shareOGD – an Approach for Integrating Semantic Information in Open Government Data

Johannes SCHOLZ[1], Roland GRILLMAYER[2] and Manfred MITTLBÖCK[3]

[1]Research Studios Austria, Studio iSPACE, Salzburg/Austria· johannes.scholz@researchstudio.at
[2]University of Applied Sciences Wiener Neustadt, Department of Geomatics, Wr. Neustadt/Austria
[3]Research Studios Austria, Studio iSPACE, Salzburg/Austria

Abstract

Open Government Data are data that are published in digital format by administrative bodies. The Open Government Data Guidelines should cover how data have to be published to fulfill the requirements of the linked open data approach. The idea of open linked data is that resources published on the Internet can be interlinked and this combination leads to new information by utilizing semantic web technologies (Bizer et al. 2009). One essential premise to cover linked open data requirements is that the semantics of the data must be published in order to define the meaning of the objects of events represented by the data. At the moment these requirements are not covered within the existing Open Government Data guidelines. Hence, Open Government Data are published in a digital format but they do not follow a certain data specification, nor share a semantic model. Hence, the published data of various administrative bodies can hardly be compared, due to the lack of a standardized data model and missing semantic description. This paper attempts to list current shortcomings of Open Governmental Data and presents data modelling and presentation concepts for Open Governmental Data in order to follow linked data rules (Berners-Lee 2006).

1 Extended Abstract

Open Governmental Data (OGD) are data published by administrative bodies in a digital format and comprise of all data generated by public authorities. The goal of the OGD initiative is to improve transparency in the political processes and to foster the involvement of the public in decision making. These data have an inherent spatial and temporal component, as a number of statistical data are published or data have an effect on the geographic space as such. In Austria the OGD are published using the portal http://data.gv.at.

The economic potential of the OGD Initiative was analysed by MEPSIR (2011) and HUIJBOOM & VAN DEN BROEK (2011). These studies quantified the potential market impact of OGD with 68 billion EUR. In order to reveal the economic power of the OGD initiative several publications suggest improving the fitness for use in business applications (FORNEFELD et al., 2003; HUIJBOOM & VAN DEN BROEK, 2011). KALTENBÖCK et al. (2011) listed criteria for OGD necessary to improve their usability for applications directly:

- harmonized spatial reference of published OGD;
- near real-time availability of newly generated OGD, or at least defined update intervals;
- comparableness of data originating from different sources;
- description of published datasets with metadata;

The criteria mentioned above by KALTENBÖCK et al. (2011) are underpinned by the fact that different administrative authorities do not follow a shared data model or semantic annotation strategy when publishing data on the same topic. In addition, the spatial reference is not harmonized as such, which is hindering the computer-based creation of spatial datasets out of OGD. These arguments are justified by the example of OGD birth data in Austria, which are described in the following paragraph.

Searching for the birth related OGD using the portal data.gv.at results in the datasets given in Table 1. The map depicted in figure depicts the spatial extent and granularity of the OGD birth data sets. In Figure 2 the dataset of the City of Linz is provided, whereas in figure 3 the dataset offered by the province of Vorarlberg is presented. The birth data of the City of Linz provide a separate CSV file for each year that lists the births in relation to the age of the mother. The Vorarlberg example lists for each year and community the numbers of male and female births – and thus represents a time series. Hence, both datasets obviously do not share a common data model and a semantic model that would allow a direct comparison of the datasets. In addition the spatial reference of the datasets is not directly comparable – one dataset covers the City of Linz with approx. 191.000 inhabitants, and the other one describes 96 spatial reference units (communities) with a total of approx. 373.000 inhabitants (average population per community: approx.. 3.900).

