SemAI: A Novel Approach for Achieving Enhanced Semantic Interoperability in Public Policies

George Manias, Argyro Mavrogiorgou, Athanasios Kiourtis, Dimosthenis Kyriazis

SemAI: A Novel Approach for Achieving Enhanced Semantic Interoperability in Public Policies

George Manias[1], Argyro Mavrogiorgou[1], Athanasios Kiourtis[1] and Dimosthenis Kyriazis[1]

[1] University of Piraeus, Piraeus, Greece

Abstract. One of the key elements in several application domains, such as policy making, addresses the scope of achieving and dealing with the very different formats, models and languages of data. The amount of data to be processed and analyzed in modern governments, organizations and businesses is staggering, thus Big Data analysis is the mean that helps organizations to harness their data and to identify new opportunities. Big Data are characterized by divergent data coming from various and heterogeneous sources and in different types, formats, and timeframes. Data interoperability addresses the ability of modern systems and mechanisms that create, exchange and consume data to have clear, shared expectations for the context, information and value of these divergent data. To this end, interoperability appears as the mean for accomplishing the interlinking of information, systems, applications and ways of working with the wealth of data. To address this challenge, in this paper a generalized and novel Enhanced Semantic Interoperability approach is proposed, the SemAI. This approach primarily focuses on the phases of the translation, the processing, the annotation, the mapping, as well as the transformation of the collected data, which have major impact on the successful aggregation, analysis, and exploitation of data across the whole policy making lifecycle. The presented prototype and its required subcomponents associated with this approach provide an example of the proposed hybrid and holistic mechanism, verifying its possible extensive application and adoption in various policy making scenarios.

Keywords: Semantic Interoperability, Neural Machine Translation, Semantic Web, NLP, Policy Making.

1 Introduction

Data have long been a critical asset for organizations, businesses, and governments and their analysis is of major importance for every stakeholder in order to be able to handle and extract value and knowledge from these data. The advances in the fields of IoT, cloud-computing, edge-computing and mobile-computing have led to the rapidly increasing volume and complexity of data, thus the concept and term of Big Data has experienced an enormous interest and usability over the last decade. The spectacular growth in the creation, storage, sharing and consumption of data during the last decade indicates the need for modern organizations to fuse advanced analytical tech-

niques with Big Data in order to deal with them and to get significant value from them. Hence, Big Data and analysis of them can enable companies, organizations, and governments to increase operational efficiencies, identify business risks, predict new business opportunities, reduce costs and to provide added value and innovative strategies and mechanisms across their whole policy lifecycle and their policy making techniques. A recent survey has estimated the global market for Big Data at US$70.5 Billion in the year 2020, which is further projected to reach the size of US$86.1 Billion by 2024, growing at a CAGR of 5.25% [1].

Moreover, the term of Big Data defines a two-fold meaning in these data. In one hand it describes a change in the quality and type of data that organizations dispose of, which has potential impacts throughout the entire policy lifecycle. While, in the other hand it describes a massive volume of both structured and unstructured data that is so huge and complicated in order to be processed using traditional database and software techniques [2]. On top of this, unstructured data can be defined as data that do not comfort in predefined data models and traditional structures that can be stored in relational databases. Data generated by conversations, opinions, texts or posts on modern social networks are such type of unstructured data and their main characteristic is that they include information that is not arranged according to a predefined data model or schema. Therefore, these types of data are usually difficult to manage, and as a result analyzing, aggregating, and correlating them in order to extract valuable information and knowledge is a challenging task. Hence, deriving value and knowledge from this type of data based on the analysis of their semantics, meanings and syntactic is of major importance [3]. Data interoperability is the ability to merge data without losing meaning and is realized as a process of identifying the structural, syntactical, and semantic similarity of data and datasets and turn them into interoperable domain-agnostic ones. In practice, data are said to be interoperable when they can be easily reused and processed in different applications, allowing different information systems to work together and share data and knowledge. Hence, Semantic Interoperability is a key enabler for the policy makers in order to enhance the exploitation of Big Data and to better understand data, by extracting and taking into account parameters and information they were not aware of, thus creating efficient and effective policies in terms of good governance. The latter demonstrates the need for the modern stakeholders to implement techniques, mechanisms, and applications that focus their operations on the concept of data interoperability and more specific on Semantic Interoperability [4], for increasing their performance and enhance their entire policy making approach.

