Spatial Linked Data Approach for Trace Data in Digital Humanities

Franziska Hübl[1] and Johannes Scholz[1][0000−0002−3212−8864]

Graz University of Technology, Institute of Geodesy, Research Group Geoinformation, Steyrergasse 30/I, 8010 Graz, Austria
{franziska.huebl,johannes.scholz}@tugraz.at

Abstract. Question answering in Geohumanities within the Semantic Web is a broad re-search field. This work incorporates the investigation, use and extension of existing patterns and ontologies having a space and time dimension to combine historical Linked Places and Linked Traces data. Linked Traces are way to collect historical events and places in a so-called trace, in order to describe events that consist of several single places/events – e.g., the expeditions of Charles Darwin. The paper deals with the integration of Linked Data repositories like World Historical Gazetteer for trace data. By linking different historical data sources new historical findings can be inferred. That can be managed by creating a Knowledge Graph, a semantically enriched contextual triplestore, which functions as a basis for e.g., Geographic Question Answering. The data-driven method of creating a Knowledge Graph can be realized by combining ontologies and data sources of interest. The resulting spatial Linked Data approach is tested and evaluated with regard to geosemantic capabilities in form of a Knowledge Graph investigation. To follow the motivation of "linking" a SPARQL Endpoint is used to show a standardized proof-of-concept implementation.

Keywords: Semantic Web · Linked Data · Knowledge Graph · Trace Data.

DOI: https://doi.org/10.25436/E2T882.

1 Introduction

As [1] highlighted, "The Semantic Web isn't just about putting data on the web". It is about making the web grow to a single global data space on the basis of a formalized knowledge representation. Therefore, data is structured as objects or entities and stored within tree or graph structures to make it linkable and easy to access. One example of such data linking, including also spatio-temporal data, is described by [19]. The approach focuses on utilizing existing dialect data and publishing it by using a virtual Resource Description Framework (RDF) graph. The Linked Open Data Cloud already includes data in form of Knowledge Graphs (KG) like DBpedia [3] and thus makes data available as

structured knowledge bases consisting of data triples (subject, predicate, object). By combining Linked Data and Geoinformation a geospatially enriched Semantic Web comes into existence [5, 10]. Basically, a KG is a combination of ontologies and data. Thus, for Geographic Question Answering semantically enriched contextual data are required [12]. Within this work especially the Linked Pasts Ontology [11] is used as basis to create a Geographic KG (GeoKG) for historical events and traces. The research question of addressed here is as follows: "Which extensions to the Linked Pasts Ontology are necessary to integrate the Linked Pasts Ontology, Linked Place and Linked Traces format into a single GeoKG supporting historic trace data.". The challenge is to support the Linked Places and Linked Traces format [9, 7] within the GeoKG approach, while reusing existing patterns and ontologies. Additional objectives are to evaluate the geosemantic capabilities by investigating the GeoKG and to perform spatio-temporal queries to demonstrate the usability of the GeoKG with respect to question answering in the Geohumanities. To evaluate the approach a test data set is developed and published with a standardized GeoSPARQL Endpoint.

2 Background

The Linked Pasts Ontology is based on given standards like the Comité international pour la documentation (CIDOC) Conceptual Reference Model (CRM) ontology [4] and lawdi/LAWD [2]. The Linked Places format was defined by Grossner et al. [7]. A detailed description of the format and sample datasets are available on GitHub. The goal behind the Linked Places format is to link different gazetteers in a uniform way and include just enough metadata to support the requirements to search across different gazetteers, find enough information to identify and disambiguate places and annotate data with stable Uniform Resource Identifiers (URI). The Linked Places format supports the Pelagios Gazetteer Interconnection Format (PGIF) which itself serves as a template for contributions for Pelagios [16] and World-Historical Gazetteer [23]. The Linked Places syntax is primarily based on JavaScript Object Notation for Linked Data (JSON-LD). That incluedes RDF as well as Extensible Markup Language (XML), Turtle etc. and JavaScript Object Notation (JSON) valid frameworks, languages, and formats. It also includes the GeoJSON format, which is necessary for spatial data. In fact, the GeoJSON-T format is implemented, which extends GeoJSON and standardizes the representation of temporal attributes. Linked Traces are a continuation of the Linked Places model. It is a W3C Web Annotations format and a proposal for a set of patterns for use by historical researchers and systems – and aims at describing overarching events (e.g., the expeditions of Charles Darwin) with linked fine-grained events and/or places. The Linked Places format is annotating web resources with identifiers for places relevant to the phenomena they describe. That might be historical entities like artifacts, events, people or works. World Historical Gazetteer intends to develop the format further and uses already example traces within the platform [7].

