Accepted Manuscript

Enriching integrated statistical open city data by combining equational knowledge and missing value imputation

Stefan Bischof, Andreas Harth, Benedikt Kämpgen, Axel Polleres, Patrik Schneider

PII: S1570-8268(17)30035-5
DOI: https://doi.org/10.1016/j.websem.2017.09.003
Reference: WEBSEM 445

To appear in: *Web Semantics: Science, Services and Agents on the World Wide Web*

Received date : 2 February 2017
Revised date : 1 August 2017
Accepted date : 17 September 2017

Please cite this article as: S. Bischof, A. Harth, B. Kämpgen, A. Polleres, P. Schneider, Enriching integrated statistical open city data by combining equational knowledge and missing value imputation, *Web Semantics: Science, Services and Agents on the World Wide Web* (2017), https://doi.org/10.1016/j.websem.2017.09.003

Enriching Integrated Statistical Open City Data by Combining Equational Knowledge and Missing Value Imputation

Stefan Bischof[a,*], Andreas Harth[b], Benedikt Kämpgen[c], Axel Polleres[d,e], Patrik Schneider[a]

[a]*Siemens AG Österreich, Siemensstrasse 90, 1210 Vienna, Austria*
[b]*Karlsruhe Institute of Technology, Karlsruhe, Germany*
[c]*FZI Research Center for Information Technology, Karlsruhe, Germany*
[d]*Vienna University of Economics and Business, Vienna, Austria*
[e]*Complexity Science Hub Vienna, Austria*

Abstract

Several institutions collect statistical data about cities, regions, and countries for various purposes. Yet, while access to high quality and recent such data is both crucial for decision makers and a means for achieving transparency to the public, all too often such collections of data remain isolated and not re-usable, let alone comparable or properly integrated. In this paper we present the Open City Data Pipeline, a focused attempt to collect, integrate, and enrich statistical data collected at city level worldwide, and re-publish the resulting dataset in a re-usable manner as Linked Data. The main features of the Open City Data Pipeline are: (i) we integrate and cleanse data from several sources in a modular and extensible, always up-to-date fashion; (ii) we use both Machine Learning techniques and reasoning over equational background knowledge to enrich the data by imputing missing values, (iii) we assess the estimated accuracy of such imputations per indicator. Additionally, (iv) we make the integrated and enriched data, including links to external data sources, such as DBpedia, available both in a web browser interface and as machine-readable Linked Data, using standard vocabularies such as QB and PROV.

Apart from providing a contribution to the growing collection of data available as Linked Data, our enrichment process for missing values also contributes a novel methodology for combining *rule-based inference* about equational knowledge with inferences obtained from *statistical Machine Learning* approaches. While most existing works about inference in Linked Data have focused on ontological reasoning in RDFS and OWL, we believe that these complementary methods and particularly their combination could be fruitfully applied also in many other domains for integrating Statistical Linked Data, independent from our concrete use case of integrating city data.

Keywords: open data, Linked Data, data cleaning, data integration

1. Introduction

The public sector collects large amounts of statistical data. For example, the United Nations Statistics Division[1] provides regularly updated statistics about the economy, demographics and social indicators, environment and energy, and gender on a global level. The statistical office of the European Commission, Eurostat[2], provides statistical data mainly about EU member countries. Some of the data in Eurostat has been aggregated from the statistical offices of the member countries of the EU. Even several larger cities provide data in on their own open data portals, e.g., Amsterdam, Berlin, London, or Vienna[3]. Increasingly, such data can be downloaded free of charge and used under liberal licences.

Such open data can benefit public administrations, citizens and enterprises. The public administration can use the data to support decision-making and back policy decisions in a transparent manner. Citizens can be better informed about government decisions, as publicly available data can help to raise awareness and underpin public discussions. Finally, companies could develop new business models and offer tailored solutions to their customers based on such open data. As an example for making use of such data, consider Siemens' Green City Index (GCI) [1], which assesses and compares the environmental performance of cities. In order to compute the KPIs used to rank cities' sustainability, the GCI used qualitative and also quantitative indicators about city performance, such as for instance CO_2 emissions or energy consumption per capita. Although many of these quantitative indicators had been openly available, the respective datasets had to be collected, integrated, and checked for integrity violations mostly manually because of the following reasons: (i) heterogeneity: ambiguous data published by different Open Data sources in different formats, (ii) missing data, that needed to be added manually through additional research in text documents

*Corresponding author
Email addresses: bischof.stefan@siemens.com (Stefan Bischof), harth@kit.edu (Andreas Harth), kaempgen@fzi.de (Benedikt Kämpgen), axel.polleres@wu.ac.at (Axel Polleres), patrik@kr.tuwien.ac.at (Patrik Schneider)

[1] http://unstats.un.org/unsd/

[2] http://ec.europa.eu/eurostat/

[3] http://data.amsterdam.nl/, http://daten.berlin.de/, http://data.london.gov.uk/, and http://data.wien.gv.at/