At the same time, the volume and continuous growth of the produced data, whether in the form of real-time data or stored data, in various relational and non-relational databases, has led the scientific and business communities to develop sophisticated Big Data applications based on the utilization of techniques and methods from the field of Artificial Intelligence (AI) [5]. Undoubtedly, progress in making the computer systems more human-friendly, requires the inclusion of Natural Language Processing (NLP) techniques, a subfield of the modern area of AI, as integral means of a wider communication interface. NLP leverages linguistics and computer science to make human language intelligible to machines. Therefore, NLP has been used in several

applications and domains to provide enhanced approaches for generating and understanding the natural languages of humans. Speech recognition, topic detection, opinion analysis, and behavior analysis are only a few of such applications and approaches [6]. However, it can be used in concert with text mining to provide policy makers with new filtering systems and more comprehensive analysis tools [7]. In addition, NLP can be utilized as an aide to extract entities and to develop controlled vocabularies, especially enterprise schema that represent classification of a proprietary content set, which can be further utilized for creating proper ontologies [8]. To this end, an approach is being proposed in this paper which utilizes NLP and other AI techniques and tools, such as Neural Networks, and integrates them with Semantic Web technologies, such as controlled vocabularies and ontologies, for achieving Enhanced Semantic Interoperability.

The rest of the paper is organized as follows. Section 2 introduces background knowledge and information from the domain of interoperability in the policy making sector, as well as the related work that has been implemented in the fields of Semantic Interoperability, focused on the utilization of Semantic Web and NLP technologies. Moreover, Section 3 presents the overall methodology and the proposed holistic mechanism for achieving Enhanced Semantic Interoperability based on a hybrid mechanism which couples the utilization of advanced NLP tasks, such as Neural Machine Translation (NMT) and Named Entity Recognition (NER), along with Semantic Web technologies and approaches, such as controlled vocabularies, Resource Description Framework schemas (RDFs)[1], SPARQL[2], and ontologies. Finally, Section 4 concludes the paper and states the future work.

2 Related Work

2.1 Background

Nowadays, policy makers publish an increasing amount of their data on the Web in an effort with double fold meaning. In one hand, to comply with the emerging Open Data movement and in the other hand in order to optimize and improve their policy management and making lifecycle. A key to realizing Open Data and providing advanced open policies is the ability to merge divergent data and datasets coming in the majority of the cases from heterogeneous data sources and in several formats. Moreover, linking new data sources with established data sources is another one key factor for providing and improving policy making procedure. Hence, interoperability is the key "back office" element across the whole policy making lifecycle and Open Data semantics and the mean which can enable policy makers to monitor and improve their performance and trust levels [9].

On top of this, achieving true interoperability entails different representations, purposes, and syntaxes and will enable improved access to assets, records, datasets, and policies. The European Commission, through their program ISA[2] has defined the

[1] https://www.w3.org/2001/sw/wiki/RDFS
[2] https://www.w3.org/TR/rdf-sparql-query/

European Interoperability Framework (EIF) which defines interoperability across four layers: (i) organizational interoperability, (ii) semantic interoperability, (iii) technical interoperability and (iv) legal interoperability [10]. More specifically:

- **Organizational Interoperability**. Ensures public administrations align their processes, responsibilities and expectations to achieve commonly agreed and mutually beneficial goals. Moreover, it aims at addressing the requirements of the end-users by making services and policies available and easily identifiable.
- **Semantic Interoperability**. Ensures that the precise format and meaning of exchanged data and information is preserved and understood throughout by any other application that was not initially developed for this purpose.
- **Technical Interoperability**. Covers the technical issues of linking computer systems and services and includes key aspects such as open interfaces, interconnection services, data integration, data exchange, accessibility and security services.
- **Legal Interoperability.** Ensures that organizations operating under different legal frameworks, policies and strategies are able to work together and in compliance with different laws and regulations.