3 The Linked Traces Knowledge Graph Approach

According to literature, KGs are usually created by starting with the data basis and defining requirements for the questioning. Because of the requirement to integrate Linked Places and Traces formats in form of sample data, we follow existing methods. During the process of preparing, integrating, and mapping the data, the limits of the ontology are surfacing. Hence, we choose to create the GeoKG within GraphDB [14] by mapping the data and extend the Linked Pasts Ontology as needed. The main steps of this data-driven approach are described in the following sections.

3.1 Steps to Build the Geo-Knowledge Graph

Geo-Knowledge Graph Requirements. The requirements are mostly defined within the research objectives. The GeoKG builds upon the Linked Pasts Ontology and include the Linked Places and Linked Traces format. Further the GeoKG should be linkable and usable to query spatio-temporal data concerning historic events. These requirements are handled by integrating the available Linked Pasts Ontology, and spatio-temporal test datasets of places and traces.

Test Data and Data Pre-Processing. The focus of this work is on the integration of datasets published in the Linked Places and Linked Traces format. Thus, the basis for the GeoKG is the available sample datasets on GitHub generated by Grossner et al. [8]. The Linked Places datasets originate from different data sources themselves, like WHGazetteer, ToposText and Historical Geographic Information System (HGIS) de las Indias [23, 21, 20]. Auxiliary datasets can be considered as suitable if they are linkable to the given places and traces. Thus, GeoNames [6] and Natural Earth [13] countries, plus Art & Architecture Thesaurus (AAT) feature types [18] are additionally included in order to merge different data sources within a GeoKG - which in turn is utilized for spatial queries. The GeoSPARQL plugin of GraphDB supports two geometry formats, GML and WKT, thus, data sources with geometries need to be adapted where appropriate before the data integration.

Extending the Linked Pasts Ontology. Basically, to harmonize the data, a semantic data model has to be created by analyzing different data schemata. In this case the given Linked Pasts Ontology is reused, and formalized by using standards like RDF Schema and OWL. Fig. 1 depicts the latest Linked Pasts Ontology as a visual graph. Within the Linked Pasts Ontology, a place is defined as “Attestation” and traces are not yet included. Thus, the classes “Place” and “Trace” are added. By taking a closer look at the trace data, the need of adding an “Event” class needs to represent the relation between places and traces. An event could also be handled as “historical entity” of type “event” in the case of further development. That would give the opportunity to define more types like “artifact”, “person” or “work”. For now, the defined event class is a transitional

solution to show the relation between traces and places. We are aware of the fact that an event ontology exists [22], which is planned to be integrated later. As of now we tested the capabilities to related traces and places accordingly.

