

Publish and Enrich Geospatial Data as Linked Open Data

Claire Ponciano¹^a, Markus Schaffert¹, Falk Würriehausen² and Jean-Jacques Ponciano¹^b

¹*i3mainz – Institute for Spatial Information and Surveying Technology, Mainz University of Applied Sciences, Lucy-Hillebrand-Straße 2, 55128 Mainz, Germany*

²*Federal Agency for Cartography and Geodesy, Richard-Strauss-Allee 11, 60598 Frankfurt am Main, Germany*


Keywords: Semantic Interpretation, Linked Open Data, SPARQL, Ontology, Spatial Data, Semantic Web, Neural Machine Translation.


Abstract: The rapid growth of geospatial data (at least 20% every year) makes spatial data increasingly heterogeneous. With the emergence of Semantic Web technologies, more and more approaches are trying to group these data in knowledge graphs, allowing to link data together and to facilitate their sharing, use and maintenance. These approaches face the problem of homogenisation of these data which are not unified in the structure of the data on the one hand and on the other hand have a vocabulary that varies greatly depending on the application domain for which the data are dedicated and the language in which they are described. In order to solve this problem of homogenisation, we present in this paper the foundations of a framework allowing to group efficiently heterogeneous spatial data in a knowledge base. This knowledge base is based on an ontology linked to Schema.org and DCAT-AP, and provides a data structure compatible with GeoSPARQL. This framework allows the integration of geospatial data independently of their original language by translating them using Neural Machine Translation.

1 INTRODUCTION

Nowadays, we live in a world lead by the information. In the geospatial domain, it exists a lot of various sources (e.g. governmental data, crowd sourcing data, Linked Open Data, etc). Several approaches (Guan et al., 2012; Gannon et al., 2020; Simon and Fröhlich, 2007) have attempted to implement frameworks in order to be able to gather and share geospatial data more easily. With the emergence of Semantic Web technologies, new approaches based on knowledge graphs (such as (Karam and Melchiori, 2013; Abbas and Ojo, 2013; Larkou et al., 2013)) have proposed to define open data structures that can be linked together in order to facilitate data integration and sharing. The major issue of data integration is to allow to enrich data with other data from other sources (such as OpenStreetMap, DBpedia, and Wikidata). Each source has its own way of structuring the information. Moreover, this information is composed of vocabulary that can be from different languages. The diversity of structures and vocabularies that can be used in geospatial information have pushed approaches to specialize in

few sources to reduce the complexity of integration. It is, therefore, necessary to homogenize the different geospatial data structure and to homogenize the different vocabularies in different languages. Such a homogenization requires to define a data structure and a vocabulary sufficiently flexible to take into account the evolution of data over time, language variations, and allowing to retrieve geospatial information as quickly as possible. To meet these needs, in this paper we propose the foundations of a framework based on Semantic Web technologies, allowing us to unify the different structures and vocabularies most used in the geospatial domain in a single knowledge base. This unification allows a universal sharing of geospatial data and facilitates collaboration between different states and organizations worldwide. The goal of this project is to enable the Federal Agency for Cartography and Geodesy in Germany to (i) bring together the geospatial information contained in the data held by the different *Bundesländer* (data in different formats and structures) into a single platform in RDF form, (ii) enrich this data and link it to the Linked Open Data, and (iii) query and use this platform to create the most informative maps possible. To reach this goal, the main contributions made are the following:

^a  <https://orcid.org/0000-0001-8883-8454>

^b  <https://orcid.org/0000-0001-8950-5723>

- ontology defining a geospatial data structure that can be universally used
- method of homogenization of the different geospatial vocabularies
- integration of heterogeneous geospatial data into linked data.

The article is organized as follows: Section 2 presents the state of the challenges related to the creation of such a framework. Section 3 presents the proposed solutions and section 4 concludes with a discussion and future work to be done.

2 RELATED WORK AND PROBLEM STATEMENTS

Storing geospatial data for widespread use and sharing is a challenging research problem, primarily because of the rapid growth of data, at least by 20% every year (Lee and Kang, 2015).

