City Data Pipeline

A System for Making Open Data Useful for Cities

Stefan Bischof[1,2], Axel Polleres[1], and Simon Sperl[1]

[1] Siemens AG Österreich, Siemensstraße 90, 1211 Vienna, Austria
{bischof.stefan,axel.polleres,simon.sperl}@siemens.com
[2] Vienna University of Technology, Favoritenstraße 9, 1040 Vienna, Austria
stefan.bischof@tuwien.ac.at

Abstract. Some cities publish data in an open form. But even more cities can profit from the data that is already available as open or linked data. Unfortunately open data of different sources is usually given also in different heterogeneous data formats. With the City Data Pipeline we aim to integrate data about cities in a common data model by using Semantic Web technologies. Eventually we want to support city officials with their decisions by providing automated analytics support.

Keywords: open data, data cleaning, data integration

1 Introduction

Nowadays governments have a big arsenal of data available for decision support. But also city administrators need this kind of data to make better decisions and policies for leading cities to a greener, smarter, and more sustainable future. Having access to correct and current data is crucial to advance on these goals. Printed documents like the Green City Index [3] are helpful, but outdated soon after publication, thus making a regularly updated data store necessary.

Even though there is lots of data available as open data, it is still cumbersome to collect, clean, integrate, and analyze data from different sources, with different specifications, written in different languages, and stored in different formats. Sources of city data can be widely known linked open data sources like DBpedia, Geonames, or Eurostat via Linked Statistics. Urban Audit[3] for example, provides almost 300 indicators on several domains for 258 European cities. But there are also many smaller data sources which provide data in a narrow domain only, like oil prices or stock exchange rates. Furthermore several larger cities provide data from their own databases, e.g., London[4], Berlin[5], or Vienna[6]. Data is available in different formats following different data models. One can find data in RDF, XML, CSV, RTF, XLS, or HTML. The specification of the individual data fields

[3] http://www.urbanaudit.org/
[4] http://data.london.gov.uk/
[5] http://daten.berlin.de/
[6] http://data.wien.gv.at/

is often implicit only (in free text documents) and has to be processed manually for understanding. Small and medium sized cities often do not have the resources to handle these kinds of data heterogeneity and thus often miss relevant data.

With the *City Data Pipeline* we aim at providing an extensible platform to support citizens and city administrators by providing *city key performance indicators* (KPIs) based on diverse publicly available open data sources.

The project QuerioCity [5] uses partly similar techniques, but does not include an analytics component which is one of the main features of our system.

2 Architecture and Main Features

The City Data Pipeline collects data, organizes this data into indicators, and shows these indicators to the user. The system is organized in several layers which this section explains in more detail: crawler, wrapper components, semantic integration, data storage, analytics, and user interface (see Figure 1).

Crawler. The City Data Pipeline (semi-)automatically collects data from various registered open data sources in a periodic manner dependent on the specific source. The crawler currently collects data from 32 different sources, e.g., DBpedia, UN open data, Urban Audit, as well as datasets of several cities. Adding new data sources is a semi-automatic process where manual effort is necessary.

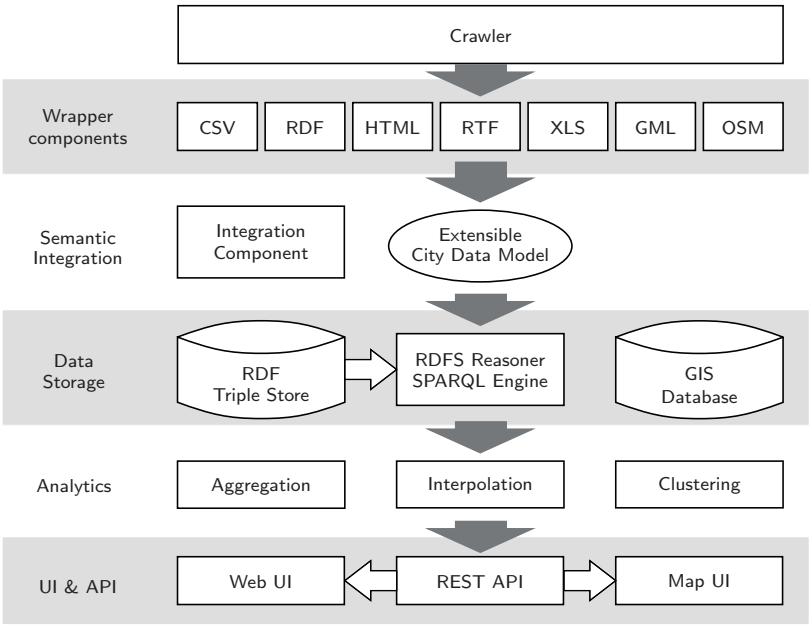

Fig. 1. City Data Pipeline architecture showing components for crawling wrapping, cleaning, integrating, and presenting information

Wrapper components. As a first step of data integration, a set of wrapper components parses the downloaded data and converts it to a source specific RDF.

