Automated Knowledge Graph Approach for Dataset Metadata Harmonisation

Paula Peña-Larena, María del Carmen Rodríguez-Hernández, Luis García-Garcés, Rosa M. Montañés-Salas and Rafael del-Hoyo-Alonso

Technological Institute of Aragon (ITA), María de Luna, 7–8, Zaragoza, Spain

Abstract

The importance of data in Europe's economy, industry, and society is widely recognized, with data-driven innovation playing a crucial role in fostering competitiveness. To ensure more data availability for business use, the concept of data spaces has emerged as a key strategy. In this context, delivering high-quality data-driven services has become imperative. This paper presents an automatic semantic approach for harmonise data for seamless integration by enriching a reference ontology using metadata and textual content. Although Natural Language Processing (NLP) techniques and transformer-based linguistic models are employed for this purpose, our approach goes further by leveraging additional knowledge bases to identify and incorporate new interconnected concepts and relationships into a knowledge graph. As part of the European Data Innovation Hub initiative, our approach's effectiveness in automating data harmonization and enriching the knowledge domain is validated by experimental results derived from extensive dataset analysis and assessment of the generated knowledge graph's quality.

Keywords

Knownledge graph, Natural Language Processing, Transformers, Data spaces

1. Introduction

The importance and impact of data on the European economy, industry and society is nowadays unquestionable. The European Data Market (EDM) monitoring tool [1] reveals the effect of the pervasiveness of data products and services on innovation in the whole economy, so that the benefits from using data and introducing innovation are widespread. Enterprises are increasingly recognizing that business transformation and business benefit is dependent on improved use of data. Early adopter's companies of data technologies, such as big data analytic and artificial intelligence, would have greater benefits than companies in other non-European countries.

Several initiatives towards data-driven innovation as a key driver of growth and jobs to boost European competitiveness in the global market have been promoted by the European Commission. In addition, the European contractual Public Private Partnership on Big Data Value to build a data-driven economy across Europe has already mobilised large amount of private investments [2].

In this context, to bring data-driven innovation closer to the industry (small and medium companies, entrepreneurs and startups), the role of competence centers and digital innovation hubs (DIH), offering support around big data services, products and applications is crucial. According to the DIHs Catalogue of the European Commission[1], there are around 180 DIHs in Europe (out of around 226) specialized in data technologies, such as artificial intelligence and big data, data analytics and data handing.

Moreover, with the aim of breaking down silos, finding synergies and fostering collaboration between DIHs in different technologies and domains, some European initiatives such as *Euhubs4Data* are ongoing[2]. According to the European Commission[3], common European data spaces will ensure that more data becomes available for use in the economy and society, while keeping companies and individuals who generate the data in control. Data space, although still in their early stages, represent a paradigm shift, offering openness and flexibility while being regulated by shared principles and standards[3].

Apart from common principles, aspects of trustworthy and data sovereignty are required to realize data sharing in data spaces. It is crucial to ensure the interoperability data in and between data spaces. In times of digital transformation, data catalogues are gaining prominence to make strategic use of information. A catalogue enables collaboration around data sources, and it might include datasets coming from different data sources, with differ-

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain

ppena@ita.es (P. Peña-Larena); mcrodriguez@ita.es (M. d. C. Rodríguez-Hernández); lggarcia@ita.es (L. García-Garcés); rmontanes@ita.es (R. M. Montañés-Salas); rdelhoyo@ita.es (R. del-Hoyo-Alonso)

0000-0001-5750-6238 (P. Peña-Larena); 0000-0002-0062-9525 (M. d. C. Rodríguez-Hernández); 0000-0003-4636-5868 (R. M. Montañés-Salas); 0000-0003-2755-5500 (R. del-Hoyo-Alonso)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] New DIH Catalogue

[2] AI DIH network, MIDIH, DIHELP and Euhubus4Data project

[3] European data strategy

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

ent levels of information depending on the availability and accessibility of the data and offering a single access point to a common data space. For this reason, it is necessary to devise a strategy or methodology for harmonising dynamically data sources, datasets and metadata.