In (Jung et al., 2013), the authors present an ontology-enabled framework to find geographic services. They aim to solve geospatial problems. The presented ontology for GIS data has the benefit of respecting “OGC Abstract Specification Topic 5: Feature” that stipulates a geospatial feature should contain a particular type of geometry with an SRS and attributes. However, their ontology does not provide a common vocabulary to describe geographic features and attributes. In (Sun et al., 2019), the authors present the GeoDataOnt ontology that provides a general semantic basis for integrating and sharing geospatial data. This ontology has the benefit of allowing semantic issues to be resolved at a coarse level. However, the descriptions of some features of some geospatial data are not integrated or not precise enough, which limits the resolution of the corresponding feature semantic problems using this ontology. In (Budak Arpinar et al., 2006), the authors provide an interesting ontology for geospatial semantic analytics but do not deal with geospatial attributes integration and uniformization. In (Dsouza et al., 2021), the authors present the knowledge graph WorldKG, which aims at representing geographic entities in OpenStreetMap semantically. This work facilitates using OpenStreetMap, a rich source of openly available geographic information, on the Semantic Web and among Linked Open Data, thanks to its links with Wikidata and DBpedia. However, they do not deal with the integration of heterogeneous spatial data. In (Karalis et al., 2019), the authors present YAGO2geo, an extension of YAGO2 (knowledge graph based on Wikipedia, WordNet, and GeoN-

ames). It aims at adding geospatial information represented by geometries encoded by Open Geospatial Consortium standards. They also integrate other attributes; however, they add properties specific to the dataset into YAGO2geo and do not use a common vocabulary. Although these research projects do not group heterogeneous spatial data in a knowledge base at an advanced level and using a common vocabulary, projects such as (Jung et al., 2013; Budak Arpinar et al., 2006; Dsouza et al., 2021) demonstrate that intelligent data provisioning using Semantic Web methods appears to be an effective solution for simplified (from a technical point of view) linking of official and unofficial data for geodata users. Moreover, projects using ontology-based frameworks such as (Dsouza et al., 2021; Sun et al., 2019; Karalis et al., 2019) highlight the growing interest in structuring geospatial data in the form of knowledge graphs. This structuring method brings more flexibility than traditional database structures, which is essential in view of the fast evolving nature of data. Such graphs are stored in triplestores (such as Apache Marmotta ¹, Apache Jena ², Eclipse RDF4J ³, Strabon ⁴, Oracle Spatial and Graph ⁵, GraphDB ⁶, Stardog ⁷, and Virtuoso Universal Server ⁸) allowing fast read and write access by SPARQL.

In the domain of geospatial, the GeoSPARQL (Battle and Kolas, 2011) approach has greatly improved access to geospatial data from a triplestore, allowing to retrieve and update geospatial data with simplified queries. However, GeoSPARQL requires that the stored data be structured according to a particular form in order to be usable, but most of the triplestores from the LOD have not added the structure proposed by GeoSPARQL. This constraint limits the use of GeoSPARQL for the LOD.

The other main issue is the heterogeneity of the geospatial data. Usually these data are stored in spe-

¹Apache Marmotta: <https://marmotta.apache.org>, accessed on 2022-07-07

²Apache Jena: <https://jena.apache.org/documentation/tdb/>, accessed on 2022-07-07

³Eclipse RDF4J: <https://rdf4j.org>, accessed on 2022-07-07

⁴Strabon: <http://strabon.di.uoa.gr>, accessed on 2022-07-07

⁵Oracle Spatial and Graph: <https://www.oracle.com/database/technologies/spatialandgraph.html>, accessed on 2022-07-07

⁶GraphDB: <https://www.ontotext.com/products/graphdb/>, accessed on 2022-07-07

⁷Stardog: <https://www.stardog.com>, accessed on 2022-07-07

⁸Virtuoso Universal Server: <https://virtuoso.openlinks.com>, accessed on 2022-07-07

cific formats (such as GeoJSON (Butler et al., 2016), Shapefile⁹, OpenStreetMap (Ramm et al., 2014)) having their own structure and being in different languages. This variation in vocabulary poses a major challenge for the integration of data from different sources. It requires either the creation of queries specific to the original structure of the data and the language used, thus limiting the sharing and exploitation of these data, or transform the data after defining a common data schema (such as INSPIRE) and convert all other schema to the chosen one and do analyses afterwards, which adds more steps to the processing.

