🏆 *reviewed paper*

Information Around Us: Questions Connected to Information and Data Heterogeneities in Planning Activities

Otakar Cerba

(Ing. et Mgr. Otakar Cerba, Ph.D, University of West Bohemia, Univerzitni 22, 30614 Plzen, Czech Republic, cerba@kma.zcu.cz)

1 ABSTRACT

Information, data and knowledge represent the most interesting and desired resources and commodities in the world. It relates also to spatial planning and urban planning activities as well as to development of cities. These processes are strongly connected to spatial plans but also to another types of supporting information and data such as Volunteered Geographic Information, social media, data provided by local authorities or by business subjects and last but not least information, knowledge and experience of particular users).

We (as a data experts) suppose that there is enough various information and data resources to satisfy all needs and requirements of participants of above-mentioned activities (planners, decision-makers, researchers, visitors, business subjects and citizens). In contrast to our opinion they feel a lack of information and data. Why can we see totally different views?

There are several reasons. Data sets are provided as self-standing products without a chance to interconnect independent data from various resources. All involved subjects want and need to have data and information, but they are not able to find them, because they are provided in very complicated and unintelligible form. For example the metadata and data catalogues based on metadata are mainly focused on experts or computers and not on common users. Similarly the conditions of using (licences, digital rights management) are also connected to expert terminology and usage. Users are confused by a quality and reliability of data and information (e.g. the latest problems related to a quality of some map products in mobile navigation systems).

Current research in the field of processing and distribution of data and information is focused on technological aspects (e.g. implementation of standards or data models). But typical users do not need sophisticated and complicated solutions, but intelligible and understandable products. For example they do not use the large possibilities of connection of web services and support of standards. They need to have connected useful and reliable data and they are not interesting in processes and technologies running in the background.

The low attractiveness is the latest factor of "inaccessibility" of information and data connected to spatial planning and urban development. Except above-mentioned orientation to technical solution it means ways of presentation and publication. Information and data producers prefer raw data (tables, database queries) instead of processed data and information (maps, graphs, infographics) connected with approaches of data journalism.

Goals of this paper are not only to show general questions related to information and data exploitation in the contemporary cities (that are indicated in previous paragraphs). We would like to present simple case study indicating some solutions. The case study covers a proposal of list of necessary harmonization and integration steps and examples of data integration and presentation to provide local information for citizens.

2 INTRODUCTION

Information, data and knowledge with spatial component represent the most interesting and desired resources and commodities in all levels of spatial planning and urban planning activities as well as in development processes of regions, cities and municipalities. These processes are strongly connected to spatial plans but also to another types of supporting information and data, including Volunteered Geographic Information (VGI), crowdsourcing activities and data provided as results of international of global activities.

We (as a spatial data experts) suppose that there is enough various information and data resources to satisfy all needs and requirements of participants of above-mentioned activities (planners, decision-makers, researchers, visitors, business subjects and citizens). In contrast to our opinion they feel a lack of information and data.

The heterogeneity of data is one and very important reason. The research in projects such as Humboldt, Plan4all, Habitats or Plan4business have been focused on technological aspects of processing and

distribution of data and information (e.g. implementation of standards, formats conversion, INSPIRE complience or data models). The main goals of this paper are not only to show general questions related to information and data exploitation in the contemporary cities and regions (they are mentioned for instance in Cerba et al., 2012a and Cerba et al., 2012b). We would like to present simple case study based on interconnection of three data sets indicating some solutions. The case study covers data description, a proposal of list of necessary harmonization and integration steps and examples of data integration and presentation to provide local information for citizens.

This paper is composed of four parts. The first (Introduction) is focused on presentation of different view of amount of accessible spatial data. In the second section (Data Sets) the three data sets mentioned in next part are introduced, including their properties important from the view of data harmonization. The third section ()Harmonization describes selected case study of harmonization of spatial data of themes "land use" and "land cover". The last part (Conclusion) summarizes results of harmonization and proposes new steps that can improve the harmonization effect and usability of data.

