reboting.com: Towards geo-search and visualization of Austrian Open Data

Erich Heil[1] and Sebastian Neumaier[2] *

[1] 23°, https://www.23degree.org/, Vienna, Austria
erich.heil@23degrees.io
[2] Vienna University of Economics and Business, Vienna, Austria
sebastian.neumaier@wu.ac.at

Abstract. Data portals mainly publish semi-structured, tabular formats which lack semantic descriptions of geo-entities and therefore, do not allow any exploration and automated visualization of these datasets. Herein, we present a framework to add geo-semantic labels, based on a constructed geo-entity knowledge graph, and a user interface to query and automatically visualize the resources from the Austrian data portals. The web-application is available at `https://reboting.com/`.

1 Introduction

Governmental Open Data portals such as the Austrian `data.gv.at` release local, regional and national data to a variety of users. The data is collected as part of census collections, infrastructure assessments or any other, secondary output data; for instance, public transport data of cities, demographic indicators, etc. Making this data accessible, searchable and analyzable to the public is vital to foster an open government [2]. However, geospatial information in Open Data – as it is currently published – mainly still comes in semi-structured and tabular formats, such as CSV or XLS [4] and geo-references in these tabular sources are not encoded structuredly or homogeneously, but using mixes of region names, country codes, or other implicit references. Therefore, these portals do not allow any geo-semantic queries; in fact, the search functionalities are limited to the metadata descriptions only, and hardly provide any visualizations to explore the datasets. Herein, we present a framework to automatically generated visualizations of open datasets based on queries for geo-entities:

1. We integrate Linked Data repositories, geo-reference datasets, and geocode standards in a hierarchical base geo-entities knowledge graph.
2. Using this knowledge graph, we label metadata and data of the Austrian data portals and index all labelled datasets. We provide an API to search over geo-entities, but also full-text search over the content.
3. The user interface at `reboting.com` offers showcase queries for Austrian geo-entities and displays automatically generated visualizations for any input dataset from the data portals.

* This work was supported by the Austrian Research Promotion Agency (FFG) under the projects ADEQUATe (grant no. 849982) and CommuniData (grant no. 855407).

2 Approach

In Figure 1 we display our overall approach: Initially, we crawl CSVs and metadata information (such as title, description, publisher) from the two Austrian data portals `data.gv.at` and `opendataportal.at` and label these using our constructed base knowledge graph (cf. Section 2.1). The data is stored and indexed in ElasticSearch.[3] The web application at reboting.com (Section 2.4) accesses the indexed data via a search API (Section 2.3).

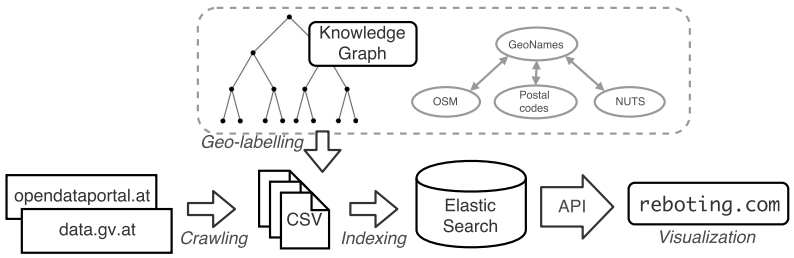

Fig. 1. Process of crawling, labelling, indexing and visualizing datasets from the Austrian Open Data portals.

2.1 Constructing a base knowledge graph of geo-entities

Similar to our approach, the 2012 project LinkedGeoData [6] uses Open Street Maps to construct a lightweight ontology and a Link Data resource. In contrast, our geo-entities knowledge graph is based on GeoNames:[4] it contains over 10 million geographical names and provides detailed hierarchical descriptions, e.g., countries, federal states, regions, cities, etc.. We enrich the GeoNames graph with three additional sources:

Postal codes: GeoNames also provides a comprehensive collection of postal codes for several countries and the respective name of the places/districts.[5] Additionally, we use codes available in Wikidata to extend and verify the graph.

NUTS is a geocode standard by the European Union (EU). It references the statistical subdivisions of all EU member states in three hierarchical levels.[6] Wikidata includes several links to the GeoNames repository as well as a property for the NUTS classifications of regions, so we can use the Wikidata SPARQL endpoint[7] to add mappings to the corresponding NUTS regions.

