Multilingual Knowledge Systems as Linguistic Linked Open Data for European Language Grid

Alena Vasilevich, Michael Wetzel

Coreon GmbH, Rungestrasse 20, 10179 Berlin, Germany

Abstract

Creation and re-usability of language resources in accordance with Linked Data principles is a valuable asset in the modern data world. In this paper we describe the contributions made to extend the LLOD stack with a new resource, Coreon MKS, bringing together concept-oriented, language-agnostic terminology management and graph-based knowledge organization. We dwell on our approach to mirroring of Coreon's original data structure to RDF and supplying it with a real-time SPARQL endpoint. We integrate MKS into the existing ELG infrastructure, using it as a platform for making the published MKS discoverable and retrievable via a industry-standard interface. While we apply this approach to LLOD-ify Coreon MKS, it can provide a relevant input for standardisation bodies and interoperability communities, acting as a blueprint for similar integration activities.

Keywords

Terminology Management, Linguistic Linked Open Data, RDF, SPARQL, Semantic Web, Multilingual Knowledge, European Language Grid

1. Introduction

In the world depending on knowledge sharing, data-driven businesses and research communities are concerned with creation, sharing, and use of language resources in accordance with Linked Data principles, which ensure better data discoverability, standardised structure, and significant cost savings for parties involved in the creation of such structured data.

Robust, coherent, and multilingual information standards are needed to enable information exchange among public organizations, similar to standards that have been fostering technical interoperability for decades [1]. European Language Grid (ELG)[1] is a shared platform for the European Language Technology (LT) landscape that tackles the fragmentation of the LT market space, bridging together services that can benefit European society, industry, and politics [2]. This initiative was launched to encourage the use of automated, easy-to-integrate mechanisms that allow such cross-platform exchange of services, models, datasets, and metadata records, providing structured and semantically aligned information about the contents of the respective platforms [3, 4].

SALLD-1: Workshop on Sentiment Analysis & Linguistic Linked Data in conjunction with LDK 2021 – 3rd Conference on Language, Data and Knowledge, September 01-03, 2021, Zaragoza, Spain

alena@coreon.com (A. Vasilevich); michael@coreon.com (M. Wetzel)

https://www.coreon.com/ (M. Wetzel)

0000-0002-9769-1885 (A. Vasilevich)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://live.european-language-grid.eu/

W3C SPARQL[2] is a protocol and an established query language widely used for the information retrieval in the Semantic Web resources. While existing SPARQL tools enable users to query knowledge graphs, they are rarely used for termbases and other terminology resources that represent core sources of data for translation and localization [5].

In this light, we extend the Linguistic Linked Open Data (LLOD) stack with a new resource, Multilingual Knowledge System (MKS)[3]. MKS caters for discovery, access, retrieval, and re-usability of terminologies and other interoperability assets organised in knowledge graphs (KG) in a taxonomic fashion. Being a semantic knowledge repository, its main forte is the ability to exchange information among acting systems, ensuring that its precise *meaning* is understood and preserved among all parties, in any language. Injecting structure into the language data and expanding the resulting KG with multilingual terminologies, Coreon uses ELG as a platform for making the published resources discoverable and retrievable via the SPARQL interface. This step makes Coreon integration into other systems tool-independent: instead of using the proprietary API, it relies on the well-pronounced LLOD standards.

The goal of our contribution is to deliver MKS resources to the Semantic Web Community, enabling it to query concept-oriented multilingual structured data with a well-established industry-standard syntax, and to promote the development of data multilingualism within the Semantic Web. In the longer perspective, MKS as a LLOD resource can provide a relevant input for standardisation bodies and interoperability communities: acting as a blueprint for similar integration activities, it can be viewed as a starting point for an international standard. We share our experience with ISO / TC37 SC3[4] working groups as a draft for a technical recommendation on how to represent TermBase eXchange (TBX)[5] dialects as RDF.