3 DATA SETS

Our case study developed in the Plan4business[1] and Habitats[2] projects has worked with three main data sets. All these data are free accessible and could be used for collecting information about "land use" and "land cover". Therefore they are important from the view of planning activities above all in regional level or in cross-border areas without any other consistent data set.

	CLC (version 2006)	UA	OSM
Coverage	EU states + European Free Trade Association (EFTA) + states Turkey,	EU Large Urban Zones with more than 100.000 inhabitants	Global
Theme	Land use / land cover	Land use / land cover	All spatial features
Provider	European Environment Agency	European Environment Agency	OpenStreetMap Foundation
Geometry	Areas (in vector version)	Areas	Points (nodes), lines (ways), areas(or polygons as closed ways)
Nomenclature	CLC nomenclature; three-levels hierarchy containing code numbers and explanations	CLC nomenclature; four-levels hierarchy containing code numbers and explanations	Tag containing key (type of element) and value (detail specification of key)
Original format	GeoTIFF	ESRI shapefile	.osm (XML based format)
Granularity	100 m resolution	100 times higher than CORINE land cover	Depend on each contributor
Updating / Last version	Temporal coverge: 2006; Last upload 7 May 2012	Temporal coverge: 2005-2007; Last upload 28 May 2010	Every day (but not systematic, it depends on an activity of particular contributors)
Download	http://www.eea.europa.eu/data-and-maps; including seamless vector versions in ESRI shapefile format	http://www.eea.europa.eu/data-and-maps; ZIP archives containing ESRI shapefile data	http://wiki.openstreetmap.org/wiki/Download; including links to external sources containing ESRI shapefile data
Licence	EEA standard re-use policy: unless otherwise indicated, re-use of content on the EEA website for commercial or non-commercial purposes is permitted free of charge	EEA standard re-use policy: unless otherwise indicated, re-use of content on the EEA website for commercial or non-commercial purposes is permitted free of charge	Open Data Commons Open Database License (data); Creative Commons CC-BY-SA (maps)
Level of Open Data (5 star deployment scheme for Open Data)	**	**	***

Table 1: Comparison of the main attributes of CLC, UA and OSM.

CORINE Land Cover (CLC; CORINE land cover, 1995) is probably the most known "land cover" (as well as "land use") data set and classification system in Europe. The project CORINE (Coordination of Information on the Environment) was realized in Europe in years 1985-1990. It was focused on collecting, coordination and ensuring of high-quality information on environment and natural resources. The regular updated database is managed by EEA (European Environmental Agency). The CLC nomenclature is composed of 44 classes that are divided into three levels. The CLC system was extended in many another

[1] http://www.plan4business.eu
[2] http://www.inspiredhabitats.eu

projects – in the paper (Feranec & Oťahel, 2004) is described the 4th level of CLC and the document (Kopecká, 2006) mentions the 5th level at that.

Urban Atlas (UA) data set provides reliable, inter-comparable, high-resolution land use maps for 305 Large Urban Zones and their surroundings (more than 100.000 inhabitants) for the reference year 2006. It is published by EEA. Urban Atlas uses the classification that is very similar to CLC, but in some categories (e.g. natural landscapes) is simplified, while categories describing artificial surfaces are more detailed. (the paragraph is based on Meirich, 2008 and Urban Atlas web page)

OpenStreetMap (OSM; http://www.openstreetmap.org) represents one of the most typical example of neogegraphy and Volunteered geographic information. The data are not focused on "land use" and "land cover" in the first place. But OSM contains many elements related to both above-mentioned themes (e.g. forest, waters). More information about OSM and another project based on OSM are found in the web page of the project or in the paper Hackly & Weber, 2008.

The table 1 summarizes main attributes (coverage, type of geometry, formats or links to download) of above-mentioned data sets. The last row contains the classification of data sets according five star deployment scheme for Open Data suggested by Tim Barners-Lee. Two stars means using of open licence (we consider the EEA rules as "open") and providing structured-data. The third star label using of non-proprietary format.

