Improving Open Data Usability through Semantics

Sebastian Neumaier

Supervisors: Axel Polleres, Jürgen Umbrich
Vienna University of Economics and Business
Problem

• Open Data is (partially) structured and tabular [1]:

- 3-star: 27%
- 2-star: 12%
- 1-star: 10%

![Bar chart showing data formats: CSV, Excel, PDF, Missing Format, 82 data portals, 160K datasets.]

- Use of existing Semantic Web technologies to improve existing 3-star Open Data
  - Integrate and interlink data
  - Allows cross-portal integration

• Focus on tabular data

State of the art

- Existing work mainly use named-entity recognition techniques
- Web/HTML tables differ from typical Open Data tables:
  - **Domain**: e.g., public administration data, statistical data, weather data, elections, ...
  - **Structure**: OD tables contain large amount of numerical columns

Research Question & Hypothesis

The semantic analysis of Open Data CSVs increases the usability of Open Data and ideally generates 5-star linked data.

- **How can we use existing Semantic Web technologies?**
  - Report and analysis of current OD in order to select/filter methods

- **How can we assign semantic labels to numerical columns?**
  - Labeling of numeric data based on background knowledge graph

- **Is there a meaningful definition of relatedness?**
  - Relatedness measure allows search and recommendation
Approach

- **Monitoring** and analysis of Open Data portals
  
  - identify quality issues and possible improvement/integration strategies

- Evaluate applicability of existing techniques
  
  - identify approaches which are applicable to our corpus of data

- Labeling and classification of numerical values
  
  - develop a method to find and rank candidates of semantic context descriptions for numerical values

- Open Data tables **recommendation** and linkage
  
  - incorporate previous steps into recommendation and linkage/integration system which allows automatic enrichment
Preliminary results

- **Monitoring and analysis of Open Data portals** [2,3]
  - Periodical monitoring and assessment
  - (Meta)data quality metrics, e.g.: Retrievability, existence of contact or license information
  - [http://data.wu.ac.at/portalwatch](http://data.wu.ac.at/portalwatch)

- **Labelling of numerical values** [4]
  - Assign multi-level semantic labels to numerical values
  - Based on knowledge graph constructed from DBpedia

Evaluation plan

• Existing Semantic label annotation and linking evaluation is based on (manually created) Gold standard datasets:
  • Crawl of the Web [5], Domain specific datasets (e.g., IMDB, MusicBrainz)

• Our numeric labelling approach:
  • Cross-validating over a sample of DBpedia
  • “In the wild” on tabular data extracted from Open Data portals to gain insights and report challenges

• Test-data corpus for further evaluations:
  • CSV data monitored by Open Data Portal Watch

• **Automated integration** and high-quality linkage of all Open Data CSV tables **not possible**

• **Partial mappings** already generate high value and increases the usability of current Open Data

• ADEQUATe project [6]:
  • “Analytics & Data Enrichement to improve the Quality of Open Data”
  • Identified the potential (and demand) for integration and semantic labelling of Open Data resources