The set of wrapper components include a CSV wrapper to parse and clean CSV data, a wrapper for extracting HTML tables, a wrapper for extracting tables of RTF documents, a wrapper for Excel sheets, and a wrapper for cleaning RDF data as well. All of these wrappers are customizable to cater for diverse source-specific issues. These wrapper components convert the data to RDF and preprocess the data before integrating the data with the existing triple store. Preprocessing contains the usual data cleansing tasks, unit conversions, number and data formatting, string encoding, and filtering invalid data.

Furthermore there is an OpenStreetMap (OSM) wrapper and a wrapper for GML [4] data, to feed the *geographic information system* (GIS) database.

Semantic integration. To be able to access a single KPI such as the population number, which is provided by several data sources, the semantic integration component *unifies the vocabulary* used by the different data sources. The semantic integration component is partly implemented in the individual wrappers and partly by an RDFS [2] ontology (extended with capabilities for reasoning over numbers by using equations [1]) called *City Data Model*. The ontology covers several aspects: spatial context (country, region, city, district), temporal context (validity, date retrieved), provenance (data source), terms of usage (license), and an extensible list of indicators for cities. For each indicator the ontology contains descriptions and a reference to an indicator category, e.g., *Demography*. To integrate the source specific indicators the ontology maps data source specific RDF properties to City Data Model properties, e.g., it maps dbpedia:population to citydata:population by an RDFS subPropertyOf property.

Data storage. For storing the processed data we use Jena TDB[7] as *triple store* for RDF data, and PostGIS/PostgreSQL as a *GIS database* for geographic information. GIS databases allow us to compute missing information such as areas of cities or districts, or lengths of certain paths. Subsequent subsystems can access the RDF data via a SPARQL interface. The SPARQL engine provides RDFS reasoning support by query rewriting (including reasoning over numbers [1]).

Analytics. When integrated, open data contains incomplete data. Different tools in the analytics layer try to complete data by using statistical or simple algebraic methods. The analytics layer also includes tools for value aggregation as well as clustering of similar cities. We plan to extend the analytics part to allow in-depth analysis of city data to reveal hidden relationships.

User interface and API. Figure 2 shows the simple Java powered web interface. The interface also provides programmatic access via HTTP GET and HTTP POST to allow external tools such as data visualization frameworks, to query the database. The web application communicates with the Jena triple store via SPARQL 1.1 by using the Jena API directly.