Hence, it is easily understandable that Semantic Interoperability is the aspect of interoperability which enables systems to combine received information with other information resources and to process it in a meaningful manner, therefore it is a prerequisite for the frontend multilingual delivery of services to the user. To this end, achieving meaningful Semantic Interoperability of data from heterogenous sources is a challenging issue for policy makers, as it is a complex procedure since it covers both semantic and syntactic aspects. The semantic aspect refers to the meaning of data elements and the relationship between them. It includes developing vocabularies, standards, models and schemas to describe data exchanges and ensures that data elements are understood in the same way by all communicating parties and systems. The syntactic aspect refers to describing the exact format of the information to be exchanged in terms of grammar and format. Agreements on reference data, in the form of controlled vocabularies, ontologies, and reusable data models are key prerequisites for achieving Semantic Interoperability. To this end, Semantic Interoperability can offer a way of enriching data with context and meaning and to extract semantic knowledge and good quality information from the data, in order to achieve enhanced understanding of the data, hence better data-driven policy making.

2.2 Semantic Interoperability Approaches

Currently, a wide range of data representation standards and Semantic Interoperability approaches in various domains have emerged as a means of enabling data interoperability and data exchange between different systems. In the recent years many approaches, standards, ontologies and vocabularies have been proposed as means of achieving various tasks of Semantic Interoperability between heterogeneous and independent datasets. One of the first approaches for addressing the issue of Semantic Interoperability was conducted in the scopes of a project in the archaeological domain, which highlights the use of RDFs to achieve Semantic Interoperability across

datasets by extracting and exposing archaeological datasets (and thesauri) in a common RDF framework assisted by a semi-automatic custom mapping tool [11]. Moreover, a recent research focused on implementing a vocabulary (i.e. VoIDext) to formally describe virtual links in order to enable Semantic Interoperability among different datasets [12]. By defining virtual links with VoIDext RDF schema and by providing a set of SPARQL query templates to retrieve them, the research team achieved to facilitate the writing of federated queries and knowledge discovery among federated datasets. In addition, a relevant research exploited semantic similarities between datasets and proposed a method for determining Semantic Interoperability by introducing three metrics to express it between two datasets: the identifier interoperability, the relevance and the number of conflicts [13]. More recently, several approaches were introduced in the sector of IoT where the plethora of divergent datasets coming from heterogeneous sources emerges the issue of achieving high performances of Semantic Interoperability. These approaches mainly focused their operations and procedures on the modeling of a set of ontologies that describe devices and establish Semantic Interoperability between heterogeneous IoT platforms [14, 15]. Moreover, another approach in the domain of IoT introduced a Lightweight Model for Semantic annotation of Big Data using heterogeneous devices in IoT to provide data annotations. To this end, RDF and SPARQL technologies from the domain of Semantic Web were utilized in order to enhance the Semantic Interoperability of the examined datasets and to extract added value and information from them [16]. Another commonly used technology for achieving and enhancing interoperability is the JSON for Linking Data[3] (JSON-LD) format, that has been a W3C recommendation since 2014 to promote interoperability among JSON-based web services. The latter has been utilized in the scopes of a research in the biological sector, which highlights the usage of a JSON-LD system by providing a standard way to add semantic context to the existing JSON data structure, for the purpose of enhancing the interoperability between APIs and data [17]. In addition, a recent research introduced the SEMPROP approach which entails automatically discovering links between datasets through a semantic matcher which leverages Word Embeddings and other components that find links based on syntactic and semantic similarities [18]. Finally, in the healthcare domain, an advanced Semantic Interoperability technique was introduced with emphasis on the utilization of Structural and Ontology Mapping services along with Terminology Linking services in order to transform the clinical information into interoperable and processable data using eHealth standards and terminologies [19].

The above introduced approaches provide the means for common representation of domain specific datasets and the means for achieving Semantic Interoperability across diverse databases and datasets. However, these approaches are insufficient for delivering a holistic mechanism for achieving Enhanced Semantic Interoperability as they lack Semantic Interoperability across datasets from different domains. Effective policy making indicates explicit rules for communication and means for the integration of heterogeneous systems and information resources. Moreover, the above presented approaches do not consider the emerging issue of multilingualism and language-

[3] http://json-ld.org

independence. In modern multicultural and multilingual environments like European Union, policy making is a complex and multilayered procedure of organizations, people, languages, information systems, information structures, rules, processes, and practices. Hence, the needs and trends in modern societies are increasingly demonstrating the need for creating multilingual and interoperable solutions and techniques that will operate in a wider and language-independent context. To this end, NMT should obtain full and effective utilization in modern interoperability systems.