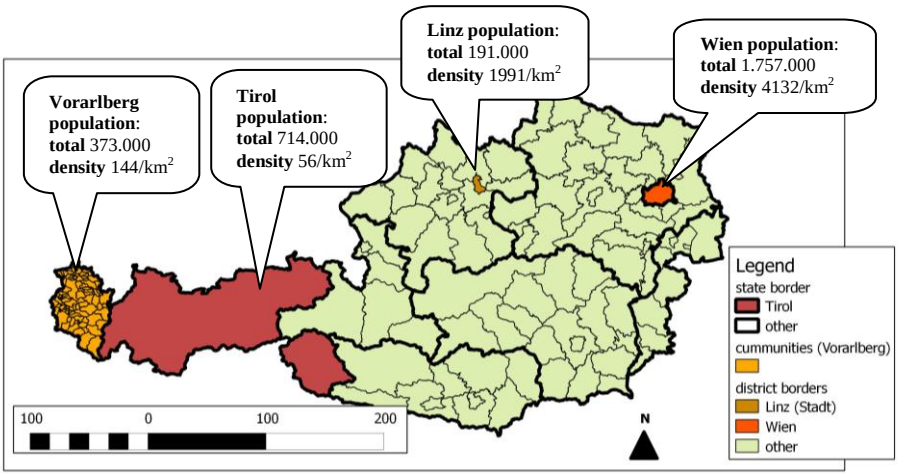
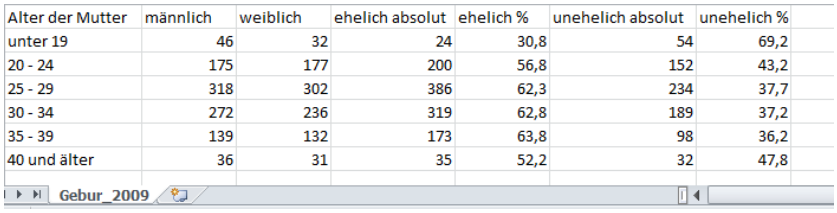


Fig. 1: Map of Austria and spatial extent and spatial resolution of available datasets concerning "birth" in data.gv.at. Datasets are available on community level in the state of Vorarlberg (the most western part of Austria), as dataset for the whole city of Linz (equals to one district), a dataset for whole province Tirol, and on census track level in the city of Vienna.

Table 1: Overview of datasets for "births" in data.gv.at.

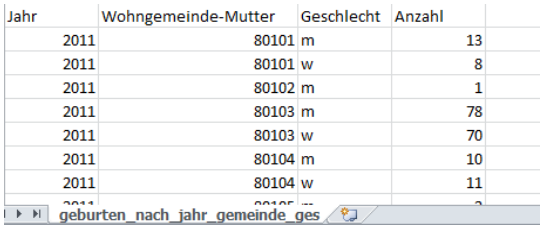
Title	Vendor	Format
Births	City of Linz	PDF, CSV
Birth statistics	Province of Vorarlberg	CSV
Birth-rate in Vienna: Sex and Census track	City of Vienna	CSV
Birth-rate in Vienna: Sex and Age	City of Vienna	CSV
Demographic indicators for Vienna – time series	City of Vienna	CSV
Population change (natural)	City of Linz	PDF, CSV
Live births by age of the mother	Province Tirol	HTML, CSV
Regional population prognosis 2005-2035	City of Vienna	MDB, CSV



Alter der Mutter	männlich	weiblich	ehelich absolut	ehelich %	unehelich absolut	unehelich %
unter 19	46	32	24	30,8	54	69,2
20 - 24	175	177	200	56,8	152	43,2
25 - 29	318	302	386	62,3	234	37,7
30 - 34	272	236	319	62,8	189	37,2
35 - 39	139	132	173	63,8	98	36,2
40 und älter	36	31	35	52,2	32	47,8

Gebur_2009

Fig. 2: Visualization of the dataset birth statistics of the City of Linz, revealing the inherent "data model". For each year a separate CSV file given that lists the births in relation to the age of the mother.



Jahr	Wohngemeinde-Mutter	Geschlecht	Anzahl
2011	80101	m	13
2011	80101	w	8
2011	80102	m	1
2011	80103	m	78
2011	80103	w	70
2011	80104	m	10
2011	80104	w	11

geburten_nach_jahr_gemeinde_ges

Fig. 3: Visualization of the dataset birth statistics of the Province of Vorarlberg, revealing the "data model". For each year and community the numbers of male and female births are given.