[3] https://www.elastic.co/products/elasticsearch
[4] http://www.geonames.org/
[5] http://download.geonames.org/export/zip/, last accessed 2018-01-05. Note, that this dataset is not linked to GeoNames entities: We heuristically map the codes to the entities by using the countries and place/district names.
[6] http://ec.europa.eu/eurostat/web/nuts/overview, last accessed 2018-03-05
[7] https://query.wikidata.org/ with the following query to get these NUTS-to-GeoNames mappings: `SELECT ?s ?nuts ?geonames WHERE {?s wdt:P605 ?nuts.?s wdt:P1566 ?geonames}`

OpenStreetMaps (OSM): To cover a more detailed and larger set of labels as it is available in GeoNames, e.g., the set of all street names and local places/POIs of a city, we extract OSM ways and nodes and map these to the GeoNames hierarchy. We use OSM's Nominatim service to get polygons for all district/city-level regions and use these to extract all street names, places, etc. from an OSM country extract (currently Austria).

2.2 Geo-labelling algorithm

Our framework crawls the datasets from the two Austrian Open Data portals `data.gv.at` and `opendataportal.at`, and annotates (i) the CSVs' columns and (ii) the metadata descriptions using the geo-entities in our base knowledge graph.

(i) *The CSVs'* columns get classified based on regular expressions for NUTS identifier and postal codes. In case a column holds potential postal codes the algorithm tries to map the values to existing postal codes and add the respective semantic labels.

In case of string columns, we first try to map the column values to GeoNames labels: We collect all possible entity mappings for all column values, and disambiguate values with multiple GeoNames candidates based on the predecessors in the knowledge graph, i.e., we sum up the aggregated counts of these predecessors to resolve a mapping.[8] If no GeoNames mapping was found we try to instantiate the values with the OSM street names and places from our knowledge graph.

(ii) *The Metadata descriptions*, which can be found on the Open Data portals, often give hints about the respective region covering the actual data. Therefore, we try to extract geo-entities from the titles, descriptions and publishers of the dataset. Table 1 lists the total number of indexed CSVs, the number of CSVs with annotated columns (by GeoNames and OSM labels) and the number of datasets where we could annotate the metadata. The source code for the graph construction and the labelling is available on GitHub.[9]

Table 1. Number of CSVs with column and metadata labels.

portal	**indexed**	**Columns**	GeoNames	OSM	**Metadata**
data.gv.at	2427	717 *(30%)*	587	185	2391 *(99%)*
opendataportal.at	442	7 *(2%)*	5	3	441 *(99%)*

2.3 Search API

An indexed document (i.e., CSV) in Elasticsearch contains all cell values of the table (arranged by columns), the potential geo-labels for the labelled columns,

[8] This algorithm is based on the assumption that values in the same column have a common context, i.e., common predecessors.

[9] https://github.com/sebneu/geolabelling

the metadata of the CSV (e.g., the data portal, title, publisher, etc.), and any additional geo-labels extracted from this metadata. The data can be accessed via the ODGraph API (http://data.wu.ac.at/odgraph/):
`/locationsearch?l={GeoNames}&offset={offset}&limit={limit}&q={keyword}`
It takes multiple instances of `GeoNames` identifiers (parameter `l`) and an optional white space separated list of keywords (`q`) as input parameters. The output consists of a list of documents that match the requested entities or keywords.

Alternatively to the document-based Elasticsearch, GeoSPARQL [5] defines a small ontology to represent and query geometries and spatial regions. In future work, we plan to make our base knowledge graph and RDFized linked data points from the CSVs also available via a GeoSPARQL endpoint.