We can deduce from this study that (i) the use of knowledge base (in the form of ontology and triplestore) is the most flexible and promising way to store large amounts of geospatial data, (ii) that there is currently no known platform to efficiently group heterogeneous spatial data in a knowledge base with a common vocabulary for all integrated datasets, (iii) that the major challenge for the integration of geospatial data is the diversity of structures and vocabularies used to organize and express this information.

3 SOLUTION PROPOSED

In order to meet the challenge of storing vast amount of geospatial data for universal sharing, we outline a framework for integrating heterogeneous geospatial data from different sources and languages in a structured and consistent manner. Figure 1 illustrates the goal of this framework. As a position paper, we focus on how to implement this framework using existing research, rather than explaining all the mechanisms that serialized it, which is part of our future work.

In section 3.1, we highlight solution for the integration of heterogeneous geospatial data into a Universal Spatial Knowledge Base (USKB). In Section 3.2 we highlight possibility to create common vocabulary, on the basis of which universal queries can be performed on data in RDF format. This common vocabulary aims to bring together vocabularies used in data from different sources such as Linked Open Data and geospatial formats.

3.1 Automatic Integration of Heterogeneous Data into a Universal Spatial Knowledge Base

Among the various approaches developed to unify geodata structure and vocabularies, the Ordnance Sur-

⁹Shapefile: <https://www.nationalarchives.gov.uk/pronom/x-fmt/235>, accessed on 2022-07-07

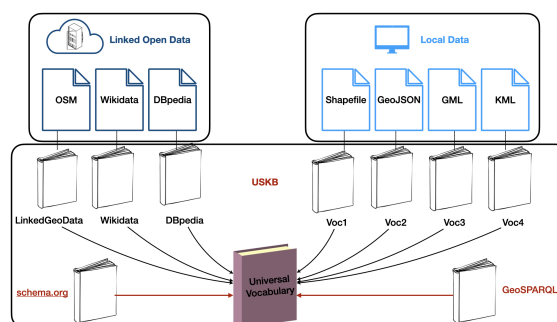


Figure 1: Goal of the proposed framework.

vey's Linked Data Platform proposes an approach using OWL (Antoniou and Harmelen, 2004) to link GeoSPARQL concepts to equivalent concepts in its ontology. GeoSPARQL has thus become a standard for representing and querying linked geospatial data for the Open Geospatial Consortium (OGC) (van Rees, 2013) Semantic Web.

Inspired by this approach, we highlight the relevance of creating an ontology combining the concepts of GeoSPARQL with the work (Quarati et al., 2021). Moreover, the work (Stadler et al., 2012) allows for describing metadata quality concepts, and link them with the LinkedGeoData.

Such an ontology aims to bring together geoinformation from different sources by converting common file formats in the geodomain and their application schemas, such as GeoJSON, Shapefile, OpenStreetMap, RDF (Pan, 2009). The work (Patroumpas et al., 2014) proposed such conversion without considering other ontologies. Thus, they are limited to the vocabulary used in the data file for creating RDF concepts. It is therefore necessary to adapt the RDF vocabulary and structure to those of the ontology for which we intend to integrate them.

The RDF structure adaptation can be performed with the help of an alignment of the ontology structure as proposed in (Li et al., 2008) using automatic integration approaches developed in the works (Prudhomme et al., 2020; Prudhomme et al., 2017).

However, such approaches required different parametrisation according to the data source quality. This quality mainly depends of the source type. Therefore an evaluation of data quality is needed to guide the integration process.

As an exemple, such evaluation can be based on the "5-star Open Data"¹⁰ principle as follows:

1. 1 star: Data available on the Web (in any format) under an open license,
2. 2 stars: Data available as structured data (spread-

¹⁰<https://5stardata.info/en/>, accessed on 2022-07-07

sheet) in a proprietary format,

3. 3 stars: Data available in a non-proprietary open format (such as CSV),
4. 4 stars: Data available linked to URIs that denote things,
5. 5 stars: Data available and linked to other data to provide context.

Automatic integration of data into an ontology without human supervision can lead to an inconsistent structure due to the accumulation of errors. Thus, to ensure the consistency of the added data, each concept is described using OWL2 (Consortium et al., 2012) constraints. These constraints allow to verify that the added data is consistent with the structure imposed by the ontology. This consistency check is performed using reasoner (such as HermiT (Glimm et al., 2014), OwlGres (Stocker and Smith, 2008), Pellet (Sirin et al., 2007), DeLorea (Bobillo et al., 2012)).