4 HARMONIZATION (UA A CLC A OSM)

The harmonization of spatial data of CLC, UA and OSM data sets is based on general principles of spatial data harmonization (see Cerba et al., 2012a, Cerba et al., 2012b). They come out of the statement that data represents just a limited model of real world and therefore the efficient spatial data harmonization has to be composed of particular harmonization steps that process particular properties of data such as spatial reference systems, portrayal rules or data models (an overview possible harmonization steps in figure Data harmonization components in D2.5, 2007).

All harmonization steps follow the first main and the most important task – why do we need harmonize? There are several reasons for the case study presented in this paper. CLC is complete (it covers most of Europe). But it is not very detailed and frequently updated. UA is more detailed but it provides isolated data (not complete coverage of interest area, but only densely populated regions). Similarly to CLC the UA is quite old. OSM is regularly updated, but this data set has with different quality (including low granularity and missing elements). Exploitation of benefits of all three data sets represents the real reason for harmonization. Users should be able to find the best information about "land use" and "land cover" themes without regard to original data.

In the next paragraphs we follow selected harmonization steps that are important from the point of view of harmonization of CLC, UA and OSM (above all information about "land use" and "land cover" as a reference data of spatial planning). Harmonization of conditions of usage arise from licences (see Table 1). The CLC and UA data sets are based on EEA standard re-use policy, that enables re-use of content on the EEA website for commercial or non-commercial purposes is permitted free of charge unless otherwise indicated. OSM is licensed by two different ways – maps by Creative Commons CC-BY-SA (it means users are free to copy, distribute and transmit the work, to adapt the work and to make commercial use of the work; they must attribute the work in the manner specified by the author or licensor and distribute the resulting work only under the same or similar license). The OSM data uses the Open Data Commons Open Database License. It has similar approach – to keep an attribution and share-alike.

All three data sets are connected to European space, but except OSM they do not cover complete Europe. CLC and UA include countries cooperating with EEA, moreover UA does not represent continuous coverage. Therefore the following course of advance of harmonization has been suggested:

(1) To take CLC as fundamental data set.

(2) To add UA objects that are not a part of CLC data.

(3) To replace these CLC objects with UA objects that cover or overlap original CLC data.

(4) To decide if OSM data in particular areas are suitable to an accuracy improvement of combination of CLC a UA data.

(5) To add OSM objects in the same way as UA data (see points 2 and 3). There is established the wiki project transforming CLC data set to OpenStreetMap[3], including nomenclature.[4]

(5) Because the CLC nomenclature is the most detailed, the UA and OSM classification should be transformed to CLC classes.

The process of harmonization will be consisted in following steps:

(1) Development of common data model – a minimal variant has to contain geometry, an id and a nomenclature code (CLC), extended model can include for example original code or another attributes.

(2) Transformation of OSM elements (ways and closed ways) to polygons to have one geometry. This step will be realized by a GIS (Geographic Information System) tool.

(3) Harmonization of classification codes – this activity could be implemented by a matching table. There is also an advanced solution based on ontology (describing particular nomenclatures, their components, properties and relations) and an ETL (Extract-Transform-Load) tool, that enable to transform source data to common data model, including changing classifications. The rules of classification transformation will be derived from ontology (more information about this approach in Cerba, 2011).

Another technical solution of proposed harmonization process could be based on application of spatial databases. This approach is very fitting for large data set, that could be processed very effectively. Databases enable to implement the majority of above-mentioned harmonization steps (e.g. transformation of nomenclatures or changing of data models). More about spatial data harmonization based on databases as well as another approaches to harmonization are published in Cerba et al., 2012a.

5 CONCLUSION

This paper shows the essential principles of spatial data harmonization in the case of integration three data sets (CLC, UA and OSM) that contain information of themes "land use" and "land cover". These types of information are important for many human activities including spatial planning or regional development. These data sets are also illustrative of example of sufficient amount of data that could be felt as apparent shortage because of difficult way of possibilities of data integration and combination.