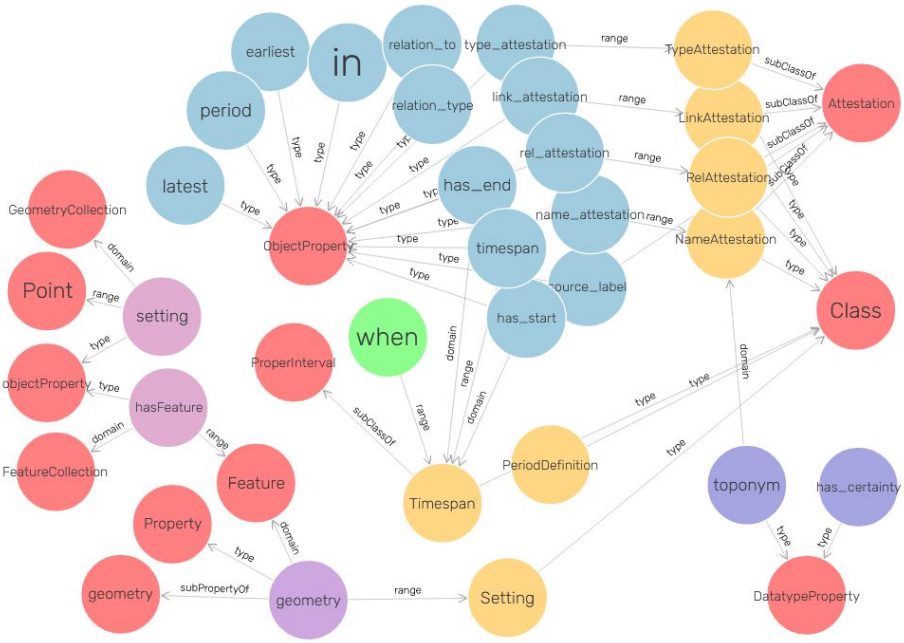

Fig. 1. Linked Pasts Ontology represented as visual graph in GraphDB. Most elements are already reused from existing ontologies - e.g. "when" is part of the Linked Pasts Ontology, taken from OWL's Time Ontology.

Data Integration and Data Mapping. For data integration and mapping GraphDB was used. The Linked Pasts Ontology and the Linked Traces extension are both integrated in the GraphDB repository as Turtle files. The OntoRefine tool is employed to import and map the CSV and JSON-LD test datasets. To generate RDF triples out of JSON sources the starting point is the root of the JSON object. The root is treated as the triple subject, the field names are the according predicates, and the values are the objects. If the value holds a JSON object, the according field name is a new starting point and thus a new subject. Thus, also lists or arrays can be mapped accordingly.

Merging the Graphs. The GeoKG subjects, but also objects and predicates, can be seen as "links". This means that e.g. a place identifier can function as a subject leading to objects, but a place identifier could also be an object of e.g.

an event that happened at a certain place. Those GeoKG "links" are used to merge different data sources, which are stored as separated graphs in GraphDB. In the case of Linked Places, the identifiers are URIs leading to e.g. the related place description.

Knowledge Graph Investigation. With the merged GeoKG, data can be queried with respect to spatio-temporal competency questions. One example competency question is: "To ask for a given trace and find all related events at certain places with their start and end time, ordered by time to show the historical process". Therefore, the GraphDB Query tool is used.

3.2 GeoSPARQL Endpoint

An important step to follow the Linked Data principles [1] is to publish the GeoKG in a standardized way, make it available and linkable. In order to test the compatibility and performance [17] with at least one other GeoSPARQL endpoint than GraphDB, we export the data as RDF-XML file and import it into Virtuoso Web Application Server [15]. For now, the GeoSPARQL Endpoint runs locally and queries are executed on a named graph IRI.

4 Results and Outlook

The paper presented intial results, highlighting extensions to the Linked Pasts Ontology that are necessary to integrate the Linked Pasts Ontology, Linked Place and Linked Traces format into a single GeoKG supporting historic trace data. With the Linked Pasts Ontology as a basis, which already includes the CIDOC CRM and lawdi/LAWD ontologies, we extended the Linked Pasts with several classes plus relevant object properties. The extension of the Linked Pasts Ontology created within this work is kind of a "workaround" to make the data mapping of the given "trace data" possible. One suggestion to adapt the Linked Pasts Ontology is to define a trace of events, which are at a certain place. This solution needs an additional event identifier or in general, the definition of a historical entity of type "event" - which could be based on the simple event ontology [22]. In addition, a "Place" is corresponding with "Attestation". One place could hold for instance more than one event taking place at different time like shown in Fig. 2. The place "Babylon" is part of five different events concerning three traces. Without the extension of the Linked Pasts Ontology described in this paper, such a representation would not be possible. The GeoKG approach indicates that the Linked Places format holds enough linkable object values to integrate it into a GeoKG. By taking a look at the geosemantic capabilities, the given Linked Places format is providing possibilities to include all kinds of descriptions and links, and also different geometry types and formats. Thus, in terms of geosemantics, places are basically well describable and can be enriched with context as needed. Also, in terms of time relevant changes of e.g. places, events, names, the format offers lots of possibilities to describe the process. In

contrast, the Linked Traces format needs further adaption to satisfy the overall requirements, e.g., by defining historical entities and their types.

A future research task include the integration of entire project data sets concerning historical events - and to publish them as GeoKG for the scientific community. The shown approach indicates that such a step would be worthwhile in order to gain and share further knowledge in the field of Geohumanities. In addition, the integration and utilization of relevant ontologies and ontology design patterns - like e.g. the simple event ontology [22] - are of particular interest. Additionally, we strive to critically evaluate the approach with a more structured approach - e.g. a defined set of competency questions - that could be defined together colleagues from the Humanities. From a GIScience perspective, the integration and definition of "places" and their spatio-temporal definition especially the context of historical places could be of interest for the GIScience and the Geohumanities community as well.

