A spatio-temporal Linked Data Representation for modeling spatio-temporal Dialect Data

Johannes Scholz, Emanual Hrastnig and Eveline Wandl-Vogt

Abstract Collections of linguistic and dialect data often lack a semantic description and the ability to establish relations to external datasets, from e.g. demography, socio-economics or geography. Based on existing projects - the Database of Bavarian Dialects in Austria and exploreAT! - this paper elaborates on a spatio-temporal Linked Data model for representing linguistic/dialect data. Here we focus on utilizing existing data and publishing them using a virtual RDF graph. Additionally, we exploit external datasources like DBPedia and geonames.org, to specify the meaning of dialect records and make use of stable geographical placenames. In the paper we highlight a spatio-temporal modeling and representation of linguistic records relying on the notion of a discrete lifespan of an object. Based on a real-world example - using the lemma ”Karotte” (engl. carrot) we show how the usage of a specific dialect word (”Karottn”) changes from 1916 until 2016 - by exploiting the expressive power of GeoSPARQL.

1 Introduction and Motivation

Language Geography and Geolinguistics are concerned with the geographic distribution of language(s) or its constituent elements. It is a field that strives to enhance the usability of digital language databases, and that works towards a visual explora-

Johannes Scholz
Graz University of Technology, Institute of Geodesy, Steyrergasse 30, 8010 Graz, Austria, e-mail: johannes.scholz@tugraz.at

Emanuel Hrastnig
Graz University of Technology, Institute of Geodesy, Steyrergasse 30, 8010 Graz, Austria, e-mail: hrastnig@student.tugraz.at

Eveline Wandl-Vogt
Austrian Academy of Sciences, Austrian Centre for Digital Humanities, Wohllebengasse 12-14/2, 1040 Wien, Austria, e-mail: eveline.wandl-vogt@oeaw.ac.at

tion of linguistic data. Languages and dialects are present in space and are mostly represented as language areas (Chambers & Trudgill, 1998).

In language geography, and especially dialectology, the basis for creating language maps are field surveys. Examples for this approach are the Wenker Atlas (Schmidt & Herrgen, 2001) or the Dictionary of Bavarian Dialects in Austria (Österreichische Akademie der Wissenschaften & Bauer, 1985). Field surveys are questionnaires that were answered by teachers or other trained persons from 1887-1888 (Wenker Atlas) and from 1913 onwards (Dictionary of Bavarian Dialects in Austria). Thus, each questionnaire is connected to a specific place - i.e. where the person lives/d and collected evidence. Subsequently, each identified dialect record - at least the word, the pronunciation and its meaning - is connected to a location. Recently, these data have been digitized and stored electronically, using contemporary object-relational database technology. These data can be analyzed by linguists who create maps with isoglosses, dialect continua and finally language dictionaries. Hence, the approach in this paper is concerned with basic data, that are necessary to create more advanced "products" - like language maps.

Nevertheless, as these basic data are not opened up for the public, it remains hard to combine other language data sets and/or to compare them with historic socio-economic or demographic datasets. Especially as most linguistic datasets lack a semantic description and the use of shared vocabularies such as e.g. place names. Yet, the ongoing project exploreAT! (exploring austrias culture through the language glass; Austrian Academy of Sciences; 2015-2019) is going to open up the data sets, interlink existing concepts, make use of semantic technologies and make citizens part of the scientific process. Bird, Klein, and Loper (2009) formulated three fundamental questions concerning the design and distribution of language resources. Of these three questions, number 3 is of importance for the current paper: "What is a good way to document the existence of a resource we have created so that others can easily find it?" (Bird et al., 2009, p. 407).

We propose a spatio-temporal Linked Data approach to model and publish data on linguistics and dialects. As there are a number of local linguistic data sets in Austria (e.g. dialect database of Upper Austria [1], dialect database of Salzburg [2]) a Linked Data approach helps integrate different datasets in an ad-hoc manner and to facilitate an integrated analysis of different datasets. Based on the Dictionary and Database of Bavarian Dialects in Austria (DBÖ) (Österreichische Akademie der Wissenschaften & Bauer, 1985), and the results of the research project "Dictionary Bavarian Dialects in Austria electronically mapped" (e.g. Scholz et al., 2008) our objective is to develop a spatio-temporal Linked Data representation for dialect data. In addition, we present preliminary results of the Linked Data approach that enable spatio-temporal query capabilities in conjunction with external Linked Data sets, utilizing data originating from the DBÖ and dialect database of Salzburg.