The real harmonization (presented in this article) represents a sequence of simple harmonization steps depending of spatial data aspects that have to be harmonized. Just a design of the whole harmonization process (including particular sub-steps and their parameters) means the key (and usually underestimated) part of the spatial data harmonization and integration. It is connected to essential question – What is the target and meaning of concrete harmonization?

In the introduction there is mentioned a feeling a lack of information and data by real users and providers of spatial planning activities. The spatial data harmonization (as the main topic of this research) does not represent a final solution of this problem. There are another reasons lack of information and data. (1) Data sets are provided as self-standing products without a chance to interconnect independent data from various resources. (2) All involved subjects want and need to have data and information, but they are not able to find them, because they are provided in very complicated and unintelligible form. For example the metadata and data catalogues based on metadata are mainly focused on experts or computers and not on common users. (3) Similarly the conditions of using (licences, digital rights management) are also connected to expert terminology and usage. These reasons do not represent the scope or this paper, but a subject of following research. (4) Spatial data problems can include also a transformation to KML (Keyhole Markup Language) format instead of using of very complicated GML (Geography Markup Language) format for spatial data storage or (5) links to semantic tools such as controlled vocabularies, gazetteers or thesauruses. The next research in the field of making spatial data more accessible to common users can be focused on an (6) improvement of the low attractiveness which is the last (but not least) factor of "inaccessibility" of information and data connected to spatial planning and urban development. Except above-mentioned orientation to technical solution it means various ways of presentation and publication. Information and data producers prefer raw data (tables, database queries) instead of processed data and information (maps, graphs, infographics) connected with approaches of computer graphics (including 3D objects) and data journalism.

[3] http://wiki.openstreetmap.org/wiki/WikiProject_Corine_Land_Cover
[4] http://wiki.openstreetmap.org/wiki/Corine_Land_Cover

6 REFERENCES

CERBA, Otakar: Ontologie jako nastroj pro navrhy datovych modelu vybranych temat priloh smernice INSPIRE. Dissertation, Univerzita Karlova v Praze, 2011.

CERBA, Otakar, JEDLICKA, Karel, CHARVAT, Karel, JEZEK, Jan, JANECKA, Karel, MILDORF, Tomas: The Overview Of Spatial Data Harmonisation Approaches And Tools. In Proceedings Vol. 1. 4th International Conference on Cartography and GIS, Albena, 2012. ISSN: 1314-0604.

CERBA, Otakar, JEDLICKA, Karel, CHARVAT, Karel: Ontologies and ETL Tool in Free Spatial Data Integration. In GeoCart'2012 and ICA Regional Symposium on Cartography for Australasia and Oceania, Auckland, 2012. ISBN: 978-0-473-22313-7.

CORINE land cover. Commission of the European Communities, 1995.

Drafting Team "Data Specifications" – deliverable D2.5: Generic Conceptual Model. Drafting Team "Data Specifications", 2007.

FERANEC, Jan., OTAHEĽ, Jan: The 4th Level CORINE Land Cover Nomenclature For The PHARE Countries. In Developments in strategic landscape monitoring for the Nordic countries. Groom, G. (ed.). Nordic Council of Ministers, Copenhagen (Dánsko), 2004. p. 54-63.

HAKLAY, Mordechai, WEBER, Patrick: OpenStreetMap: User-Generated Street Maps, In Pervasive Computing, IEEE , vol.7, no.4, pp.12-18, Oct.-Dec. 2008.

KOPECKÁ, Monika: Identifikácia a hodnotenie zmien krajiny vo veľkej mierke (na príklade okolia Trnavy) . In Geografický časopis, roč. 58, č. 2, 2006. s. 125-148. ISSN 0016-7193.

MEIRICH, Susanne. Mapping Guide for a European Urban Atlas. GSE Land Consortium, 2008.

ISBN: 978-3-9503110-4-4 (CD-ROM); ISBN: 978-3-9503110-5-1 (Print)
Editors: Manfred SCHRENK, Vasily V. POPOVICH, Peter ZEILE, Pietro ELISEI