In order to be regarded as linked data, OGD should follow the four principles/rules defined by BERNERS-LEE (2006). Due to the fact, that certain OGD policies do not entirely follow the linked data rules to a certain degree, OGD implementations stop when a machine readable format of the data can be published, and neglect the issue of semantics. Based on the OGD examples (see figure 2 and figure 3) the authors conclude that automatic data mining and harmonization – also on a spatial-temporal level – is hardly possible without any auxiliary information. Hence, if OGD data are published in a pure machine-readable format without a semantic annotation, any application built on top of OGD cannot reveal the full information depth that is hidden in the data. Furthermore, any computer-based

georeferencing of OGDsets and subsequent analysis is hardly possible. Based on these findings the authors suggest to develop a concept for OGD Data that envisages integrated semantic information for each dataset that helps to avoid becoming (semantically) trapped in purely machine-readable data formats. In order to overcome the semantic heterogeneity of OGD and to make them more usable with respect to semantics and spatial reference we suggest applying methods from semantic engineering (KUHN 2002; KUHN 2009; LUTZ et al. 2009; VISSER & STUCKENSCHMIDT 2002; FICHTINGER et al 2011; SCHARFFE et al 2008). By adding semantic information to OGD, these datasets could be of use for a wide range of new spatial-enabled applications. A first step towards achieving this goal is done by the definition of an OGD metadata schema (http://reference.e-government.gv.at/uploads/media/OGD-Metadaten_2_1_2012_10.pdf) and provided metadata sets.

References

BERNERS-LEE, T. (2006), Linked Data - Design Issues. URL: http://www.w3.org/DesignIssues/LinkedData.html, last visited: 31-01-2013.

BIZER, C., HEATH, T. & BERNERS-LEE, T. (2009), Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), 5(3): 1-22.

BUNDESMINISTERIUM DES INNERN (2011), Open Government Data Deutschland. URL: http://bit.ly/ZDFmlf; last visited: 25-09-2012.

FICHTINGER, A., RIX, J., SCHÄFFLER, U., MICHI, I., GONE, M. & REITZ, T. (2011), Data Harmonisation Put into Practice by the HUMBOLDT Project. International Journal of Spatial Data Infrastructures Research, Vol.6, pp. 234-260.

HUIJBOOM, N. & VAN DEN BROEK, T., 2011. European Journal of ePractice. 12(March/April 2011), pp. 1-13. ISSN: 1988-625X

KALTENBÖCK M. & THURNER T. (HRSG) (2011), Open Government Data Weißbuch; Donau- Universität Krems; ISBN-13: 978-3902505231; URL: http://issuu.com/semwebcomp/docs/ogd_weissbuch_2011_web; last visited:14-08-2012

KUHN, W. (2002), Modeling the Semantics of Geographic Categories through Conceptual Integration. In: In GIScience. Lecture Notes in Computer Science, Springer, pp. 108-118.

KUHN, W. (2009), Semantic Engineering. In G. Navratil (Ed.): Research Trends in Geographic Information Science. Springer-Verlag Lecture Notes in Geoinformation and Cartography: 63- 74.

LUTZ, M., SPRADO, J., KLIEN, E., SCHUBERT, C. & CHRIST, I. (2009), Overcoming semantic heterogeneity in spatial data infrastructures, Computers & Geosciences, 35(4): 739-752.

MEPSIR (2006), Measuring European Public Sector Information Resources, Final Report of Study on Exploitation of public sector information – benchmarking of EU framework conditions, prepared for the European Commission, Brussels: HELM and Zenc.

SCHARFFE, F., EUZENAT, J. & FENSEL, D. (2008), Towards design patterns for ontology alignment. In: Proceedings of the 2008 ACM symposium on Applied computing (SAC '08). ACM, New York, NY, USA, 2321-2325. DOI=10.1145/1363686.1364236 http://doi.acm.org/10.1145/1363686.1364236

VISSER, U. & STUCKENSCHMIDT, H. (2002), Interoperability in GIS - Enabling Technologies. In: Ruiz,M., M. Gould & J. Ramon (eds.): 5th AGILE Conference on Geographic Information Science: 291-297.