This work presents a scalable semantic technique for automatically harmonising and enriching datasets from data catalogues and textual content. It utilizes a reference ontology to model datasets, extracting metadata through an API and leveraging textual data, such as the "description" property, through semantic and NLP techniques, and transformer-based linguistic models. The result is a structured knowledge graph that not only maps metadata but also enriches it by extracting additional insights from textual information. Furthermore, the approach facilitates enhancing descriptions of entities by integrating concepts from other ontologies like DBpedia related to the name entities extracted from unstructured text.

This paper is organized as follows. In Section 2, the proposed approach is described. The reference ontology used is described in Section 3. We discuss experimental results in Section 4. Finally, in Section 5, conclusions and future work are presented.

2. *Datasets Metadata Harmonisation* Approach

In the current landscape, the delivery of data-driven services requires a strategy approach to harmonise and standardise data, ensuring seamless integration across diverse sources and domains. These sources cover a broad spectrum of regions and applications domains (open, industrial, personal, research, and more). As a result, datasets often exhibit a multitude of formats, data types and licensing arrangements. In addition, various metadata or schemas (e.g., Dublin Core, VoID, OMV, DCAT, MOD) are applied to describe these datasets, adding another layer of complexity to the data landscape.

In this context, the approach followed by ITA involves adding a semantic layer to enhance data harmonisation and enrich datasets with additional insights extracted from text. A dataset is conceptualized as a *DataResource* using a RDF/OWL ontology based on the widely recognized International Data Spaces (IDS) Information Model[4]. This ontology delegates domain modelling to share vocabularies and data schemes (DCAT, SKOS, FOAF, Owl-Time, among others). This model pursues the goal of establishing an ecosystem that facilitates secure, trusted, and semantically interoperable data exchange. It defines essential concepts for describing actors within a data space, their interactions, the resources exchanged and data usage restrictions.

Considering the significant challenge of enhancing knowledge graphs automatically[4][5][6], without human supervision, semantic approaches have been adopted. A novel harmonised algorithm, detailed in several steps (Figure 1) has been proposed to enrich the IDS Model by leveraging metadata or textual data from datasets to extract additional insights. Typically, knowledge graphs are constructed manually, which demands extensive time and expertise from domain specialists. Hence, the importance and implications of this challenge are profound, particularly given the complexity of automatically building a knowledge graph from unstructured content. Moreover, the quality evaluation of the resulting knowledge graph is crucial, as it directly influences the provision of superior content.

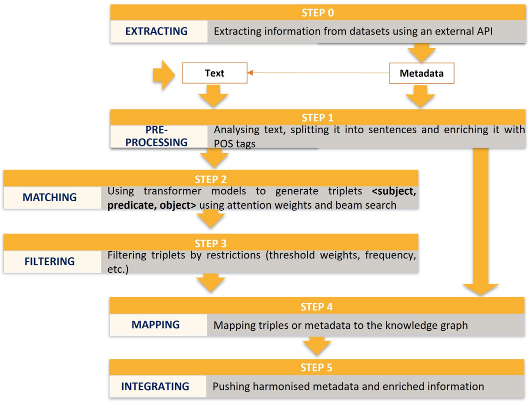

STEP 0
EXTRACTING
Extracting information from datasets using an external API
Text
Metadata
STEP 1
PRE-PROCESSING
Analysing text, splitting it into sentences and enriching it with POS tags
STEP 2
MATCHING
Using transformer models to generate triplets <subject, predicate, object> using attention weights and beam search
STEP 3
FILTERING
Filtering triplets by restrictions (threshold weights, frequency, etc.)
STEP 4
MAPPING
Mapping triples or metadata to the knowledge graph
STEP 5
INTEGRATING
Pushing harmonised metadata and enriched information

Figure 1: Harmonisation algorithm.

