. Spatial data, their integration and analysis are the main topics of this paper. The authors present the technological issues of a service platform for data and information reuse that is being developed within the frame of the FP7 project Plan4business (04/2012 – 03/2014).

The next chapter gives and insight into the problem of data interoperability and examples of initiatives aiming to data sharing in an interoperable manner.

2 Interoperability of Spatial Planning Data

Spatial data, sometimes referred to as geographic data, geodata or geospatial data, are defined by INSPIRE as “data with a direct or indirect reference to a specific location or geographic area.” [1] The possibility to locate objects and phenomena is an opportunity to interpret, analyse and visualise data in an understandable way familiar to humans. Spatial data create a reference layer capable of linking other non-spatial data, such as statistical data. The relations between objects and phenomena can be then interpreted and analysed in the context of spatial information.

As it was already mentioned, data coming from various countries and data providers are heterogeneous, unless they comply with semantic and technical specifications, and legal and organisational frameworks agreed between all parties involved. The application of common specifications leads to certain level of data interoperability.

The term interoperability is defined by the International Organisation for Standardization (ISO) as the "capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units." [3] As stated by David Schell, founder and chairman of the Open Geospatial Consortium (OGC), “Interoperability seems to be about the integration of information. What it’s really about is the coordination of organizational behaviour.” [6]

Examples of particular issues categorised according to the European Interoperability Framework are showed in Figure 1. Heterogeneity of spatial data was well documented by the European Plan4all project that collected data specifications from most of the European countries. The Plan4all results show the complexity of the situation.

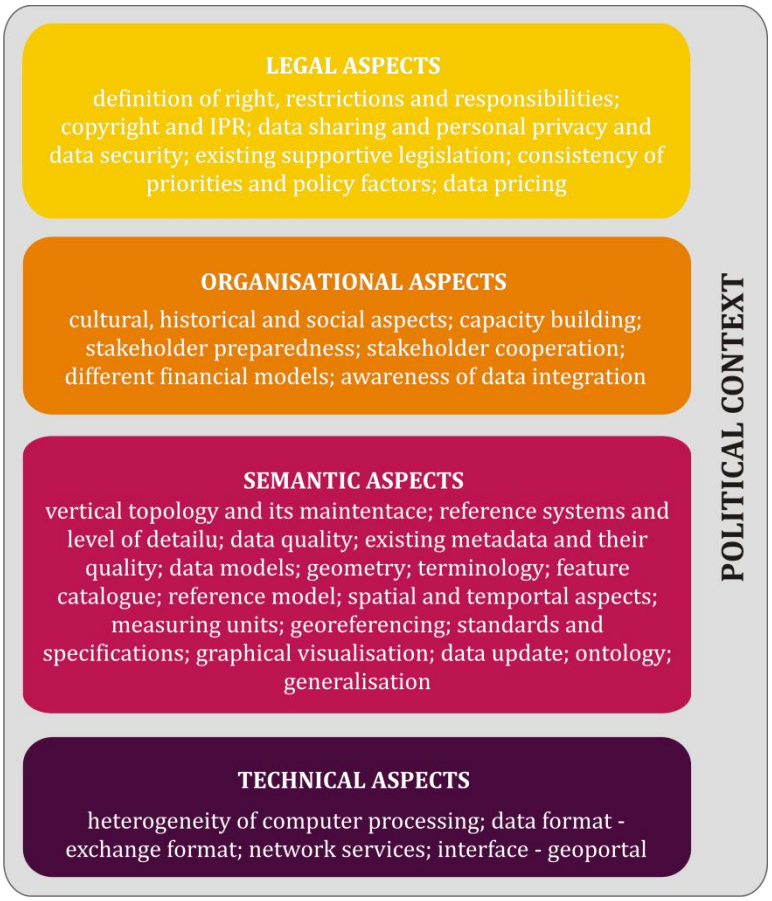

Fig. 1. Aspects of data interoperability [5]

There are several initiatives in Europe aiming to bridge the differences between datasets from various sources by providing legal and organisational frameworks as well as semantic and technical specifications. The already mentioned INSPIRE Directive is the most important initiative as far as spatial data are concerned. Other initiatives include the Shared Environmental Information System (SEIS), Copernicus (previously known as Global Monitoring for Environment and Security - GMES) or the Global Earth Observation System of Systems (GEOSS).

The Plan4business project aims to provide a technical solution for data integration based on the INSPIRE principles, OGC specifications and ISO (International Organization for Standardization) standards of the 19100 series Geographic Information.

Plan4business focuses on integration of spatial planning data from various sources. Spatial planning data represent a valuable source of information that could be used not only for spatial planning activities but also as a reference layer for e.g. real estate agencies, providing tools for data analysis across a certain geographical area where differences between administrative units are diminished. Currently, it is difficult to

use results of spatial planning for any other purposes than for printing or simple publishing by the authorities that create them. Creating time series, comparative analyses or data mining on these data sets is not yet possible. However, such capabilities are needed by researchers, spatial planners and professionals from the real estate world.