3 Proposed Approach

As presented in the above section many frameworks and approaches have accomplished to provide a level of Semantic Interoperability in the data. In contrary to these introduced approaches, which focused their usability on specific domains, the proposed approach seeks to enhance Semantic Interoperability and to address the issues and lack of Semantic Interoperability across datasets from different domains and the lack of multilingualism and language-independence. Hence, SemAI seeks to address these issues based on technologies from the field of Semantic Web, such as linked data technologies (e.g. JSON-LD and RDF), standards-based ontologies and controlled vocabularies, coupled with the utilization of advanced AI and NLP tasks, such as NMT and Topic Detection. The main goal of this hybrid mechanism is to design and implement a holistic semantic layer that will address data heterogeneity. To this end, this hybrid approach aims to enhance both semantic and syntactic interoperability of data based on the aggregation, correlation, and transformation of incoming data according to the defined schemas and models. The knowledge that will be derived from these processes, shaped in a machine-readable way, can be used latter from other tools for providing Big Data analytics, i.e. Sentiment Analysis etc.

SemAI hybrid mechanism seeks to address the need for interpretable and meaningful data by providing a holistic and multi-layered mechanism from the very beginning of the data lifecycle. Most machine learning algorithms work well either with text or with structured data, but those two types of data are rarely combined to serve as a whole. Semantic web is based on machine understandable languages (RDF, OWL, JSON-LD) and related protocols (SPARQL, Linked Data, etc.), while on the other hand NLP tasks and services focuses on understanding natural languages and raw texts. To this end, the combination of Semantic Web Technologies and NLP tasks will provide a state-of-the-art approach for achieving semantic and syntactic interoperability and will enhance the ability of the SemAI mechanism to combine structured and unstructured data in multiple ways. For example, the utilization of Named Entity Recognition (NER), one of the most widely used tasks of NLP, coupled with the utilization of text mining methods based on semantic knowledge graphs will enhance the Semantic Interoperability and the final linking of divergent data and datasets. To this end, this integrated and hybrid approach will ultimately lead to a powerful and more Enhanced Semantic Interoperability. On top of this, linked data based on W3C Standards can be served as an enterprise-wide data platform and can help to provide training data for machine learning in a more cost-efficient way. The latter further enables

the linking of data even across heterogeneous data sources to provide data objects as training datasets which are composed of information from structured and unstructured data at the same time. To this end, instead of generating datasets per application or use case, high-quality data can be extracted from a knowledge graph. Through this standards-based approach, also internal data and external data can be automatically linked and can be used as rich datasets for any machine learning task. Finally, the utilization of SemAI, in other words the combination of NLP and Semantic Web technologies, will provide the capability of dealing with a mixture of structured and unstructured data that is not possible using traditional, relational tools.

To this end, the hybrid SemAI mechanism incorporates and integrates three different subcomponents: the NMT component, the Semantic & Syntactic Analysis with NLP component, and the Ontology Mapping component, as presented in the below figure (see Fig. 1). One of the preliminary steps of this mechanism is to deal with the very different languages of incoming data. Hence, a NMT component is the first phase and subcomponent that is introduced and will be invoked in this hybrid mechanism, in order to translate data derived in divergent languages into a common language, e.g. English. In next phases, SemAI seeks to identify relevant, publicly available, and widely used classifications and vocabularies, such as the Core Person Vocabulary provided by DCAT Application Profile for Data Portals in Europe (DCAT-AP), that can be reused to codify and populate the content of dimensions, attributes, and measures in the given datasets [20]. Hence, this mechanism aims to adopt standard vocabularies and classifications early on, starting at the design phase of any new data collection, processing or dissemination system. Through the utilization of advanced NLP techniques and tools, such as Text Classification, NER, and Topic Detection, it is feasible to identify and classify same entities, their metadata and relationships from different datasets and sources and finally create cross-domain vocabularies in order to identify every new incoming entity. Likewise, in order to create and enhance semantic interoperability between classifications and vocabularies this component seeks to engage in structural and semantic harmonization efforts, mapping cross-domain terminology used to designate measures and dimensions to commonly used, standard vocabularies and taxonomies with final aim to provide an enhanced Ontology Mapping component. Thus, by implementing a “JSON-LD context” to add semantic annotations to SemAI mechanism’s output, the system will be able to automatically integrate data from different sources by replacing the context-depended keys in the JSON output with URIs pointing to semantic vocabularies, that will be used to represent and link the data [17]. Hence, added information can be expressed by connecting data piece by piece and link by link, allows for any resource (people, policies, articles, search queries etc.) to be identified, disambiguated, and meaningfully interlinked.