Our algorithm discriminates between metadata and textual data obtained through an API from datasets. The dataset information encompasses various attributes such as name, if it is a repository, description, domain, location, license, formats, privacy, publisher, language, etc. To enhance the metadata pre-processing and mapping stages, we introduce NLP techniques aimed at identifying semantic similarities and retrieving knowledge graph properties to improve information filtering.

During the metadata mapping process, a search is carried out for data type properties within the reference ontology that are most akin to a given metadata property, particularly those related to the *DataResource* class and its super-classes. If the input metadata matches any property in the reference ontology, the knowledge graph is enriched with the property description. Otherwise, knowledge-based semantic similarity metrics are applied to compute similarity scores between concepts, words, and entities. If these scores surpass a predefined threshold, the knowledge graph is enriched with the metadata-provided information. Certain datasets include a property that describes their contents, typically denoted

[4] IDS Information Model

as the *description* property. In such cases, the process starts with the extraction of textual data, followed by pre-processing steps such as parsing, segmentation into sentences, and augmentation with metadata attributes like Part-of-Speech (POS) tags.

In the matching step, attention mechanisms are used to infer candidate triplets, leveraging transformer-based language models like BERT[7]. In the first instance, we have used *bert-base-uncased* for English text and *bert-base-spanish* for Spanish text from the Huggingface library. An attention mechanism replicates human selective attention in neural networks, focusing on pertinent information while disregarding irrelevant details. It operates by converting two sentences into a matrix, where words from one sentence form columns and words from another form rows, and then identifies relevant contextual relationships. The attention weights, learned by pre-trained transformer-based models during training, connections between terms in the text (see Figure 2). These weights suggest potential RDF triplets (subject, predicate, object), which are further refined through filtering using constraints such as threshold or frequency (STEP 3).

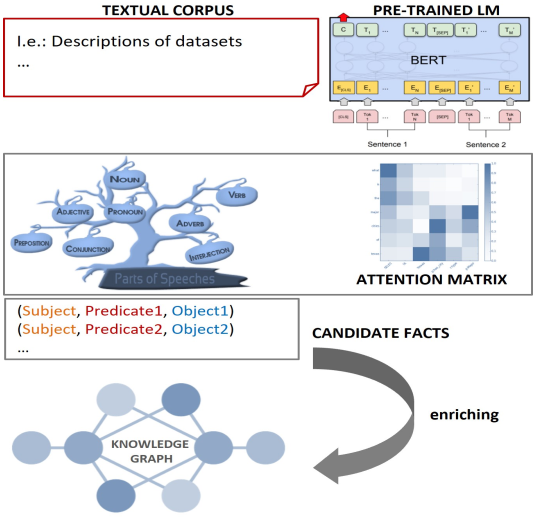
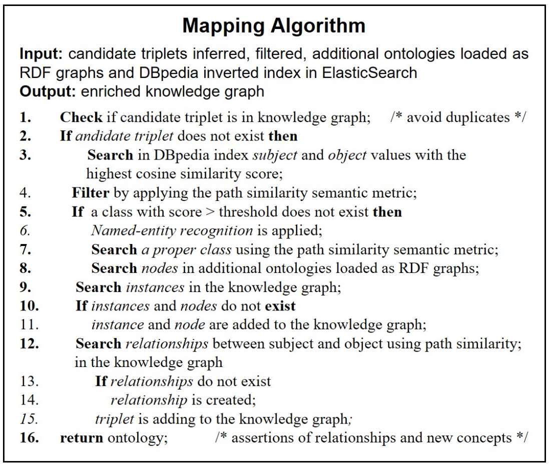


Figure 2: Matching process.