3 Architecture of the Plan4business Service Platform

3.1 General Architecture

Plan4business develops a platform that can serve to users as a full catalogue of planning data such as transport infrastructure, regional plans, urban plans and zoning plans in relation with other Open Data including statistics, OpenStreetMap, Corine Land Cover or Urban Atlas. The Plan4business platform offers to clients not only the data itself in an integrated, harmonised and thus ready-to-use form, but also offers rich analysis and visualisation services via an API and an interactive web frontend.

The Plan4business platform consists of several different technical components grouped into three layers. The first layer contains human-machine interfaces for planning data management, integration and conversion as well as for accessing the analytical functions of the platform. The second layer provides two groups of processing engines, again for integration and harmonisation on the one hand and for analytical processing on the other hand. In addition, this layer provides the Plan4business application programming interface (API). The final layer is a storage layer, which contains a storage manager and two different data bases, each optimised for different goals. All the three layers, their components and relations are depicted in Figure 2.

The design and implementation of the platform are partly based on components developed in previous projects and by various initiatives. There are three central components of the server side of the Plan4business platform including Integration, Storage and Analysis Engines (see Figure 3). These components are described in more detail in the next sections.

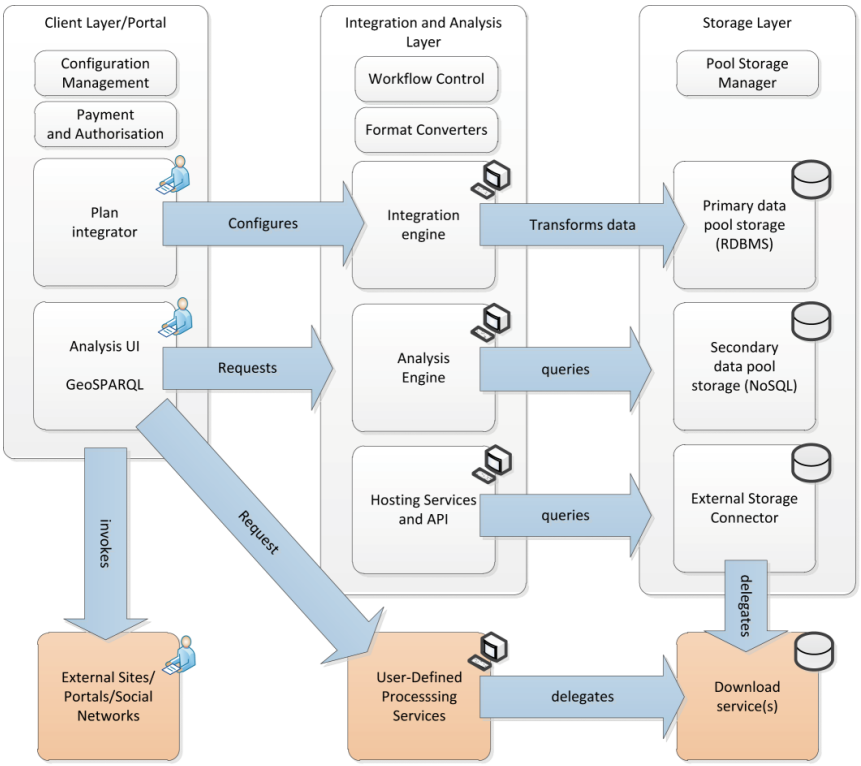

Fig. 2. General architecture of the Plan4business platform [8]

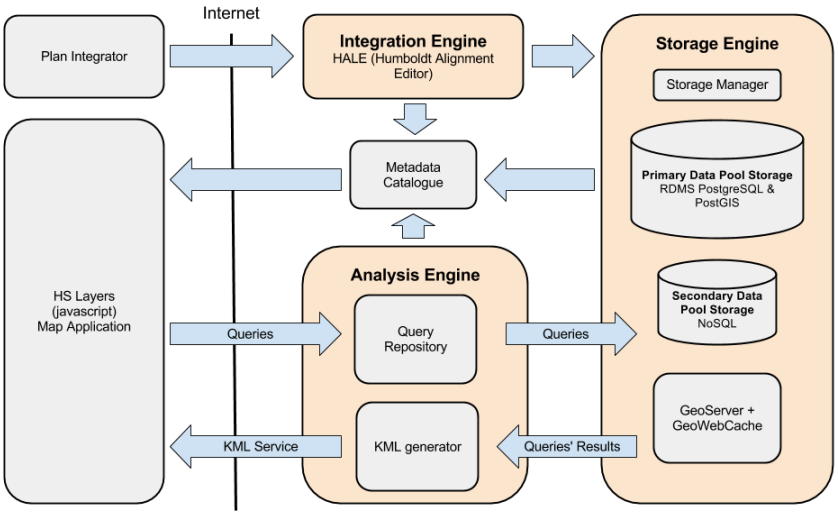

Fig. 3. Main components of the server side of the Plan4business platform [4]

The central part of the components is a metadata catalogue. The catalogue enables to store not only metadata about existing datasets, but also about performed analyses, created map compositions and predefined integration services.