We elaborate on relevant work in Section 2. Section 3 deals with the approach to develop a Linked Data representation for the linguistic and dialect data with a

[1] http://www.stifter-haus.at/sprachforschung

[2] https://www.sprachatlas.at/salzburg

focus on the DBÖ. Subsequently, we elaborate on preliminary results in Section 4 and critically discuss them in Section 5.

2 Related Work and Background

An overview of mapping techniques in the field of linguistics and dialectology is given in Lameli, Kehrein, and Rabanus (2010). Contemporary atlases on dialectology and/or languages present their data using point symbols or thematic maps (e.g. Schmidt & Herrgen, 2001). An additional element often used in language maps are isoglosses and isographs - critically discussed by Pi (2006). Some papers suggest the usage of honeycomb maps around observation points (similar to Voronoi diagrams or Delaunay triangulation) (Goebl, 2010; Nerbonne, 2010). Rumpf, Pickl, Elspaß, König, and Schmidt (2010) proposed an analysis of language data using kernel density estimation and elaborated on geographical similarity evaluations on area-class maps.

In the field of linguistics several publications utilize GIScience methods to analyze linguistic data. However, only a handful papers in GIScience deal with linguistics. Among those papers are publications by Hoch and Hayes (2010), Sibler, Weibel, Glaser, and Bart (2012), and Scholz, Lampoltshammer, Bartelme, and Wandl-Vogt (2016). Jeszenszky and Weibel (2015) postulate four research questions to analyze and describe the nature of language boundaries.

Buccio, Nunzio, and Silvello (2014) describe an approach to publish linguistic data of the Syntactic Atlas of Italy, but do not mention any spatial and temporal modeling and/or analysis capabilities . A number of geolinguistic and linguistic projects are of interest for this paper. The first ontology designed to support the publishing and description of linguistic data in the semantic web is mentioned in Farrar and Langendoen (2003). An ontology-based mapping between different linguistic datasets is presented in Chiarcos et al. (2008). Xie et al. (2009) present an outcome of the research project LL-Map, highlighting the integration of language related data with data from the physical and social sciences with the help of a GIS. In addition the Open Linguistics Working Group is working towards a Linguistic Open Data Cloud, making use of semantic web methodologies (Chiarcos, Hellmann, & Nordhoff, 2011). Lee and Hsieh (2015) present an example of linguistic Linked Data by publishing the Chinese Wordnet as part of the Linguistic Linked Data Cloud. Frontini, Gratta, and Monachini (2016) report on the transformation of GeoNames ontology concepts, into a GeoDomain WordNet-like resource in English, and its translation into Italian.

3 Linked Data 4 Dialects: Concept & Development Approach

The approach followed here is based on an existing relational database model developed for the DBÖ. Since this database serves as the main storage of linguistic data, we use it as much as possible, avoiding redundant storage of data sets wherever possible. Based on the relational data model we developed an ontology, for modeling and representing geolinguistic resources. The OWL ontology is divided in three parts: derivation, tagging and geographic. The derivation part deals with people speaking a language, whereas the geographic part deals with the locations where a certain language is spoken (and by whom). The tagging part is concerned with language specific classes, properties, like documents, questions, lemma or meaning. The basic structure is given in Figure 1. The most important classes of the ontology are source, record, lemma, meaning, location, time and geodata. The class *lemma* contains the canonical form, dictionary form, or citation form of a set of words. Each individual of the class *record* is related to a lemma, and has a certain meaning as well. This is necessary, as each usage of a lemma is embodied in a context that influences the meaning. An example is the lemma "mouse", which can be used in the context of computers or biology. The class *source* contains the source (evidence) of each record. The DBÖ relies on a database of 5 million paper sources (vouchers with dialect words), which were digitized from the early 1990ies.