2. Making Coreon Data Structure LLOD-compatible

Resource Description Framework (RDF)[6] and Web Ontology Language (OWL)[7] are standardised formats for representing Semantic Web information. They support data integration and offer a plethora of tools and methods for data access. SPARQL, being among the popular ones, acts on RDF/OWL resources allowing users to retrieve structured answers based on submitted queries. To express queries, it utilises triple patterns that are to be matched by RDF/OWL triples and filter conditions, imposing ranges for literals [6]. However, despite the emerging interest in publishing terminological resources as linked data, LLOD stack has not been heavily utilised for this purpose so far [7].

Seeing this as a lost opportunity for the reach of terminology resources, we implemented a solution for Coreon MKS, making termbases discoverable and accessible for systems powered by the best Semantic Web and LLOD practices [8]. Normally data owners would investigate, select, and deploy an alternative technology like an RDF triple store for their terminology tool, often developing and/or setting up a tedious data-mirroring process. We describe a way to overcome

[2]https://www.w3.org/TR/rdf-sparql-query/
[3]https://www.coreon.com/
[4]https://www.iso.org/committee/48136.html
[5]https://www.tbxinfo.net/
[6]https://www.w3.org/RDF
[7]https://www.w3.org/OWL

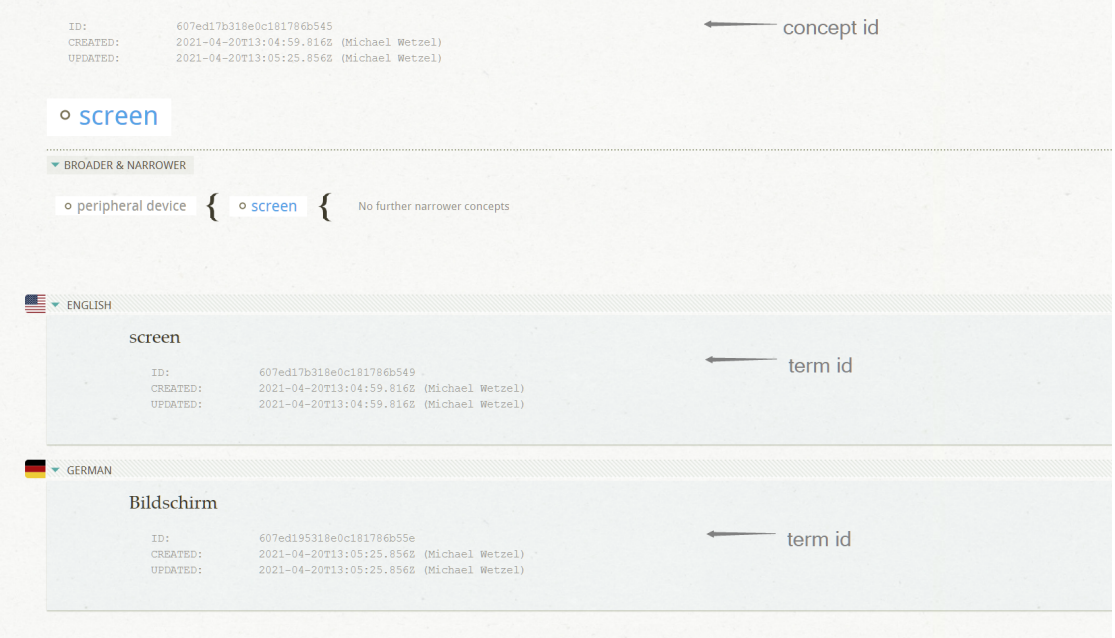

Figure 1: Sample Concept in Coreon GUI.

the limits of RDF/knowledge graph editors, which tend to be good at relation modeling but have weaknesses when it comes to capturing linguistic information.

At the core of the MKS lies a language independent knowledge graph. Unlike in other popular solutions within the terminological and taxonomic management, the linking is performed **not** at the *term* but at the *concept* level. Thus, abstracting from terms, we can model structured knowledge for phenomena that reflect the non-deterministic nature of the human language, such as word sense ambiguity, synonymy, and multilingualism. Linking *per concept* also ensures smooth maintenance of relations without additional data clutter: relation edges are independent from labels, terms and their variants, and other metadata.