[7] `http://jena.apache.org/documentation/tdb/`

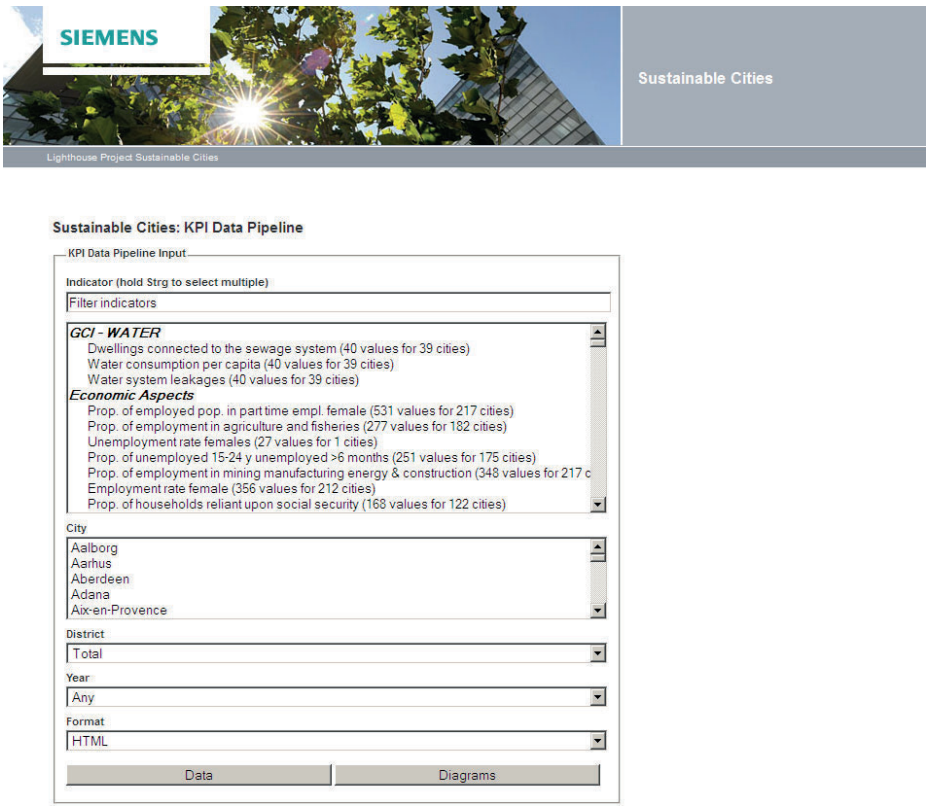

Fig. 2. Web interface for querying the City Data Pipeline, which also provides programmatic access via HTTP GET/POST

Users can select one or more of the 475 *indicators* from a list sorted by categories like *Demography*, *Geography*, *Social Aspects*, or *Environment*. The list also shows how many data points are available per indicator and for how many cities data points are available for this indicator. Next the user can select one or several of the 350 European *cities* for which we collected data. For a few cities we even have information on the individual districts available. In these cases the user can select one or several of the districts. Optionally the user can specify a *temporal context*, for which year the database should be queried. This feature allows to compare several cities with each other at a certain point of time instead of listing data of all available times. The user interface also allows the computation of *complex KPIs*. These KPIs are specified by a set of formulas in an Excel sheet and are computed on demand. Finally the system can *output* the query results as HTML report but also as XML document for further processing. With the XML export option, the web application can actually be used straightforwardly by external tools, providing for example more sophisticated visualization. One

visualizer of this kind is already implemented, showing selected data points for different cities on an interactive world map.

Currently the City Data Pipeline stores an average of *285 data points per city*. Since bigger cities tend to have a wider coverage of domains, with finer granularity of time and space, the number of available data points per city is unequally distributed. While we are currently not able to provide data, ontology, or web interface for public access, we hope this changes in the future.

3 Conclusions and Outlook

The *City Data Pipeline* provides seamless access to indicators of over 30 open data providers. The system integrates data from different domains, in different formats with different data models. The City Data Pipeline allows querying and comparing indicators for many European cities thus making analytics easier.

Currently we are working on more methods for estimating missing values and predicting selected indicators based on multiple criteria. For this purpose and other kinds of data analytics we extend the data mining tool RapidMiner[8].

Furthermore we are in the process of improving the user interface to make the application more intuitive. For this purpose we use the Google Web Toolkit together with several libraries for more advanced information visualization like different kinds of interactive charts or world maps.

Acknowledgements. Stefan Bischof has been partially funded by the Vienna Science and Technology Fund (WWTF) through project ICT12-015.

References

1. Bischof, S., Polleres, A.: RDFS with Attribute Equations via SPARQL Rewriting. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) The Semantic Web: Semantics and Big Data, LNCS, vol. 7882, pp. 335–350. Springer Berlin Heidelberg (2013)
2. Brickley, D., Guha, R., (eds.): RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation (2004), http://www.w3.org/TR/rdf-schema/
3. Economist Intelligence Unit (ed.): The Green City Index. Siemens AG (2012), http://www.siemens.com/press/pool/de/events/2012/corporate/2012-06-rio20/gci-report-e.pdf
4. ISO: Geographic information – Geography Markup Language (GML). ISO standard 19136, International Organization for Standardization (2007)
5. Lopez, V., Kotoulas, S., Sbodio, M., Stephenson, M., Gkoulalas-Divanis, A., Aonghusa, P.: Queriocity: A linked data platform for urban information management. In: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (eds.) The Semantic Web – ISWC 2012, LNCS, vol. 7650, pp. 148–163. Springer Berlin Heidelberg (2012)

[8] http://rapid-i.com/