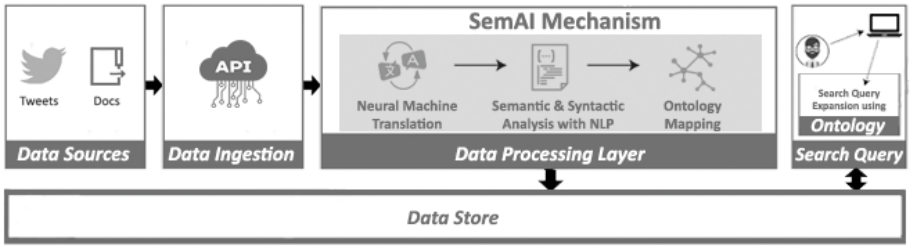

Fig. 1. SemAI Mechanism

3.1 Neural Machine Translation (NMT)

Nowadays, the overarching goal of NLP is to enable communication between humans and computers without resorting to memorable and complex processes. Modern chatbots, automatic translation engines, search engines and more are included in these applications [21]. However, the needs and trends of modern intercultural and multilingual societies are increasingly demonstrating the need for creating multilingual and language-independent solutions and techniques that will operate in a wider context. Thus, the techniques of NMT will obtain full and effective utilization in the scopes of this proposed approach. Recent advances in the field of NMT have proven to be competitive with the encoder-decoder architecture based on the utilization of Recurrent Neural Networks (RNNs), which encode the length of the variable input into unstable dimensions vector and use its encoding to then decode the desired output sequence. Hence, NMT models are often based on the seq2seq architecture [22], which is an encoder-decoder architecture and consists of two Deep Neural Networks: the encoder and the decoder [23]. The input to the encoder is the sentence in the original language, while the input to the decoder is the sentence in the translated language with a start-of-sentence token. The output is the actual target sentence with an end-of-sentence token.

Moreover, new advancements in the field of NMT introduce and propose the utilization of Transformers to solve the Machine Translation problem that relies mostly on the attention mechanism to draw the dependencies between the language models [24]. The attention mechanism enables the decoder to look backward on the whole input sequence and selectively extract the information it needs during processing. Like RNNs, the Transformer is an architecture for transforming one sequence into another using the encoder-decoder mechanism, but it differs from the previous existing seq2seq models because it does not imply any Recurrent Network (GRUs, LSTMs, etc). Yet, unlike the RNNs the Transformer stacks several identical self-attention based layers instead of recurrent units for better parallelization, while it also handles the entire input sequence in at once and does not iterate word by word [25, 26].

Both above introduced approaches and technologies will be utilized and their performance and overall functionality will be evaluated under the scopes of SemAI hybrid mechanism.

3.2 Semantic & Syntactic Analysis with NLP

To exploit what the SemAI offers, translated data first needs to be structured and annotated. To this end, in the second subcomponent, the Semantic & Syntactic Analysis with NLP, translated data will be analyzed, transformed and annotated with appropriate URI metadata and controlled vocabularies will be identified and designed through the utilization of Semantic Web technologies coupled and enhanced by the utilization of NLP techniques, such as Named Entity Recognition (NER), Part-of-Speech Tagging etc, through the utilization of advanced and multilingual NLP tools such as spaCy[4], NLTK[5] and CoreNLP[6]. These coupled functionalities are being utilized into three different layers/steps of the overall subcomponent as shown in the next figure (see Fig. 2). In next phases and steps, semantic and syntactic URI annotated data will be interlinked through the task of Ontology Mapping. The main objectives of this second subcomponent of SemAI mechanism is the identification and recognition of entities, which will be further used for interconnection and interlinking with widely used knowledge bases. Moreover, classifying named entities found in translated data into pre-defined categories, such as persons, places, organizations, dates etc, will make feasible the identification, design and utilization of proper widely used and controlled vocabularies and standards. In addition, the subtask of Named Entities Linking (NEL) will allow to annotate translated data with URIs pointing into corresponding widely used and known knowledge databases, such as Wikidata and DBpedia, while an automatic topic identification and dataset classification task will safeguard that datasets topic of interest are proper identified. Proper topic analysis should be facilitated in order to organize and fully understand the large collections of text data and the correlations among them.

