Toward Automatic Data Curation for Open Data

by Thilo Stadelmann, Mark Cieliebak and Kurt Stockinger

In recent years large amounts of data have been made publicly available: literally thousands of open data sources exist, with genome data, temperature measurements, stock market prices, population and income statistics etc. However, accessing and combining data from different data sources is both non-trivial and very time consuming. These tasks typically take up to 80% of the time of data scientists. Automatic integration and curation of open data can facilitate this process.

Most open data has scientific or governmental origins and much of it resides in isolated data stores. In data warehousing, data integration as a discipline provides best practices concerning data preparation and data management tasks by offering standardized processes and tool chains. However, with the recent popularity of Big Data, an unprecedented number of new data sources contribute to an increasingly heterogeneous trove of data. Hence, 'data curation' – a fully automated means of intelligently finding and combining possible data sources in the unified space of internal and open data sources – is in high demand [1].

We recently finished a market research and architectural blueprint, funded by the Hasler Stiftung, to evaluate requirements and business cases concerning the development of such an automatic data curation system in Switzerland.

Market Research

The Swiss ICT sector is healthy and innovative, sustained by strong players in research (e.g., universities and privately held research institutions) and industry (e.g., finance and pharma) as well as a large ecosystem of SMEs and startups. All surveyed parties among this group of stakeholders responded by stating their need for better data curation support of open data that has not been previously integrated within internal data sources.

As examples, we identified several use cases that rely on the existence of such a service:

- Economic research would be significantly improved by using automatic data curation for unifying tax and rent data of Swiss municipalities
- In a transportation research project, the overall project cost increased by 25% because of scattered and heterogeneous public data.
- Scientific digital media archives could fully index their work, thereby creating new research and application possibilities.

For these reasons, national funding agencies are very keen to support the development of such a service based on a solid model of business and operations.

Such a business model could consist of offering automatic data curation as software-as-a-service for open data. In order to comply with open data standards, the access to the data itself has to be free, while additional services could be offered on a freemium basis. To be of interest to industrial customers, private installations to curate confidential internal data could also be offered.

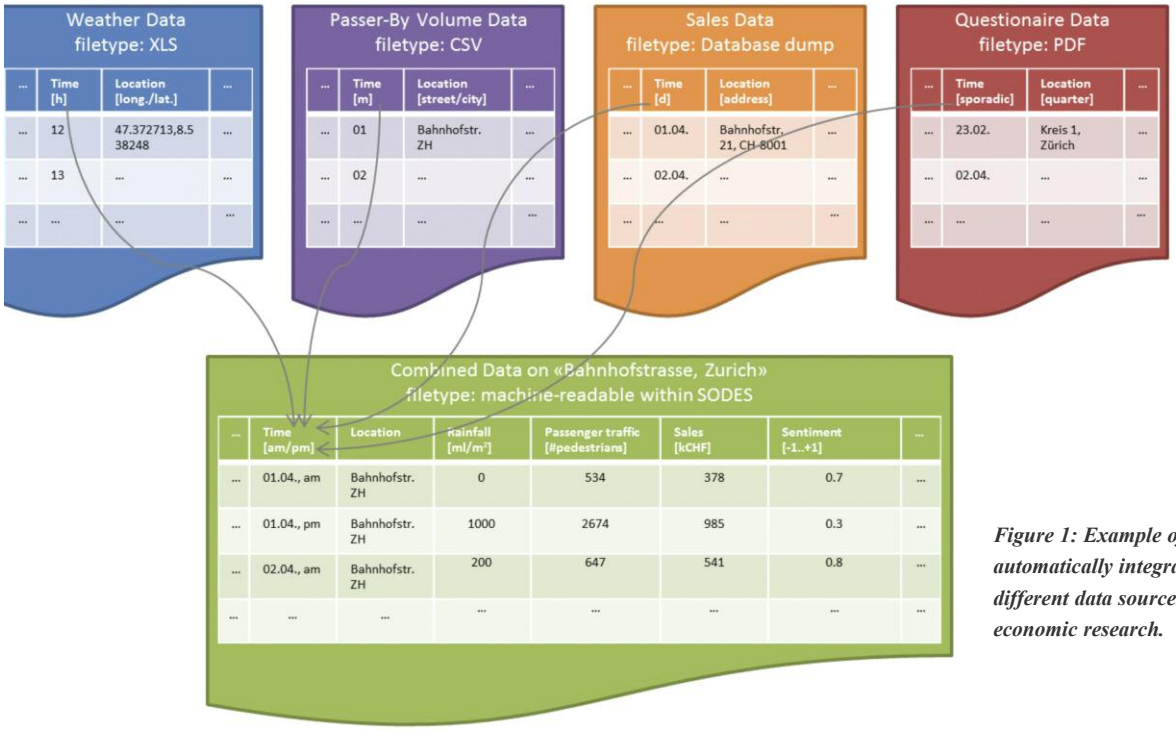

Figure 1: Example of automatically integrating four different data sources for socio-economic research.