Fig. 2. Events (blue) of certain traces (red) taking place at "Babylon" represented as visual graph in GraphDB.

References

1. Berners-Lee, T.: Linked Data - Design Issues (2006), https://www.w3.org/DesignIssues/LinkedData.html
2. Cayless, H., Heath, S., Dubin, D.: Lawd: An ontology for linked ancient world data (2016), https://github.com/lawdi/LAWD
3. DBpedia Association: Home - DBpedia Association (2021), https://www.dbpedia.org/
4. Doerr, M.: The CIDOC CRM – An Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, Special Issue on Ontologies (2002)
5. Egenhofer, M.J.: Toward the semantic geospatial web. Proceedings of the ACM Workshop on Advances in Geographic Information Systems pp. 1–4 (2002). https://doi.org/10.1145/585147.585148
6. GeoNames: Geonames country codes (2021), https://www.geonames.org/countries/
7. Grossner, K., Simon, R., Light, R., Scholz, J.: Linked traces annotations v0.2 (2019), https://github.com/LinkedPasts/linked-traces-format
8. Grossner, K., Zijdeman, R., Shaw, R., Elwert, F.: Linked places data (2021), https://github.com/LinkedPasts/linked-places-format/tree/master/data
9. Grossner, K., Zijdeman, R., Shaw, R., Elwert, F.: Linked places format (2021), https://github.com/LinkedPasts/linked-places-format
10. Janowicz, K., Scheider, S., Adams, B.: A geo-semantics flyby. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **8067 LNAI**, 230–250 (2013). https://doi.org/10.1007/978-3-642-39784-4_6, https://link.springer.com/chapter/10.1007/978-3-642-39784-4_6
11. Light, R., Grossner, K.: Linked pasts ontology (2021), https://github.com/LinkedPasts/linked-pasts-ontology
12. Mai, G., Janowicz, K., Zhu, R., Cai, L., Lao, N.: Geographic question answering: Challenges, uniqueness, classification, and future directions. AGILE: GI-Science Series **2**, 1–21 (6 2021). https://doi.org/10.5194/AGILE-GISS-2-8-2021, https://doi.org/10.5194/agile-giss-2-8-2021
13. Natural Earth: Natural earth: 1:50m cultural vectors (2021), https://www.naturalearthdata.com/downloads/50m-cultural-vectors/
14. Ontotext: Graphdb™ - ontotext (2021), https://www.ontotext.com/products/graphdb
15. OpenLink Software: Openlink software: Virtuoso homepage (2021), https://virtuoso.openlinksw.com/
16. Pelagios: Welcome to Pelagios Network (2021), https://pelagios.org/
17. Raza, A.: Comparison of geospatial support in RDF stores: Evaluation for ICOS carbon portal metadata. Master's thesis, Department of Physical Geography and Ecosystem Science, Centre for Geographical Information Systems, Lund University (2019)
18. Rouse, S.: Review of The Art and Architecture Thesaurus; Guide to Indexing and Cataloging with the Art and Architecture Thesaurus, by T. Petersen & P. J. Barnet. The American Archivist **59**(3), 367–370 (1996), http://www.jstor.org/stable/40293998
19. Scholz, J., Hrastnig, E., Wandl-Vogt, E.: A spatio-temporal linked data representation for modeling spatio-temporal dialect data. vol. 0, pp. 275–282. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-63946-8_44, https://link.springer.com/chapter/10.1007/978-3-319-63946-8_44

20. Stangl, W.: 'The Empire Strikes Back'?: HGIS de las Indias and the Postcolonial Death Star. International Journal of Humanities and Arts Computing **12**, 138–162 (10 2018). https://doi.org/10.3366/ijhac.2018.0219, https://www.euppublishing.com/doi/10.3366/ijhac.2018.0219
21. ToposText: Topostext (2021), https://topostext.org/
22. van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the simple event model (sem). Journal of Web Semantics **9**(2), 128–136 (2011). https://doi.org/https://doi.org/10.1016/j.websem.2011.03.003, https://www.sciencedirect.com/science/article/pii/S1570826811000199, provenance in the Semantic Web
23. World Historical Gazetteer: World Historical Gazetteer (2021), https://whgazetteer.org/