Based on inferred and filtered candidate triplets, and additional ontologies loaded as RDF graphs and to enrich the knowledge graph, we proceed with the mapping process detailed in Figure 3. To enrich the knowledge graph with data from other collective repositories such as DBpedia, ElasticSearch serves as a search engine enabling comprehensive searches across its resources. In addition, inverted indexes have been constructed in ElasticSearch by indexing classes from DBpedia. When querying candidate triplet subject and object values against the inverted index, suitable classes are identified to populate the knowledge graph stored in a Neo4j graph database.



```
Mapping Algorithm
Input: candidate triplets inferred, filtered, additional ontologies loaded as
RDF graphs and DBpedia inverted index in ElasticSearch
Output: enriched knowledge graph
Check if candidate triplet is in knowledge graph;     /* avoid duplicates */
If andidate triplet does not exist then
      Search in DBpedia index subject and object values with the
      highest cosine similarity score;
   Filter by applying the path similarity semantic metric;
   If a class with score > threshold does not exist then
       Named-entity recognition is applied;
       Search a proper class using the path similarity semantic metric;
       Search nodes in additional ontologies loaded as RDF graphs;
    Search instances in the knowledge graph;
    If instances and nodes do not exist
       instance and node are added to the knowledge graph;
    Search relationships between subject and object using path similarity;
    in the knowledge graph
       If relationships do not exist
          relationship is created;
       triplet is adding to the knowledge graph;
return ontology;          /* assertions of relationships and new concepts */
```

Figure 3: Mapping algorithm.

The process applies, among others, the cosine similarity algorithm to DBpedia classes. If values differ, the path similarity semantic metric, which calculates the shortest path between two words in WordNet's representation (*synsets*), is used. When candidate triplet values cannot be mapped to the IDS reference ontology or DBpedia, a multilingual Named-Entity Recognition (NER) model[5] is integrated. This model searches for nodes in additional ontologies related to the NER predictions, generating new relationships and expanding the knowledge graph.

In STEP 5, harmonised metadata and enriched information might be transferred to another system, such as CKAN, for further use. CKAN serves as an open-source Data Management System (DMS), facilitating the publication, sharing, and utilization of data in hubs and portals.

3. Reference ontology

Our research explores how artificial intelligence (AI) technologies, specifically ontologies, can enhance data harmonization and availability for societal and economic use. Semantic technologies, including ontologies and knowledge graphs, are increasingly crucial for establishing data exchange standards and addressing challenges related to heterogeneity and interoperability. Not only introduce a shareable and reusable knowledge representation, but can also add new knowledge about a domain. Ontologies serve as semantic data models, defining types of entities and their properties within a domain. They consist

[5] https://huggingface.co/xlm-roberta-large-finetuned-conll03-english

of *classes* (types of entities), *relationships* (connections between classes), and *attributes* (properties of entities).

By using a reference ontology framework, real-world data can be integrated to form a knowledge graph. For the scenario involving EuHubs4Data project, our base knowledge graph is built on the IDS Information Model ontology, which allows modeling datasets as *DataResource* entities and incorporates various metadata attributes such as dataset representation, dates, language, and source. This ontology establishes fundamental concepts and extends them using external ontologies like DCAT, SKOS, FOAF, Owl-Time, and PROV. It also enables the definition of metadata for resources, such as datasets in our case. This reference model provides a structured framework for modeling datasets, as illustrated in Figure 4.

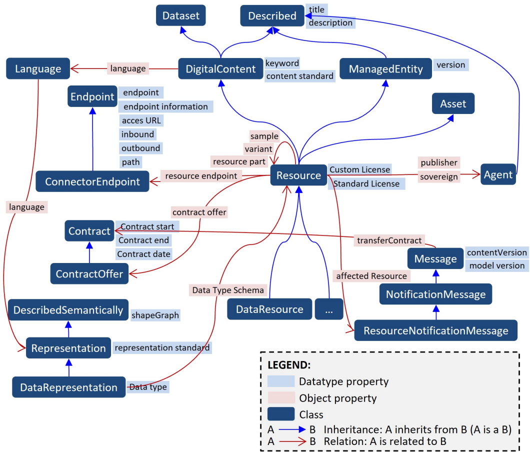


Figure 4: IDS Information Model.