3.2 Integration Engine

The Integration Engine is intended to perform all data transformations necessary to seamlessly integrate heterogeneously structured and externally provided spatial vector data into the Plan4business data pool. Based on the input data and schema alignment provided by the user through the platform interface (Plan Integrator in Figure 3), the Integration Engine, deployed on the server side of the Plan4business platform, will transforms input data into the schema used by the Storage Engine.

Various conversions and transformation steps such as mapping between different data formats, geometric representations and conceptual schemas will be performed by the Integration Engine. This functionality is being implemented on top of the Humboldt Alignment Editor (HALE), a desktop software that was initiated within the HUMBOLDT project (FP6) and is continuously developed as an open source software.

The Integration Engine will enable several data transformation processes that can be initiated by different users at the same time. Furthermore, the transformed data will not be provided as an output file within the file system but will be directly saved in a database to enable concurrent data retrieval, analysis or extension in a resource-efficient and consistent way. [4]

3.3 Storage Engine

The Storage Engine consists of four subcomponents. The first subcomponent is the Primary Data Pool Storage. This is a hybrid combination of a relational database and a document-oriented storage. This storage provides access to the original (non-harmonised) spatial data sets and to the integrated spatial vector data sets.

The second subcomponent is the Secondary Data Pool Storage which is intended to augment the capabilities that are offered by the Primary Data Pool Storage. While the relational database management system used for the primary storage can be expected to scale well with a growing amount of data and to provide fast data retrieval, the choice of the database management system for the secondary storage needs to be optimised for complex graph-based queries or complex analysis. This can be most likely achieved by using a NoSQL database that breaks with common design decisions of RDBMS.

The third subcomponent of the Storage Engine is the Storage Manager that enables to control the data that are available in the Primary and Secondary Data Pool Storages. The Storage Manager can be used to remove data from the primary data pool or to copy data that is needed for a certain analysis from the Primary Data Pool Storage into the Secondary Data Pool Storage. [4]

The fourth subcomponent supports spatial data publication. The main software solutions include GeoServer, a Java software server that allows users to view and edit

geospatial data, and GeoWebCache, a Java web application used to cache map tiles coming from a variety of sources such as OGC Web Map Service (WMS).

3.4 Analysis Engine

The Analysis Engine which is closely related to the Storage Engine ensures processing of analytical queries made by the platform user and retrieval of their results. It enables accessing and processing spatial planning data and their retrieval for visualisation in the map client. The Analysis Engine provides access to all analytical functions of the spatial databases, i.e. the Primary and Secondary Data Pool Storages.

For debugging and testing purposes a web application for performing predefined and user defined queries and displaying query results without any major configuration was implemented (Figure 4).

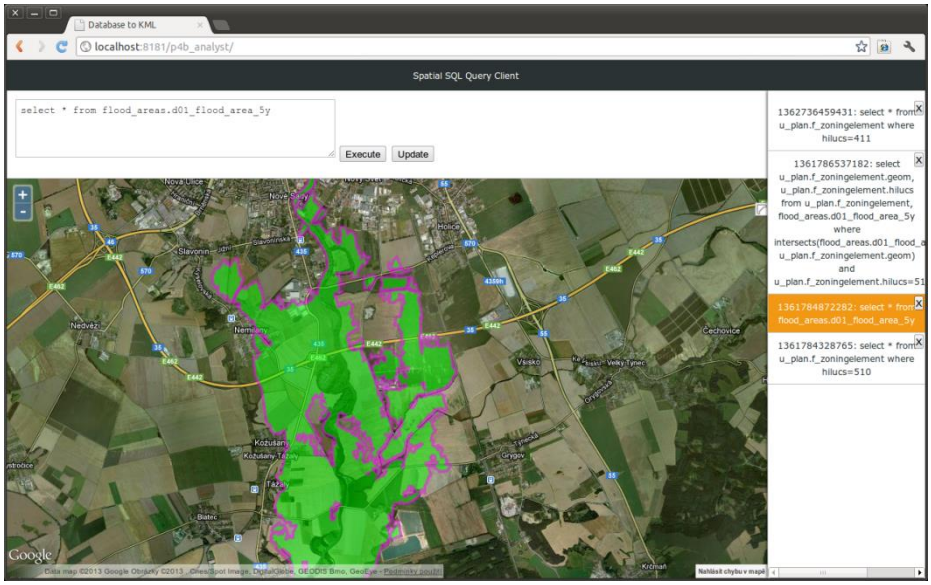

Fig. 4. A web application for querying the Plan4business database and displaying the results

The results of the analysis are in the KML format that can be visualised in any map client supporting KML such as the Plan4business platform or Google Earth. This feature will ensure interoperability with other solutions and platforms.