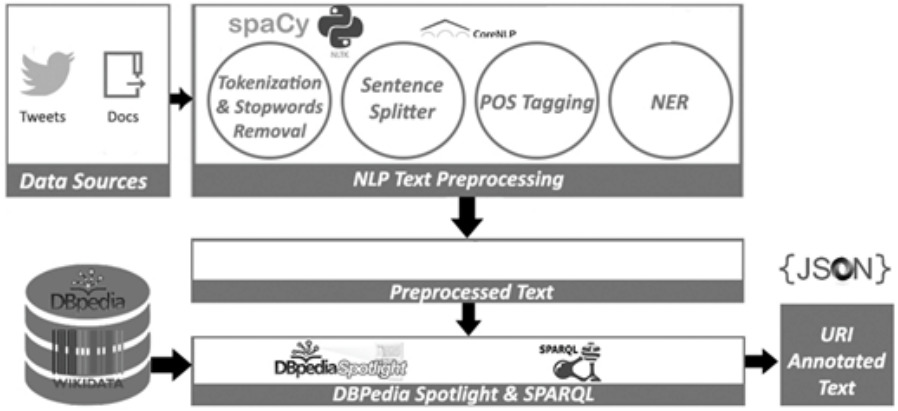

Fig. 2. Metadata and URI annotation of Data

4 https://spacy.io/
5 https://www.nltk.org/
6 https://stanfordnlp.github.io/CoreNLP/

3.3 Ontology Mapping

The overall SemAI mechanism will be further enhanced and completed in the next step by the utilization of Ontology Mapping subcomponent, where an Ontology and Structuring Mapping service will be utilized in order to interlink not only URI annotated data with proper ontologies, but also to interlink and correlate datasets among them. Successful annotation, transformation and mapping of data and corresponding ontologies in terms of semantic and syntactic interoperability of data is one of the key elements of SemAI mechanism. To this end, one of the main objectives of the Ontology Mapping subcomponent is to save correlated, annotated and interoperable data in JSON-LD format and as linked ontologies. Hence, it will be feasible the storage of semantic facts and the support of the corresponding data schema models. Moreover, this subcomponent seeks to map concepts, classes, and semantics defined in different ontologies and datasets and to achieve transformation compatibility through extracted metadata. In addition, a data modelling subtask by standard metadata schemas will be defined in order to specify the metadata elements that should accompany a dataset within a domain. To this end, semantic models for physical entities / devices (i.e. sensors related to different policy sectors) and online platforms (e.g. social media) will be identified. These models will be based on a set of transversal and domain-specific ontologies and could provide a foundation for high-level Semantic Interoperability and rich semantic annotations across policy sectors, online systems and platforms. As shown in the below figure (see Fig. 3) there are several levels of structuring before reaching proper ontologies. At the beginning, the annotation and creation of metadata representations through the utilization of JSON-LD technology is a key point. Afterwards, vocabularies and taxonomies expressed by RDFs are created and in the final step they are correlated and interlinked into ontologies with high semantic expressivity through the utilization of OWL technology.

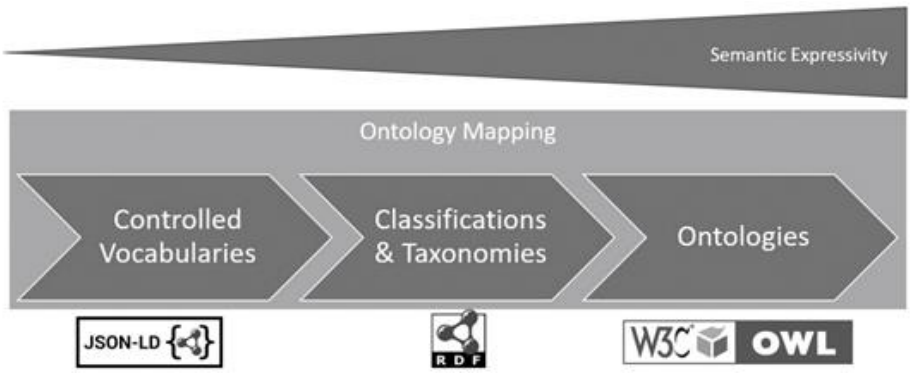

Fig. 3. Ontology Mapping subcomponent in SemAI

On top of this, ontologies are central to the SemAI as they allow applications to agree on the terms that they use when communicating and they enable the correlation

of divergent data and datasets from various sources. To this end, the utilization of ontologies under the scope of SemAI facilitates communication by providing precise notions that can be used to compose messages (queries, statements) about the policy making domain. In stakeholders and user level, the ontology helps to understand messages by providing the correct interpretation context. Thus, ontologies, if shared among stakeholders, may improve system interoperability across Information Systems (ISs) in different organizations and domains. The overall approach that will be followed brings together techniques in modeling, computation linguistics, information retrieval and agent communication in order to provide a semi-automatic mapping method and a prototype mapping system that support the process of Ontology Mapping for the purpose of improving and enhancing Semantic Interoperability during the whole data and policy lifecycle.