The ontology inherits vocabulary from other domains and uses its own domain *dboe*. The inherited vocabularies are *geonames* and *DBpedia* for now. *DBpedia* is used to define the meaning of a dialect record. In the future we plan to include *BabelNet* or *Wikidata* and connect with historical gazetteers, e.g. the project PELAGIOS [3]. *Geonames* is used to reference place and region names contained in the linguistic datasets.

The spatio-temporal context is related to each source - i.e. to each voucher. Each source has a certain location, as each dialect word is spoken at a specific physical place. In addition, a location has three subclasses, not depicted in Figure 1 : town, community, region. Towns are populated places represented as points, whereas communities and regions are polygons. Towns and communities are inherited from geonames.org. Regions are defined from a linguistic perspective and are not identical with administrative regions. Hence, spatial data on linguistic regions cannot be inherited from an external source. Since language has a dynamic nature (see e.g. Wandl-Vogt, 1997; Birlinger, 1890; Nerbonne, 2010), language phenomena may move, emerge, end, expand or shrink - similar to other real world objects (e.g. Nixon & Stewart Hornsby, 2010). Thus we added a valid time span to each source, representing the timeframe a specific word is found at a location. This approach is intended to represent the temporal changes in linguistic phenomena - using a discrete representation. Thus, it is not possible to model a gradual change with this approach. Currently, this fulfills the requirements of linguistic phenomena, as surveys are not done in a continuous, high frequent manner.

[3] http://commons.pelagios.org

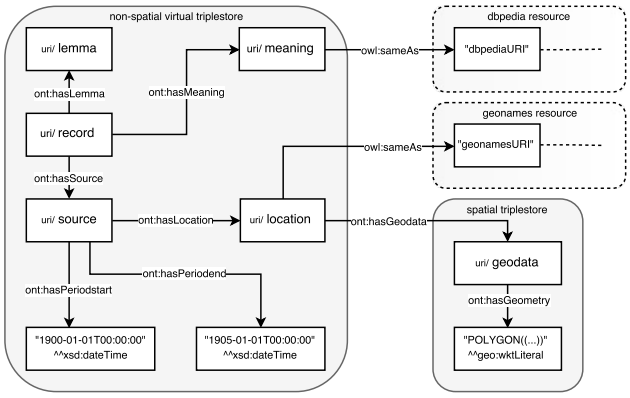

Fig. 1 Excerpt of the developed Ontology - showing only the important classes their relationships, and the inherited vocabulary from external sources. The spatio-temporal aspect is modeled as time period of each source, and the spatial data available for each location associated to each source document.

4 Preliminary Results

The preliminary results are a proof-of-concept implementation and spatio-temporal SPARQL queries (i.e. GeoSPARQL), based on the data present in the DBÖ. The implementation publishes the existing dialect data as SPARQL endpoint with the help of virtual RDF graph (Bizer & Seaborne, 2004). Figure 2 depicts the architecture of the proof-of-concept implementation. We utilize the existing relational database with the help of a virtual RDF graph using D2RQ[4]. Spatial data on the linguistic regions of the DBÖ are published in a spatial triple store - here Strabon [5].

Preliminary results are based on existing datasets of the DBÖ and the Salzburger Sprachatlas (dialect database of Salzburg) [6]. Here we are focusing on a dataset describing the dialect representation of ”carrot” in the province of Salzburg. In the specific Bavarian dialect a carrot (”Karotte”) is represented by the word ”Karottn”. Nevertheless, the dialect word ”Karottn” was not used in Salzburg around 1916. Figure 3 shows the spatial-temporal change of the lemma ”Karotte” (engl. carrot) between 1916 and 2016. The communities where the dialect word ”Karottn” is present in 1916 are colored in red. The communities where switched to ”Karottn” from 1916 until 1966 are colored in orange - hence the region where ”Karottn” is used in 1966 contains of red and orange areas. In 2016, the dialect word ”Karottn” is used in most communities, where the data on the lemma ”Karotte” (carrot) are present.