2.4 User interface & Visualization

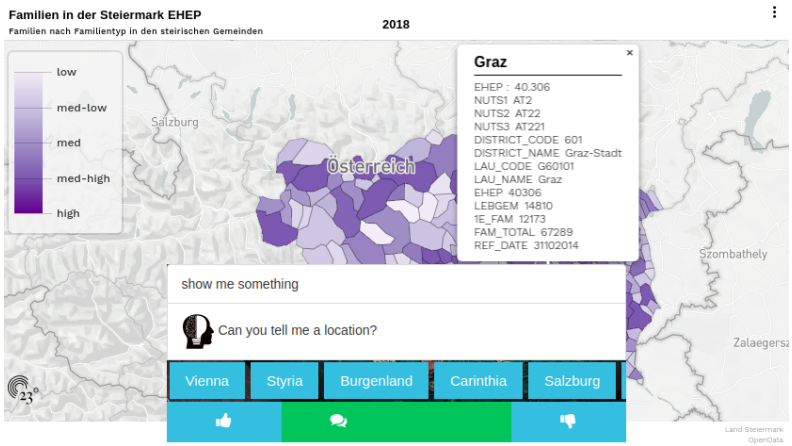

Fig. 2. Example visualization by reboting.com: Number of families in Styria, Austria. The color indicates the density; details can be displayed by selecting a subregion.

Visualizations for Open data are an active area of research. IBM [7] created a free visualization tool for user uploaded open data. Cerami et al. [1] enable everybody to explore cancer data visually. Furthermore, open data portal providers, such as `opendataportal.at`, provide their own visualization tools.

To enable non-technical users to explore the results of our search API in the context of the nine Austrian federal states, our web application `reboting.com` parses for categorical and numeric columns and then scans for geo-references and time components that can be visualized on a map or a barchart by the 23degree visualization API.[10] In theory the amount of visualizations that can be generated is huge. For one visualization type it is the amount of string columns multiplied with the amount of numeric columns. To reduce the space of possible visuals we use the string column with the maximum string length as sum of

[10] https://23degrees.io

all rows. To further reduce possible visuals we assigned random colors and for maps we use a fixed distribution for the legend depending on the minimum and maximum of the visualized numeric column. With this approach we generated 6117 visuals for 393 datasets. Out of 1321 results for the nine Austrian federal states, 928 results could not be visualized either because of download/parsing errors, or because of the structure of the data. The example visualization in Fig. 2 illustrates the possibility for users to rate shown visualizations. We save the rating information and plan to use it to evaluate and improve our visual generation process. Furthermore, we are evaluating the use of A/B testing [3] to further improve on our knowledge of visualization preferences. Currently, the user cannot choose a specific dataset and the visualization is limited to a random selection. We chose this random approach to get some initial ratings, which will hopefully shed some light on dataset features that are useful for meaningful visualizations.

3 Conclusions & Outlook

Herein, we have presented `reboting.com`, an interface that allows geo-queries for Austrian federal states and automatically generates visualization of respective open datasets. The geo-labelling of the dataset is based on a base knowledge graph of geo-entities. In future versions of this framework, we plan to integrate information gathered from the user ratings of the visualizations: In case of inadequate representations we will adapt the visualization (i.e. change the input columns). Also, we plan to scale our systems to datasets/data portals worldwide, so that users can query for any geo-entity/location. Complementary, users might benefit from other dimensions such as temporal and topic filters.

References

1. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al.: The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data (2012)
2. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, adoption barriers and myths of open data and open government. Information Systems Management 29(4), 258–268 (2012), https://doi.org/10.1080/10580530.2012.716740
3. Kohavi, R., Longbotham, R.: Online controlled experiments and a/b testing. In: Encyclopedia of Machine Learning and Data Mining, pp. 922–929. Springer (2017)
4. Neumaier, S., Umbrich, J., Polleres, A.: Automated quality assessment of metadata across open data portals. J. Data and Information Quality 8(1), 2:1–2:29 (Oct 2016), http://doi.acm.org/10.1145/2964909
5. Perry, M., Herring, J.: OGC GeoSPARQL - A geographic query language for RDF data. OGC Implementation Standard. Sept (2012)
6. Stadler, C., Lehmann, J., Höffner, K., Auer, S.: Linkedgeodata: A core for a web of spatial open data. Semantic Web 3(4), 333–354 (2012)
7. Viegas, F.B., Wattenberg, M., Van Ham, F., Kriss, J., McKeon, M.: Manyeyes: a site for visualization at internet scale. IEEE transactions on visualization and computer graphics 13(6) (2007)