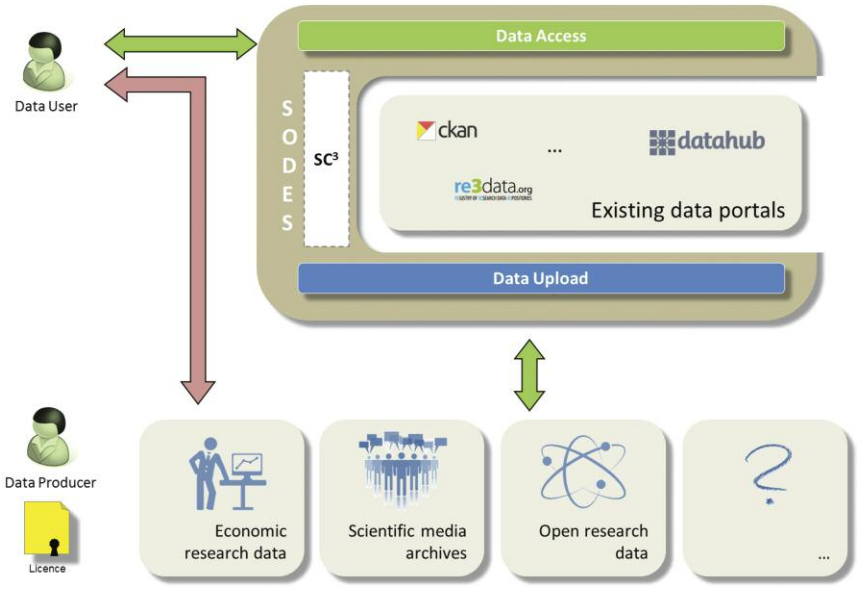

Figure 2: The high-level architecture of SODES in interaction with data providers and data portals.

While academic institutions see their general mandate to offer such services to Swiss researchers, they typically do not target industry use cases. The business models of current cloud computing providers do not necessarily coincide with offering a data curation platform to researchers, although synergies exist, e.g. through the FIWARE project.

Architecture Blueprint

We thus propose SODES – the blueprint for a Swiss Open Data Exploration System - that enables easy and intuitive access, integration and exploration of different data sources. Figure 1 shows an example of the effect: Four different data sources on Zurich's Bahnhofstrasse are automatically integrated based on common 'Time' and 'Location' columns despite the different data types and granularities therein.

SODES is designed as a platform that offers content-based search, automatic data curation and integrated data preview on top of established data technologies such as CKAN or Linked Open Data. As a service on the Internet or within any organization, SODES is envisioned to enable data scientists to do most of their data exploration in one place.

SODES can be viewed as a wrapper around existing data portals that focus on collecting resources (see Figure 2): While the user is free to still retrieve the data directly from the producer, SODES particularly adds easy-to-use data upload and data access user interfaces. These are enabled through the semantic context comprehension component SC3, powered by advanced machine learning. Therefore, SODES must process the data in machine-readable form. For advanced data analytics, users download the combined data to their tool of choice or access it remotely through standard APIs.

Future Activities

SODES is driven by a consortium of research institutions (Zurich University of Applied Sciences, University of Zurich, ETH Zurich), NGOs (Foundation opendata.ch) and industry partners (Liip AG, itopia AG) and targets Swiss data scientists in academia and industry. We are open to additional partners with interests in operating the service in order to continue development.

Links:

Swiss Open Data: http://opendata.ch
FIWARE: http://www.fi-ware.org

Reference:

[1] M. Stonebraker et al.: "Data Curation at Scale: The Data Tamer System", CIDR 2013.

Please contact:

Mark Cieliebak
ZHAW School of Engineering
Tel. +41 58 934 72 39
E-mail: ciel@zhaw.ch

An Interactive Tool for Transparent Data Preprocessing

by Olivier Parisot and Thomas Tamisier

We propose a visual tool to assist data scientists in data preprocessing. The tool interactively shows the transformation impacts and information loss, while keeping track of the applied preprocessing tasks.

Data analysis is an important topic for several domains in computer science, such as data mining, machine learning, data visualization and predictive analytics. In this context, scientists aim at inventing new techniques and algorithms to handle data and identify meaningful patterns within datasets.

Usually, data analysis techniques are worked out by using benchmark datasets: for instance, the UCI Machine Learning Repository contains a lot of material for different tasks (regressions, prediction, classification, etc.). This repository is widely used; many academic papers in the domain refer to its datasets in order to allow meaningful comparisons with the state of the art.

In practice, preprocessing is often necessary to adjust the benchmark data to the specificity of new algorithms or methods [1]. More precisely, 'data preprocessing' refers to a collection of different data transformation techniques: 'cleansing' (treatment of noise, etc.), 'dimensionality alteration' (filtering of features, etc.) and 'quantity alteration' (sampling of the data records). Moreover, a preprocessing process could drastically affect the original data, and the results of a data analysis could