4 Czech Pilot Combining Statistical Data

Spatial planning data represent a valuable source of information. In order to fully exploit their potential, it is necessary to combine them with other (non-spatial) data such as socioeconomic or demographic data. The combination is an added value

which can be beneficial for end users performing analysis on the Plan4business platform.

Plan4business tested the possibility of inclusion of statistical data from EUROSTAT on European, national and regional levels and from national statistical offices on local level. The spatial coverage of the testing included the territory of the Czech Republic.

EUROSTAT publishes data up to NUTS[1] 3 level. NUTS 1 level includes major socioeconomic regions, NUTS 2 level consists of basic regions for the application of regional policies and NUTS 3 level contains small regions for specific diagnoses.

In order to make the analyses through the Plan4business platform interesting for real estate agencies, investors and other user groups, the information from EUROSTAT was complemented by national data from LAU[2] 1 and LAU 2 levels (formerly NUTS 4 and NUTS 5 levels). LAU 1 level includes tens or hundreds of municipalities which are represented as LAU 2 level.

The structure of data published by national statistical offices is not standardised and differs in countries. The data in the Czech Republic were acquired from the Czech Statistical Office (CZSO). The statistical data of CZSO are not published as Open Data but through a Public Database[3]. The Public Database was developed for presentation (viewing) of statistical data mainly for the public where data are presented as a set of predefined tables that vary depending on the type of statistical data. This makes automatic machine collection and processing of these data rather difficult. In the Czech Republic there are currently 6253 municipalities. Obtaining statistical data for each municipality means downloading and processing 6253 files.

The data from individual tables can be exported to XML files. The problem is that the XML files are not accessible through a URL and the files can't be simply generated in a bulk. Manual downloading of such a number of files is practically impossible. Fortunately, every page containing tables with statistical data is accessible through a unique URL which can be generated using codes of municipalities from the CZSO code list. A script for downloading and processing statistical data was created and for each type of statistics the script was modified according to the structure of the data on the HTML pages. This process is rather time consuming and prevents automation of data acquisition.

5 Conclusions

The development of the Plan4business service platform goes in line with the collection and analysis of user requirements and creation of a business plan that ensures the sustainability of the platform after the end of the project.

The project consortium is currently facing challenging problems of data availability and smooth access to data in a machine readable way. Due to different national legislation and other barriers hindering access to data on the pan-European level, the

[1] Nomenclature of territorial units for statistics

[2] Local Administrative Units

[3] http://vdb.czso.cz/vdbvo/en/uvod.jsp

platform will be tested in selected pilot areas. The first one is the Czech Republic where access to statistical, cadastral, topographic and environmental data enables to test the platform functionality in real world scenarios. As described in this paper, the process of data acquisition shown on the example of statistical data is time consuming and cannot be automated.

The primary data sources for the Czech pilot will be data from public administration and governmental bodies. The data will be harmonised and integrated into the Plan4business platform and a set of services will be offered to the main target groups including real estate agencies, citizens and municipalities.

Acknowledgments. All the authors are supported by the FP7 project Plan4business (Grant agreement no: 296282). The authors from the University of West Bohemia are additionally supported by the European Regional Development Fund (ERDF), project "NTIS – New Technologies for the Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090.

References

1. European Commission: INSPIRE Glossary, http://inspire-registry.jrc.ec.europa.eu/registers/GLOSSARY
2. INSPIRE: INSPIRE website, http://inspire.ec.europa.eu/ (2012)
3. International Organization for Standardization: ISO/IEC 2382-1 Information technology -- Vocabulary -- Part 1: Fundamental terms (1993)
4. Ježek, J. et al.: D5.1 Interim Report on Integration, Analysis and Storage Engines, plan4business consortium (2013)
5. Mildorf, T.: Modelová generalizace pozemkového datového modelu (Model Generalisation of a Land Data Model). University of West Bohemia, Pilsen (2012)
6. Open Geospatial Consortium: Open Geospatial Consortium website, http://www.opengeospatial.org [Accessed February 18, 2012]
7. Rapant, P. ed.: Terminologický výkladový slovník pojmů z oblasti geoinformací. Věstník Úřadu pro veřejné informační systémy, II, (2001)
8. Fraunhofer-Gesellschaft zur Foerderung der Angewangten Forschung E. V.: Seventh Framework Programme, Grant Agreement No 296282 plan4business - A service platform for aggregation, processing and analysis of urban and regional planning data, Annex I - Description of Work (2012)