In addition, our knowledge graph remains dynamic, as it continuously incorporates new concepts and relationships identified from various external sources, including DBpedia and other ontologies, aligning with NER prediction labels like organization, person, location, and time. Furthermore, our approach can be extended to integrate with other collective knowledge bases. With the approach presented before, alternative reference ontologies could be employed in different contexts (i.e., Tourism).

4. Experimental Results

To evaluate our automated approach for harmonizing and enriching datasets through knowledge graph generation, 231 datasets from the catalogue[6] provided by Euhubs4Data project members and open data sources have been used. This collection comprises 83 repositories and 148 single datasets exhibiting diversity in domains, coverage, formats, personal data protection levels, available URLs (public or private), and languages.

Comparing basic metrics between the reference ontology and the automatically enriched knowledge graph (such as number of classes, instances, attributes, and class relationships), as presented in Table 1, seems to reveal notable differences. The automated knowledge graph has evolved significantly with a population of elements. This enrichment is attributed to extracting additional insights from textual information within dataset metadata properties and leveraging supplementary knowledge repositories, vocabularies, and ontologies (i.e., DBpedia).

Table 1
Reference and Enriched ontology base metrics

Metrics	Reference	Enriched
Classes	271	1239
Instances	586	2206
Use properties	771	13705
Attributes	1695	3618
Class relations	293	1256
Leaf classes	190	1158
Annotations	435	435
Paths	291	1254

In an experiment showcasing our approach, we examine results from a subset of 18 datasets within the catalogue containing the term *Barcelona*, mapped to the knowledge graph as *DataResource*. These datasets include various topics like *Barcelona-Territory*, *Barcelona-Population*, *Barcelona-Administration*, *Barcelona - Economy and Business* and *Barcelona-City and services* with descriptions ranging from housing and town planning to demographics and education.

Through analysis of the datasets' metadata and textual content, we observe the enrichment of the knowledge graph with new classes and relationships. Notably, these datasets are associated with the new concept of the "Barcelona City Council," which in turn links to concepts like *Housing*, *Urban_planning*, *Demography*, *Education*, and more. This demonstrates the ability of our approach to capture and represent relevant concepts and relationships from datasets, facilitating a deeper understanding of the data landscape.

Similarly, as depicted in Figure 5, the concept of *human resources* is linked to the previously mentioned *Barcelona-Administration* and *Barcelona City Council* classes, along with other new concepts, particularly *Open data*. Upon further exploration, it becomes evident that the graph is enhanced with additional sub-graph derived from insights extracted from other datasets or repositories regarding the *Open data* concept, using the aforementioned