3.2 Universal Vocabulary

The data vocabulary provided varies depending on the source of the Linked Open Data (e.g. Wikidata (van Veen, 2019), DBpedia¹¹) and the data format (e.g. GML, KML, GeoJSON). The establishment of a common vocabulary allows for gathering geospatial knowledge in one place and thus, facilitating the retrieval of geographic data. Access to the data knowledge can then be done via a web interface.

To be universal, this language can be based on the English lexicon provided by WordNet (Miller, 1995). Related works such as (Frontini et al., 2016; Bond and Bond, 2019) have succeed in linking GeoNames Ontology and WordNet. Other approaches such as Schema.org (Guha et al., 2016) and KBpedia¹² are fully exploiting the power of linked open data to provide a varied and structured vocabulary in the English language. Recently DCAT- AP (Kirstein et al., 2019) and GeoDCAT-AP¹³ are planning to link their knowledge to Schema.org¹⁴, making Schema.org a promising base for the geospatial domain. Moreover, Schema.org currently support complex geometries and WKT literals, which would make the adoption of GeoSPARQL straightforward.

Since the vocabulary comes from various languages, we propose to use a Neural Machine Trans-

lation approach to automatically translate each term into English while being able to automatically detect the original language. The works (Koehn, 2020) and (Stahlberg, 2020) provide a review of different Neural Machine Translation approaches. Recently, the work of (Zhao et al., 2021) proposes to improve these approaches by combining them with a knowledge graph, thus improving the accuracy of the translation.

The translated vocabulary is then aligned with the knowledge base vocabulary using an ontology alignment approach such as (Zhang et al., 2014). Following this alignment, the aligned terms between the vocabulary used in the RDF files and the vocabulary of the knowledge base is submitted to the user for approval. The user can then validate and complete the vocabulary match in order to allow the integration of the data to enrich the knowledge base. The alignment of the terms is then stored to allow for better automation in the future.

Figure 2 summarizes this process of supervised learning and alignment.

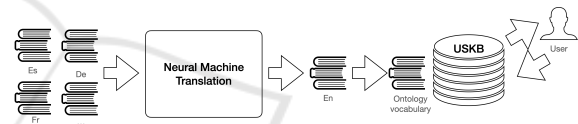


Figure 2: Overview of supervised learning and alignment process.

Let us take as an example a spatial data containing the geometries and attributes of schools in the *Bundesland Rheinland-Pfalz* in Germany in order to illustrate the proposed solution. In this dataset, a school has a geometry represented by a point and the following list of attributes in German: *Öffnungszeiten*, *FaxNummer*, and *Name*. The names of these attributes will be translated into English as: opening hours, fax number and name respectively. An individual of type *dcat:Dataset* will be created to represent the data. Then for each school contained in the data, an individual will be created. This individual will be of type *geo:Feature* (from GeoSPARQL) and *schema:School* (from Schema.org). Then, an equivalence to the attributes will be matched with Schema.org in order to add the information related to this individual. In our example, the values of the attributes *Öffnungszeiten*, *FaxNummer*, and *Name* will be added respectively with the properties *schema:openingHours*, *schema:faxNumber*, *schema:name* thanks to the English translation done before. Finally, the individual will be linked to its geometry with the property *geo:hasGeometry*. Its geometry will be of type *geo:Point* and will have a value linked by the property *geo:asWKT*.

¹¹DBpedia: <https://www.dbpedia.org>, accessed on 2022-07-07

¹²<https://kbpedia.org>, accessed on 2022-07-07

¹³<https://inspire.ec.europa.eu/good-practice/geodcat-ap>, accessed on 2022-07-07

¹⁴https://www.w3.org/2015/spatial/wiki/ISO_19115_-_DCAT_-_Schema.org_mapping, accessed on 2022-07-07

