Publishing Socio-Economic Territory Indices as Linked Data and their Visualization for Real Estate Valuation

Dina Sukhobok[1], Divna Djordjevic[2], Diego Sanvito[2], Javier Paniagua[3] and Dumitru Roman[1]

[1] SINTEF, Forskningsveien 1a, 0373 Oslo, Norway
{dina.sukhobok,dumitru.roman}@sintef.no
[2] CERVED, Via della Unione Europea, 6/A-6/B, 20097 San Donato Milanese, MI,
Italy {divna.djordjevic,diego.sanvito}@cerved.com
[3] SpazioDati S.r.l., Via A. Olivetti 13, 38122, Trento (TN), Italy
paniagua@spaziodati.eu

Abstract. The correct estimation of the real estate value facilitates decision making in various sectors, such as public administration or the real estate market. In this paper we demonstrate a method to manage territory scores and property valuation estimations as Linked Data with the help of the proDataMarket technical framework. The demo illustrates how the proDataMarket technical framework can be used to generate, maintain and serve territory and property valuation estimation data with the help of semantic technologies.

Keywords: Linked Data, socio-economic indices, property data, real estate data

1 Introduction

The correct estimation of the real estate value of residential properties owned by companies and individuals is one of the crucial elements for understanding their economic behavior and for predicting their financial stability. In order to improve current real estate property evaluation process, Cerved (an Italian data-driven company and a credit rating agency, providing services as credit information, marketing and credit management)[4] introduced a new algorithm for the evaluation of residential property that uses property data, open data and third-party data. The algorithm was implemented in the Cerved Cadastral Report Service (CCRS) enabling automatic estimate of the current market price for Italian residential properties following the Automated Value Model (AVM) methodology [5]. The final outcome of the service is a set of property valuation scores for the entire Italy, at the level of census sections from the 15th population and

---

[4] http://www.cerved.com/en
[5] https://en.wikipedia.org/wiki/Automated_valuation_model

housing census by ISTAT[6]. The approach estimates current properties values by taking into consideration values of comparable properties, number of rooms, property conditions, indication of value by revenue agency, comparable sales analysis of similar properties, surrounding sociodemographic and economic phenomena indicators such as schooling, pollution, type of industry, traffic, health care facilities, type of employment, revenue estimate, etc. In addition, higher level territory estimations have been created to analyze the marketing potential of a specific territory. The calculated property values and territory scores are then made available as Linked Data. The targeted customer segment is represented by Italian banks using the property valuation service for mortgage issuing or mortgage portfolio revaluation. The main goal of the service is to provide an accurate and objective evaluation of the real estate properties, contextualized to the market and to the territory of belonging, and updated in real-time. Cerved has also developed the Cerved Scouting the Terrain (CST) service, a Web-based map application that supports visualization and data aggregation of territory scores and property valuations from CCRS. The application was developed for Cerved's internal property appraisal department for increasing efficiently and quality of service when providing a range of products by exploring a selected area and a comparable set of properties.

In addition, all the developed territory scores and property valuation estimation data have been integrated, analyzed and can be easily visualized through a technology framework developed as part of the proDataMarket project[7].

2 Approach and Implementation

Data Sources. The estimated values of the properties derived by the CCRS algorithm are based on property data, numerous land-based socio-demographic scores, historical data on real estate appraisal from Cerved's proprietary database, property market values aggregated by the Italian revenue agency (i.e. OMI zones database), etc. The data used for the CCRS include various open, proprietary and 3rd party datasets. Including cadastral database, OMI zone database, company database including types of industry, and managers and shareholders database. Open data includes ISTAT data from the 2011 census, and OpenStreetMap (OSM) database[8]. The ISTAT data consists of tabular and shape data covering the Italian census from 2011 and contains data for 366,000 census sections with residents. These sections, in general, correspond to one district or a part of it, and are used as territory bases for the various developed scores. The disaggregated scores for the census sections include:

- *Social demographic score*: A score developed by Cerved using numerous socio-demographic variables from the ISTAT national census of 2011 and validated with the proprietary property appraisal dataset.

---

[6] http://www.istat.it/en
[7] https://prodatamarket.eu/
[8] http://wiki.openstreetmap.org/wiki/Database

- *Index of Social Distress (IDS)*: A score defined by the decree of the president of the ministers council from 2015[9], based on employment, unemployment, juvenile concentration and education rates from the ISTAT national census of 2011.
- *Index of Economic Distress (IDE)*: A score from the above referenced decree and variables from the 2011 census, based on proportion of residential properties in the urban areas in bad and medium state of preservation.
- *Manager and ShareHolders Concentration (MSHC) score*: A score based on Cerved official and proprietary data regarding people in roles of managers and shareholders.
- *Heavy Industrial Concentration (HIC) score*: A score from Cerved's official and proprietary data on industries in certain NACE[10] categories.
- Higher level integration scores: *People score* integrating the MSHC and the social demographic score, *territory score* integrating the HIC score and various proprietary features of the territory (e.g., OSM dataset), and the overall *real estate integrated score.*

The calculated scores are provisioned in a tabular format.