[6] https://euhubs4data.eu/datasets/

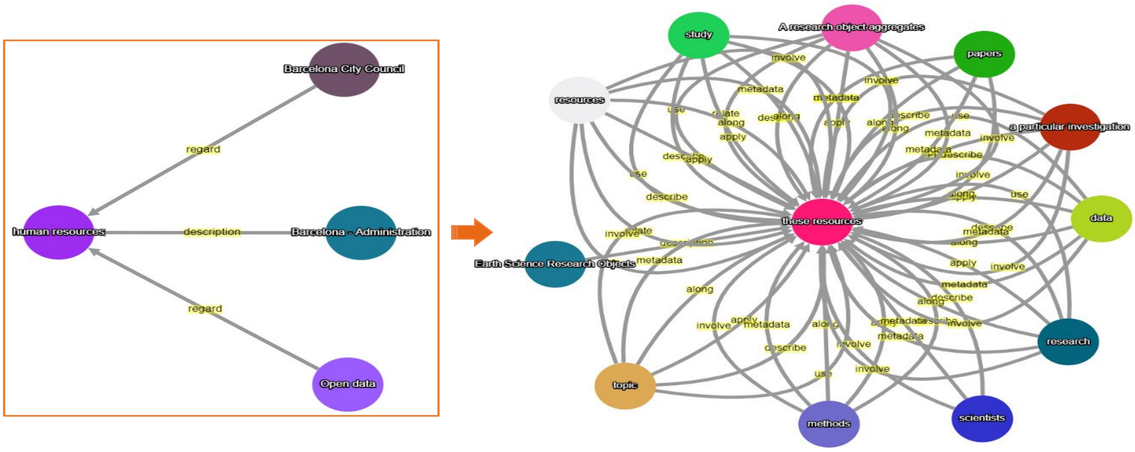


Figure 5: Topology of generated knowledge graph.

knowledge bases and additional ontologies. In addition, by navigating through classes such as *Energy* linked to the *Open Data* class, one can access further graphs concerning resource items. As a result, the enriched knowledge graph, which establishes connections not previously defined, is stored in a Neo4j graph database. This enables data retrieval and visualization through applications such as the web application developed by us[7].

Moreover, to foster higher quality and better outcomes, we have worked on assessing the quality of the generated knowledge graph. In recent years, ontology quality assessment is a key aspect in their development and reuse. The results achieved allow the expert to identify areas that might require further refinement. While various approaches exist[8], in our case, the oQuARE framework[9] is used. The oQuARE framework employs diverse metrics across various dimensions (structural, operability, reliability, transferability, functional adequacy, maintainability, and compatibility) to evaluate ontology quality. Each dimension's assessment relies on its quality subcharacteristics, which are evaluated using associated metrics. For instance, the *structural* dimension is assessed based on cohesion, consistency, formal relations support, formalization, redundancy, and tangledness, involving metrics like ANOnto, PROnto, TMOnto, and LCOMOnto.

To evaluate ontology quality, well-known metrics (LCOMOnto, WMCOnto, DITOnto, NACOnto, NOCOnto, CBOOnto, RFCOnto, NOMOnto, RROnto, AROnto, INROnto, CROnto, ANOnto, TMOnto) are depicted in a radar chart (Figure 6), comparing scores between the reference ontology (IDS Model) and the enriched knowledge graph. Additionally, metrics associated with oQuaRE dimensions are represented in another chart, with calculations similar to those in an available tool[8].

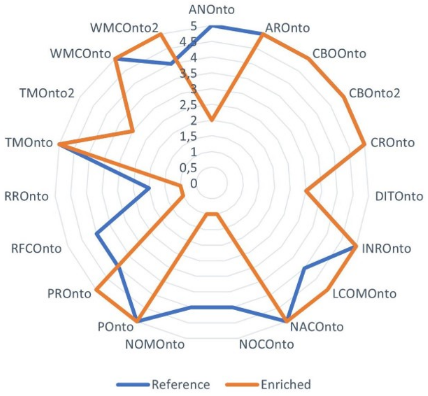


Figure 6: Quality assessment results.

The comparison between the two ontologies reveals notable quality discrepancies. The reference ontology achieved higher scores with respect to the automatically generated ontology. While some metrics related to attribute and class richness, as well as object coupling, yielded similar values, overall, the automatic ontology exhibited areas requiring improvement.

Moreover, using the *OOPS!* tool[10] to identify issues in ontologies uncovered shared pitfalls between the reference and enriched ontology. These included missing domain or range in properties, misusing ontology annotations, undeclared equivalent classes, multiple classes with identical labels, and untyped properties. Addressing

[7] https://euhub4data-graphs.itainnova.es/

[8] http://sele.inf.um.es/ontology-metrics

these issues in the reference ontology before enriching the knowledge graph could enhance overall quality. The quality of results delivered by data-driven services depends on the quality of the ontological model built.