4 DISCUSSION

In this position paper, we propose conceptual guidelines for the development of a framework to unify the integration of geospatial data within a Universal Spatial Knowledge Base (USKB) and to easily link geospatial data to Linked Open Data. This framework is composed of an ontology, whose concepts are mainly based on standards (such as GeoSPARQL) and well-known ontologies (as GeoNames). The various data sets contained in files (such as GeoJSON, Shapefile, OpenStreetMap, INSPIRE) are integrated into this ontology using structured mapping. The consistency of the data integration is ensured by a reasoning applied on the constraints defined for each concept, which ensures a continuous consistency of the data. The data vocabulary is translated into English from its original language by a Neural Machine Translation approach. The translated vocabulary is mapped to the ontology vocabulary. The resulting mapping is submitted to the user for verification, who can modify or approve the proposed mapping. The decisions taken are then saved as a mapping table to allow continuous learning of the geospatial vocabulary and thus improve the mapping. The future work consists in enriching the basic concepts of the ontology and the mapping of the vocabulary by supervised learning in order to make the integration process as automated as possible. Furthermore, the quality of the data integration must be thoroughly evaluated by experts. The exploitation of expert knowledge modelled in an ontology can be considered in order to automate this evaluation, e.g. by generating correspondence concepts to semantically map the different data systems used to the respective INSPIRE systems.

5 ONLINE RESOURCES

The framework ontology and source code is available at <https://github.com/JJponciano/SpaLod> from the authors.

ACKNOWLEDGEMENTS

This research is funded by the Federal Agency for Cartography and Geodesy in Germany.

REFERENCES

Abbas, S. and Ojo, A. (2013). Towards a linked geospatial data infrastructure. In *International Conference*

on Electronic Government and the Information Systems Perspective, pages 196–210. Springer.

Antoniou, G. and Harmelen, F. v. (2004). Web ontology language: Owl. In *Handbook on ontologies*, pages 67–92. Springer.

Battle, R. and Kolas, D. (2011). Geosparql: enabling a geospatial semantic web. *Semantic Web Journal*, 3(4):355–370.

Bobillo, F., Delgado, M., and Gómez-Romero, J. (2012). Delorean: A reasoner for fuzzy owl 2. *Expert Systems with Applications*, 39(1):258–272.

Bond, F. and Bond, A. (2019). Geonames wordnet (geown): extracting wordnets from geonames. In *Proceedings of the 10th Global Wordnet Conference*, pages 387–393.

Budak Arpinar, I., Sheth, A., Ramakrishnan, C., Lynn Usery, E., Azami, M., and Kwan, M.-P. (2006). Geospatial ontology development and semantic analytics. *Transactions in GIS*, 10(4):551–575.

Butler, H., Daly, M., Doyle, A., Gillies, S., Hagen, S., Schaub, T., et al. (2016). The geojson format. *Internet Engineering Task Force (IETF)*.

Consortium, W. W. W. et al. (2012). Owl 2 web ontology language document overview.

Dsouza, A., Tempelmeier, N., Yu, R., Gottschalk, S., and Demidova, E. (2021). Worldkg: A world-scale geographic knowledge graph. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4475–4484.

Frontini, F., Del Gratta, R., and Monachini, M. (2016). Geodomainwordnet: Linking the geonames ontology to wordnet. In Vetulani, Z., Uszkoreit, H., and Kubis, M., editors, *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 229–242. Cham. Springer International Publishing.

Gannon, B. M., Thompson, M. P., Deming, K. Z., Bayham, J., Wei, Y., and O'Connor, C. D. (2020). A geospatial framework to assess fireline effectiveness for large wildfires in the western usa. *Fire*, 3(3):43.

Glimm, B., Horrocks, I., Motik, B., Stoilos, G., and Wang, Z. (2014). Hermit: an owl 2 reasoner. *Journal of Automated Reasoning*, 53(3):245–269.

Guan, W. W., Bol, P. K., Lewis, B. G., Bertrand, M., Berman, M. L., and Blossom, J. C. (2012). Worldmap—a geospatial framework for collaborative research. *Annals of GIS*, 18(2):121–134.

Guha, R. V., Brickley, D., and Macbeth, S. (2016). Schema.org: Evolution of structured data on the web. *Commun. ACM*, 59(2):44–51.

Jung, C.-T., Sun, C.-H., and Yuan, M. (2013). An ontology-enabled framework for a geospatial problem-solving environment. *Computers, Environment and Urban Systems*, 38:45–57.

Karalis, N., Mandilaras, G., and Koubarakis, M. (2019). Extending the yago2 knowledge graph with precise geospatial knowledge. In *International Semantic Web Conference*, pages 181–197. Springer.