The novelty of the proposed Ontology Mapping subcomponent is not solely the use of formal application ontologies as an initial mechanism to achieve meaningful Semantic Interoperability, but moreover the utilization of divergent domain ontologies to support the formal application ontologies mapping process, integrated into an architectural framework.

4 Conclusion

In this paper, a novel approach for achieving Enhanced Semantic Interoperability in the domain of policy making was introduced, the SemAI. SemAI introduces a multi-layer and hybrid mechanism for Semantic Interoperability across diverse policy related datasets, which will facilitate Semantic Interoperability across related datasets both within a single domain and across different policy making domains. This requirement relates to local-regional public administrations and business domain, but it also goes beyond the national borders as it also seeks to invoke a language-independent hybrid mechanism. Moreover, IT systems and applications interoperability, sharing and reuse, and interlinking of information and policies, within and between domains are essential factors for the delivery of high quality, innovative, and seamless policies. Under this framework, SemAI and its required steps and subcomponents were presented to ease its adoption at data-driven policy making domain. Achieving high levels of Semantic Interoperability in the data can help organizations and businesses to turn their data into valuable information, add extra value and knowledge to them and finally achieve enhanced policy making through the combination and correlation of several data, datasets, and policies. The proposed approach in this paper is established as a service which can be adopted and integrated into different policy making scenarios and comprises an effort to deal with the Semantic Interoperability. The latter will be implemented and further evaluated in the context of a holistic environment for data-driven policy making as realized by the PolicyCLOUD project [27], where data from four different languages (Bulgarian, Italian, Spanish and English) and from various policymakers and domains of interests, such as public authorities, businesses, and organizations participate with the aim of turning raw data into valuable and actionable knowledge towards efficient policy making.

Acknowledgment

The research leading to the results presented in this paper has received funding from the European Union's funded Project PolicyCLOUD under grant agreement no 870675.

References

1. Big Data - Global Market Trajectory & Analytics, https://www.researchandmarkets.com/reports/2228010/big_data_global_market_trajectory_and_analytics, last accessed 2021/03/01.
2. Chavan, V., Phursule, R. N.: Survey paper on big data. Int. J. Comput. Sci. Inf. Technol, vol. 5, no. 6, pp. 7932-7939 (2014).
3. Mosley, M., Brackett, M. H., Earley, S., Henderson, D.: DAMA guide to the data management body of knowledge. Technics Publications (2010).
4. Motta, G., Puccinelli, R., Reggiani, L., Saccone, M.: Extracting Value from Grey Literature: processes and technologies for aggregating and analyzing the hidden «big data» treasure of organizations. Grey Journal (TGJ), vol. 12, no. 1 (2016).
5. Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., Vasilakos, A. V.: Big data: From beginning to future. International Journal of Information Management, vol. 36, no. 6, pp. 1231-1247 (2016).
6. Bahja, M.: Natural language processing applications in business. In E-Business. IntechOpen (2020).
7. Kao, A., Poteet, S. R.: Natural language processing and text mining. Springer Science & Business Media (2007).
8. Zheng, S., Lu, J. J., Ghasemzadeh, N., Hayek, S. S., Quyyumi, A. A., Wang, F.: Effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies. JMIR medical informatics, vol. 5, no. 2 (2017).
9. Solanas, A., Patsakis, C., Conti, M., Vlachos, I. S., Ramos, V., Falcone, F., et al: Smart health: A context-aware health paradigm within smart cities. IEEE Communications Magazine, vol. 52, no. 8, pp. 74-81 (2014).
10. New European Interoperability Framework, https://ec.europa.eu/isa2/sites/isa/files/eif_brochure_final.pdf, last accessed 2021/03/11.
11. Binding C., May K., Tudhope D.: Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction Via the CIDOC CRM. In: Christensen-Dalsgaard B., Castelli D., Ammitzbøll Jurik B., Lippincott J. (eds) Research and Advanced Technology for Digital Libraries. ECDL 2008. Lecture Notes in Computer Science, vol 5173. Springer, Berlin, Heidelberg (2008).
12. de Farias, T. M., Stockinger, K., Dessimoz, C.: VoIDext: Vocabulary and patterns for enhancing interoperable datasets with virtual links. In: OTM Confederated International Conferences "On the Move to Meaningful Internet Systems", pp. 607-625. Springer, Cham (2019).
13. Colpaert, P., Van Compernolle, M., De Vocht, L., Dimou, A., Vander Sande, M., Verborgh, R., et al: Quantifying the interoperability of open government datasets. Computer, vol. 47, no. 10, pp. 50-56 (2014).