Ontology description. On a semantic level, we used the proDataMarket ontology[5] for describing census cells geospatial attributes, socio-demographic data from the ISTAT national census of 2011 and property valuation estimation scores. Census cells were defined as geospatial objects to capture geospatial attributes of a census tract, whereas all the scores and socio-demographic data were described with the help of the generic concept of Indicator from the proDataMarket Common Vocabulary[11].

Linked Data Generation and Publication. Linked data generation and data publication was performed using the proDataMarket platform – a part of the proDataMarket technical framework, used for data cleaning, data transformation and data hosting. The proDataMarket platform includes a set of software components - DataGraft[12][2,3] and Grafterizer[4] and facilitates interactive specification of tabular data transformations and mapping of tabular data to graph data (RDF) and publishing data as a SPARQL endpoint.

3 Demonstration Outline

During the demo we will introduce the proDataMarket platform and show how it can be used to generate Linked Data from tabular data and publish it through a SPARQL endpoint, using property valuation scores as an example.

---

[9] http://www.gazzettaufficiale.it/eli/id/2015/10/26/15A08012/sg
[10] http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:
Statistical_classification_of_economic_activities_in_the_European_
Community_(NACE)
[11] http://vocabs.datagraft.net/proDataMarket/0.1/Common
[12] https://datagraft.io/

In addition, we will present the visualization of various indices in the pro-DataMarket portal and through CST Web-based map application (see Figure 1), by using data from the SPARQL endpoint and displaying several property valuation estimation scores based on census sections and properties details. In the proDataMarket portal, users have the ability to navigate the territory map, with the possibility to zoom-in and zoom-out, combine filters to fulfill more specific queries, select and visualize on map different data layers for different property datasets and to analyze the territory (through a dedicated colour scale for each layer of data).
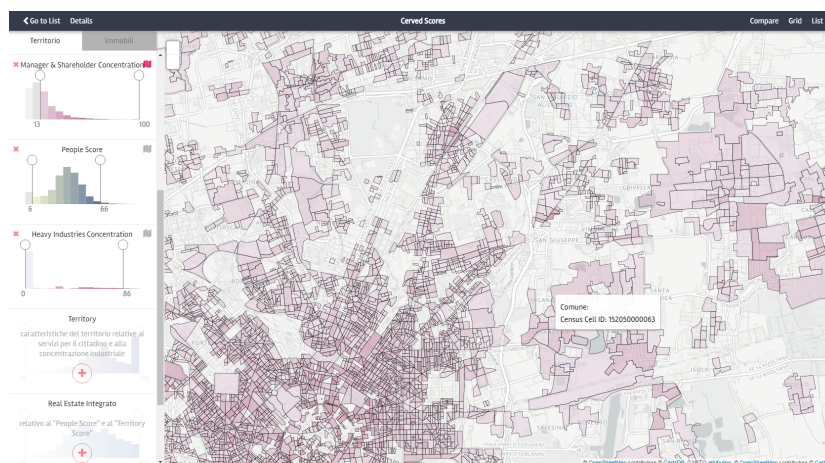

Fig. 1. Property valuation estimation scores in the CST application.

References

1. Pozzati, Stefano, et al. "Understanding territorial distribution of Properties of Managers and Shareholders: A Data-driven Approach." Territorio Italia 2 (2016), DOI: 10.14609/Ti_2_16_2e, Pages 27-40, ISSN 2499-2674.
2. Roman, Dumitru, et al. DataGraft: Simplifying Open Data Publishing. ESWC (Satellite Events) 2016: 101-106.
3. Roman, Dumitru, et al. "DataGraft: One-stop-shop for open data management." To appear in the Semantic Web Journal (SWJ) Interoperability, Usability, Applicability (published and printed by IOS Press, ISSN: 1570-0844), 2017, DOI: 10.3233/SW-170263.
4. Sukhobok, Dina, et al. "Tabular Data Cleaning and Linked Data Generation with Grafterizer." ESWC (Satellite Events) 2016: 134-139.
5. Shi, Ling, et al. The proDataMarket Ontology for Publishing and Integrating Cross-domain Real Property Data. To appear in the journal "Territorio Italia. Land Administration, Cadastre and Real Estate". n.2/2017.