[4] http://d2rq.org

[5] http://www.strabon.di.uoa.gr

[6] https://www.sprachatlas.at/salzburg

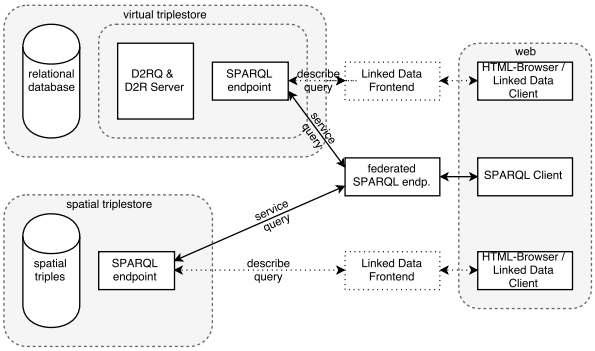

Fig. 2 System architecture of the proof-of-concept implementation.

The communities that switched to the usage of ”Karottn” from 1966 until 2016 are marked with yellow in Figure 3. Thus, the region where ”Karottn” is used in 2016, are all communities colored in yellow, orange or red.

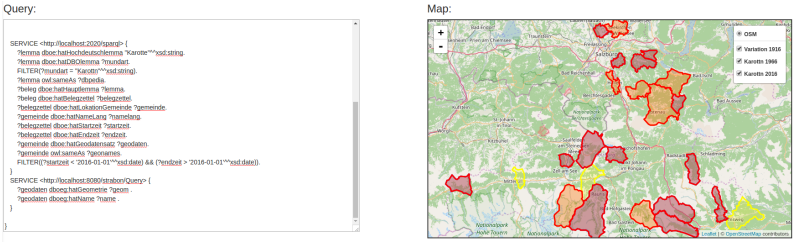

Fig. 3 Spatio-temporal analysis for the lemma ”Karotte” (carrot) for 1916-2016. The communities where dialect word ”Karottn” is present in 1916 are colored in dark red. In 1966 the communities marked with orange were added to the region where ”Karottn” is used. From 1966 until 2016 the communities marked in yellow were added to the region where ”Karottn” is spoken.

5 Discussion and Conclusion

The paper presents an approach to model and publish linguistic/dialect data as spatio-temporal Linked Data - based on the Dictionary of Bavarian Dialects in Austria. The approach followed in this paper is based on the development of an ontology for modeling and representing geolinguistic resources - focusing on dialects. This ontology forms the basis for publishing data stored in a relational data model with the help of a virtual RDF graph. The ontology models the dialect records,

their associated lemma as well as their meaning. As language is a dynamic phenomenon, we incorporate the spatio-temporal dimension in the ontology. Hence, each source (evidence) has an associated location and temporal validity. This opens up the possibility for making spatio-temporal analyses with dialect data at hand and to relate these data to other datasets published in the Linked Data cloud. A preliminary example, based on the lemma "Karotte" (carrot), shows the usage of the dialect word "Karottn" for a specific geographic area from 1916 until 2016, utilizing placenames inherited from geonames.org. Future research items - especially for the linguistic/dialect application scenario - may include the representation of gradual spatio-temporal change of e.g. linguistic phenomena, with RDF.

Acknowledgements Parts of this work were funded by Austrian Research Fund (Project Nr.: L323-G03) and Austrian Nationalstiftung (Project Nr.: DH2014/22).

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Birlinger, A. (1890). Rechtsrheinisches Alemannien. *Forschungen zur Deutschen Landes- und Volkskunde*, (4), 369–386.

Bizer, C. & Seaborne, A. (2004). D2rq-treating non-rdf databases as virtual rdf graphs. In *Proceedings of the 3rd international semantic web conference (iswc2004)* (Vol. 2004). Citeseer Hiroshima.

Buccio, E. D., Nunzio, G. M. D., & Silvello, G. (2014). A linked open data approach for geolinguistics applications. *International Journal of Metadata, Semantics and Ontologies*, *9*(1), 29–41.

Chambers, J. K. & Trudgill, P. (1998). *Dialectology*. Cambridge University Press.

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., & Stede, M. (2008). A flexible framework for integrating annotations from different tools and tagsets. *Traitement Automatique des Langues*, *49*(2), 271–293.

Chiarcos, C., Hellmann, S., & Nordhoff, S. (2011). Towards a linguistic linked open data cloud: The open linguistics working group. *TAL*, *52*(3), 245–275.

Farrar, S. & Langendoen, D. T. (2003). A linguistic ontology for the semantic web. *GLOT international*, *7*(3), 97–100.