Karam, R. and Melchiori, M. (2013). A crowdsourcing-based framework for improving geo-spatial open data.

- In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 468–473. IEEE.
- Kirstein, F., Dittwald, B., Dutkowski, S., Glikman, Y., Schimmler, S., and Hauswirth, M. (2019). Linked data in the european data portal: A comprehensive platform for applying dcat-ap. In Lindgren, I., Janssen, M., Lee, H., Polini, A., Rodríguez Bolívar, M. P., Scholl, H. J., and Tambouris, E., editors, *Electronic Government*, pages 192–204, Cham. Springer International Publishing.
- Koehn, P. (2020). *Neural machine translation*. Cambridge University Press.
- Larkou, G., Metochi, J., Chatzimilioudis, G., and Zeinalipour-Yazti, D. (2013). Cloda: A crowdsourced linked open data architecture. In *2013 IEEE 14th International Conference on Mobile Data Management*, volume 2, pages 104–109. IEEE.
- Lee, J.-G. and Kang, M. (2015). Geospatial big data: Challenges and opportunities. *Big Data Research*, 2(2):74–81. Visions on Big Data.
- Li, J., Tang, J., Li, Y., and Luo, Q. (2008). Rimom: A dynamic multistrategy ontology alignment framework. *IEEE Transactions on Knowledge and data Engineering*, 21(8):1218–1232.
- Miller, G. A. (1995). Wordnet: A lexical database for english. In *Communications of the ACM*, volume 38, pages 39–41, Cham.
- Pan, J. Z. (2009). Resource description framework. In *Handbook on ontologies*, pages 71–90. Springer.
- Patroumpas, K., Alexakis, M., Giannopoulos, G., and Athanasiou, S. (2014). Triplegeo: an etl tool for transforming geospatial data into rdf triples. In *Edbt/Icdt Workshops*, pages 275–278. Citeseer.
- Prudhomme, C., Homburg, T., Ponciano, J.-J., Boochs, F., Cruz, C., and Roxin, A.-M. (2020). Interpretation and automatic integration of geospatial data into the semantic web. *Computing*, 102(2):365–391.
- Prudhomme, C., Homburg, T., Ponciano, J.-J., Boochs, F., Roxin, A., and Cruz, C. (2017). Automatic integration of spatial data into the semantic web. In *WebIST*, pages 107–115.
- Quarati, A., De Martino, M., and Rosim, S. (2021). Geospatial open data usage and metadata quality. *ISPRS International Journal of Geo-Information*, 10(1):30.
- Ramm, F., Names, I., Files, S., Catalogue, F., Features, P., Features, N., and Cars, C. (2014). Openstreetmap data in layered gis format. *Version 0.6, 7*.
- Simon, R. and Fröhlich, P. (2007). A mobile application framework for the geospatial web. In *Proceedings of the 16th international conference on World Wide Web*, pages 381–390.
- Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: A practical owl-dl reasoner. *Journal of Web Semantics*, 5(2):51–53.
- Stadler, C., Lehmann, J., Höffner, K., and Auer, S. (2012). Linkedgeodata: A core for a web of spatial open data. *Semantic Web*, 3(4):333–354.
- Stahlberg, F. (2020). Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.
- Stocker, M. and Smith, M. (2008). Owlgres: A scalable owl reasoner. In *OWLED*, volume 432.
- Sun, K., Zhu, Y., Pan, P., Hou, Z., Wang, D., Li, W., and Song, J. (2019). Geospatial data ontology: the semantic foundation of geospatial data integration and sharing. *Big Earth Data*, 3(3):269–296.
- van Rees, E. (2013). Open geospatial consortium (ogc). *Geoinformatics*, 16(8):28.
- van Veen, T. (2019). Wikidata. *Information technology and libraries*, 38(2):72–81.
- Zhang, Y., Wang, X., Lai, S., He, S., Liu, K., Zhao, J., and Lv, X. (2014). Ontology matching with word embeddings. In Sun, M., Liu, Y., and Zhao, J., editors, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 34–45, Cham. Springer International Publishing.
- Zhao, Y., Zhang, J., Zhou, Y., and Zong, C. (2021). Knowledge graphs enhanced neural machine translation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4039–4045.