5. Conclusions and Future Work

In recent years, the exponential growth of data production has underline the significance and impact of data in the economy, industry, and society. Social trends towards openness and sharing further emphasise the transformative potential of data in the global economy and society. Providing quality data-driven services has thus become a critical business strategy. Under different application domains, regions, data types, formats and licences, data comes from multiple sources. Given the diverse sources, formats, and licenses of data, there is a need for a strategy to harmonize and standardize data for seamless integration. This paper introduces a semantic approach based on the automatic generation of a knowledge graph from metadata and textual content. Leveraging NLP techniques, transformer-based linguistic models and the support of additional knowledge bases and ontologies, the approach aims to facilitate interoperability between data-driven sources by extracting knowledge from texts and enriching a reference ontology automatically. It follows a series of steps including information extraction, preprocessing of textual content, matching to infer candidate triplets, filtering these triplets, and the mapping and integration of triplets or metadata into the knowledge graph.

The experimental results demonstrate an innovative approach for automatically building knowledge graphs from natural language text within a specific domain. However, the quality assessment conducted, combining different well-know quality metrics, has identified defects or errors in the ontology-learning process that need to be addressed in order to improve the quality of the knowledge graph and thus the quality of results to deliver data-driven services.

As future work, we aim to extract knowledge from various textual sources and automatically generate a high-quality knowledge graph. This graph will serve as a semantic layer for data harmonization and interoperability that might empower recommendation systems with semantic understanding, data quality assurance, cross-domain integration, context-awareness, explainability and adaptive learning capabilities. By harnessing the semantic richness of the knowledge graph, recommendation systems could provide more accurate, relevant, insightful and personalized recommendations that enhance user satisfaction and drive engagement across diverse domains and contexts (Tourism, Health,...). In addition, with the emergence of new large linguistic models like GPT-4, there is an opportunity to explore and test models to assess the approach and adapt it to a specific uses cases.

Acknowledgments

This work has been partially funded by the Department of Big Data and Cognitive Systems at the Technological Institute of Aragon, by IODIDE group of the Government of Aragon, grant number T1720R and by the European Regional Development Fund (ERDF).

References

[1] M. Glennon, M. Kolding, M. Sundbland, C. L. Croce, G. Micheletti, N. Raczko, L. Freitaso, European DATA Market Study 2021–2023. D2.4 Second Report on Facts and Figures, Technical Report, IDC and Lisbon Council, 2023.

[2] BDVA, Big Data Value cPPP, Technical Report, Big Data Value Association, 2019.

[3] C. Mertens, The Data Spaces Radar, Technical Report, International Data Spaces Association, 2023.

[4] N. Mellal, T. Guerram, F. Bouhalassa, An Approach for Automatic Ontology Enrichment from Texts, Informatica (Slovenia) 45 (2021).

[5] S. S. Amani Drissi, Ahmed Khemiri, R. Chbeir, A New Automatic Ontology Construction Method Based on Machine Learning Techniques: Application on Financial Corpus, Procedding 13th International Conference on Management of Digital EcoSystems (2021) 57–61.

[6] Z. Bi, S. Cheng, J. Chen, X. Liang, F. Xiong, N. Zhang, Relphormer: Relational graph transformer for knowledge graph representations, Neurocomputing 566 (2023) 127044.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings Conference of the North American Chapter of the Association for Computational Linguistics 1 (2019) 4171–4186.

[8] G. R. Roldán-Molina, D. Ruano-Ordás, V. Basto-Fernandes, J. R. Méndez, An ontology knowledge inspection methodology for quality assessment and continuous improvement, Data & Knowledge Engineering 133 (2021) 101889.

[9] A. Duque-Ramos, J. T. Fernández-Breis, M. Iniesta, M. Dumontier, et., Evaluation of the OQuaRE framework for ontology quality, Expert Systems with Applications 40 (2013) 2696–2703.

[10] M. Poveda-Villalón, A. Gomez-Perez, M. C. Suárez-Figueroa, OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation, International Journal on Semantic Web and Information Systems (IJSWIS) 10 (2014) 7–34.