We analyzed and implemented a data mirroring process from Coreon data model to an RDF graph, establishing the RDF vocabulary for classes, relations, additional term-descriptive information, and administrative meta-data, binding the elements into RDF triples that feature atomic, isolated nature. At this stage it is critical to identify information objects, mapping of predicates and literals. Figure 1 displays a snapshot of Coreon UI, featuring a sample concept with ID *607ed17b318e0c181786b545* that has two terms, English *screen* and German *Bildschirm*; whereas the code snippet in Listing 1 depicts relevant lines within the original JSON data structure that represents this sample concept, with *concept* ID and individual *term* IDs and their values highlighted. To transform this data structure into an RDF graph, the concept and its two terms are bound together in statements, i.e. RDF triples. Each triple comprises a subject, a predicate, and an object; in our case, the concept will act as the subject, the terms become objects, and the required predicate is named `hasTerm`.

Listing 1: Snippet of Coreon Data Structure.

```
{
    "created_at": "2021-04-20T13:04:59.816Z",
    "updated_at": "2021-04-20T13:05:25.856Z",
    "terms": [
        {
            "lang": "en",
            "value": "screen",
            "updated_at": "2021-04-20T13:04:59.816Z",
            "id": "607ed17b318e0c181786b549",
            "concept_id": "607ed17b318e0c181786b545",
            "properties": []
        },
        {
            "lang": "de",
            "value": "Bildschirm",
            "updated_at": "2021-04-20T13:05:25.856Z",
            "id": "607ed195318e0c181786b55e",
            "concept_id": "607ed17b318e0c181786b545",
            "properties": []
        }
    ],
    "id": "607ed17b318e0c181786b545"
}
```

Listing 2: Sample RDF Triples

```
1 coreon:607ed17b318e0c181786b545 coreon:hasTerm
2                                 coreon:607ed17b318e0c181786b549
3
4 coreon:607ed17b318e0c181786b549 coreon:value "screen"@en .
```

The resulting triple pictured on lines 1-2, Listing 2 states that the term with the given ID is a member of the concept. The second triple (line 4) shows that the value of the instance with ID `607ed17b318e0c181786b549` has a `literal` value in English, the string *screen*. Such triple statements connected via predicates make up an RDF graph. Figure 2 displays the resulting sample RDF graph, with concepts and terms as classes rendered in green and blue and predicates as graph edges. The complete sample set of triples serialised in RDF/Turtle is provided in Listing 3, with highlighted lines 18-19 indicating that the resource with ID `606336dab4dbcf018ed99308` belongs to the OWL class `coreon:Concept` and contains a term with ID `606336dab4dbcf018ed99307`. In RDF and LOD, data is stored in an atomic manner, with predicates and uniform resource identifiers (URIs) linking elements together. In our case, all instances represented as classes receive unique identifiers: the namespace `coreon:` together with unique IDs, unambiguously identify any given element, regardless of whether it is a concept, term, property, or a concept relation.

Table 1 lists our RDF vocabulary, derived from the original Coreon data structure. During the

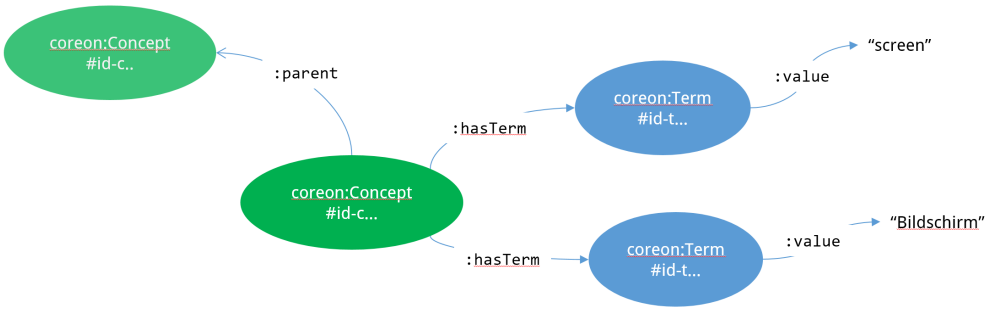

Figure 2: Resulting RDF graph.