14. Ganzha, M., Paprzycki, M., Pawłowski, W., Szmeja, P., Wasielewska, K.: Semantic interoperability in the Internet of Things: An overview from the INTER-IoT perspective. Journal of Network and Computer Applications, vol. 81, pp. 111-124 (2017).
15. Bajaj, G., Agarwal, R., Singh, P., Georgantas, N., Issarny, V. A study of existing Ontologies in the IoT-domain. arXiv preprint arXiv:1707.00112 (2017).
16. Ullah, F., Habib, M. A., Farhan, M., Khalid, S., Durrani, M. Y., Jabbar, S.: Semantic interoperability for big-data in heterogeneous IoT infrastructure for healthcare. Sustainable cities and society, vol. 34, pp. 90-96 (2017).
17. Xin, J., Afrasiabi, C., Lelong, S., Adesara, J., Tsueng, G., Su, A. I., Wu, C.: Cross-linking BioThings APIs through JSON-LD to facilitate knowledge exploration. BMC bioinformatics, vol. 19, no. 1, pp. 1-7 (2018).
18. Fernandez, R. C., Mansour, E., Qahtan, A. A., Elmagarmid, A., Ilyas, I., Madden, S., et al: Seeping semantics: Linking datasets using word embeddings for data discovery. In: 2018 IEEE 34th International Conference on Data Engineering (ICDE), pp. 989-1000. IEEE (2018).
19. Kiourtis, A., Mavrogiorgou, A., Menychtas, A., Maglogiannis, I., Kyriazis, D.: Structurally mapping healthcare data to HL7 FHIR through ontology alignment. Journal of medical systems, vol. 43, no. 3, pp. 62 (2019).
20. DCAT Application profile for data portals in Europe (DCAT-AP), https://op.europa.eu/en/web/eu-vocabularies/dcat-ap, last accessed 2021/03/11.
21. Bulut, Y. E.: AI for data science: artificial intelligence frameworks and functionality for deep learning, optimization, and beyond. Technics Publications (2018).
22. Tiwari, G., Sharma, A., Sahotra, A., Kapoor, R.: English-Hindi Neural Machine Translation-LSTM Seq2Seq and ConvS2S. In: 2020 International Conference on Communication and Signal Processing (ICCSP), pp. 871-875. IEEE (2020).
23. Yang, M., Liu, S., Chen, K., Zhang, H., Zhao, E., Zhao, T.: A hierarchical clustering approach to fuzzy semantic representation of rare words in neural machine translation. IEEE Transactions on Fuzzy Systems, vol. 28, no. 5, pp. 992-1002 (2020).
24. Bahar, P., Makarov, N., Ney, H.: Investigation of Transformer-based Latent Attention Models for Neural Machine Translation. In: Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020), pp. 7-20 (2020).
25. Pramodya, A., Pushpananda, R., Weerasinghe, R.: A Comparison of Transformer, Recurrent Neural Networks and SMT in Tamil to Sinhala MT. In: 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 155-160. IEEE (2020).
26. Lakew, S. M., Cettolo, M., Federico, M.: A comparison of transformer and recurrent neural networks on multilingual neural machine translation. arXiv preprint arXiv:1806.06957 (2018).
27. Kyriazis, D., Biran, O., Bouras, T., Brisch, K., Duzha, A., del Hoyo, R., et al: PolicyCLOUD: Analytics as a Service Facilitating Efficient Data-Driven Public Policy Management. In: IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 141-150. Springer, Cham (2020).