Frontini, F., Gratta, R. D., & Monachini, M. (2016). GeoDomainWordNet: Linking the geonames ontology to WordNet. In *Human language technology. challenges for computer science and linguistics* (pp. 229–242). Springer International Publishing. doi:10.1007/978-3-319-43808-5_18

Goebl, H. (2010). Dialectometry and quantitative mapping. In A. Lameli, R. Kehrein, & S. Rabanus (Eds.), *Language and space: An international handbook of linguistic variation: Language mapping* (Vol. 2, pp. 433–457).

Hoch, S. & Hayes, J. J. (2010). Geolinguistics: The incorporation of geographic information systems and science. *The Geographical Bulletin*, *51*(1), 23.

Jeszenszky, P. & Weibel, R. (2015, April 23). Measuring boundaries in the dialect continuum. In *Proceedings of the AGILE Conference on Geographic Information Science 2015*. Springer International Publishing.

Lameli, A., Kehrein, R., & Rabanus, S. (Eds.). (2010). *Language and space: An international handbook of linguistic variation: Language mapping*. Berlin: De Gruyter Mouton.

Lee, C.-Y. & Hsieh, S.-K. (2015). Linguistic linked data in chinese: The case of chinese wordnet. In *Proceedings of the 4th workshop on linked data in linguistics (ldl-2015),* (pp. 70–74). Association for Computational Linguistics and Asian Federation of Natural Language Processing.

Nerbonne, J. (2010). Mapping aggregate variation. In A. Lameli, R. Kehrein, & S. Rabanus (Eds.), *Language and space: An international handbook of linguistic variation: Language mapping* (Vol. 2, pp. 476–495). Mouton De Gruyter.

Nixon, V. & Stewart Hornsby, K. (2010). Using geolifespans to model dynamic geographic domains. *International Journal of Geographical Information Science*, *24*(9), 1289–1308.

Österreichische Akademie der Wissenschaften & Bauer, W. (1985). *Wörterbuch der bairischen Mundarten in Österreich (WBÖ)*. Verlag der Österreichischen Akademie der Wissenschaften.

Pi, C.-Y. T. (2006). Beyond the isogloss: Isographs in dialect topography. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, *51*(2-3), 177–184.

Rumpf, J., Pickl, S., Elspaß, S., König, W., & Schmidt, V. (2010). Quantification and statistical analysis of structural similarities in dialectological area-class maps. *Dialectologia et Geolinguistica*, *18*(1), 73–100.

Schmidt, J. E. & Herrgen, J. (2001). Digitaler Wenker-Atlas (DiWA). *Bearbeitet von Alfred Lameli, Tanja Giessler, Roland Kehrein, Alexandra Lenz, Karl-Heinz Müller, Jost Nickel, Christoph Purschke und Stefan Rabanus. Erste vollständige Ausgabe von Georg Wenkers Sprachatlas des Deutschen Reichs*.

Scholz, J., Bartelme, N., Fliedl, G., Hassler, M., Mayr, H., Nickel, J., … Wandl-Vogt, E. (2008). Mapping languages–erfahrungen aus dem projekt dbo@ ema. *Angewandte Geoinformatik*, 822–827.

Scholz, J., Lampoltshammer, T. J., Bartelme, N., & Wandl-Vogt, E. (2016). Spatial-temporal modeling of linguistic regions and processes with combined indeterminate and crisp boundaries. In *Progress in Cartography* (pp. 133–151). Springer.

Sibler, P., Weibel, R., Glaser, E., & Bart, G. (2012). Cartographic visualization in support of dialectology. *Proceeding AutoCarto 2012*.

Wandl-Vogt, E. (1997). *Alemannisch-Bairische Interferenzen im Dialekt des Tiroler Paznauns. Eine Annäherung an Mundartgrenzen: Entwicklung, Verlauf, Beurteilung* (Master's thesis, Vienna).

Xie, Y., Aristar-Dry, H., Aristar, A., Lockwood, H., Thompson, J., Parker, D., & Cool, B. (2009). Language and location: Map annotation project-a gis-based infrastructure for linguistics information management. In *Computer science and information technology, 2009. imcsit'09. international multiconference on* (pp. 305–311). IEEE.