Table 1
Derived Coreon RDF Vocabulary

	OWL Type	Coreon RDF Vocabulary
Classes	owl:Class	coreon:Admin, coreon:Edge, coreon:Concept, coreon:Flagset, coreon:Property, coreon:Term
Predicates	owl:ObjectProperty	coreon:hasAdmin, coreon:hasFlagset, coreon:hasProperty, coreon:hasTerm
Values	owl:AnnotationProperty	coreon:edgeSource, coreon:edgeTarget, coreon:id, coreon:name, coreon:type, coreon:value

Coreon-to-RDF conversion stage, there were obvious candidates for classes, like `Concept` and `Term`; yet mirroring descriptive information like `Definition` or `Term Status` and mapping of concept relations, e.g. taxonomic "broader"/"narrower" or custom associative ones, turned out to be challenging. For the predicates we had to specify what kind of information can be used, defining `owl:range` and `owl:domain`: e.g., predicate `hasTerm` can only accept resources of type `coreon:Concept` as a subject (`owl:domain`). Listing 4 provides a full specification of this predicate.

3. Real-Time Data Access via a SPARQL Endpoint

With the vocabulary ready, Coreon's export engine got equipped with the RDF publication mechanism, including RDF export in all relevant syntax flavours (Turtle, N3, JSON-LD, etc.). Coreon cloud service is supplied with a real-time accessible SPARQL endpoint via Apache Jena Fuseki[8]. It conforms to all of the published standards and tracks the revisions and updates in the under-development areas of the standard. Running as a secondary index in parallel to

[8]https://jena.apache.org/

Listing 3: Triples serialised in RDF / Turtle

```
@prefix coreon: <http://www.coreon.com/coreon-rdf#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://www.coreon.com/coreon-instance> a owl:Ontology;
  owl:imports <http://www.coreon.com/coreon-rdf>;
  owl:versionInfo "Created through Coreon export" .

coreon:607ed17b318e0c181786b547 a coreon:Edge;
  coreon:edgeSource coreon:606336dab4dbcf018ed99308;
  coreon:edgeTarget coreon:607ed17b318e0c181786b545;
  coreon:type "SUPERCONCEPT_OF" .

coreon:606336dab4dbcf018ed99307 a coreon:Term;
  coreon:value "peripheral device"@en .

coreon:606336dab4dbcf018ed99308 a coreon:Concept;
  coreon:hasTerm coreon:606336dab4dbcf018ed99307 .

coreon:607ed17b318e0c181786b545 a coreon:Concept;
  coreon:hasTerm coreon:607ed195318e0c181786b55e,
    coreon:607ed17b318e0c181786b549 .

coreon:607ed17b318e0c181786b549 a coreon:Term;
  coreon:value "screen"@en .

coreon:607ed195318e0c181786b55e a coreon:Term;
  coreon:value "Bildschirm"@de .
```

Listing 4: Specification of a Predicate

```
coreon:hasTerm
  rdf:type owl:ObjectProperty ;
  rdfs:comment "makes a term member of a concept" ;
  rdfs:domain coreon:Concept ;
  rdfs:label "has term" ;
  rdfs:range coreon:Term .
```

the repository's data store, Fuseki therefore catches any changes made by data maintainers, updating the state of the repositories in real time.

Listing 5 demonstrates a simple sample SPARQL query over a MKS that deals with wine grape varieties: here, we want to return all of the terms, including the values of the `"Usage"` flag in case the terms have them. Table 2 displays a subset of the resulting linked data structures returned by this query: there is a term's URI, its value, and usage recommendation if available.

Another typical use case is shown in the Listing 6 and its result Table 3: here we want to return counts of all instances of the unique classes, i.e. answering how many Definitions,

Listing 5: Sample SPARQL Query over MKS

```
SELECT ?t ?termvalue ?usagevalue
    WHERE {
        ?t rdf:type coreon:Term .
        ?t coreon:value ?termvalue .
        OPTIONAL {
            ?t coreon:hasProperty ?p .
            ?p coreon:key "Usage" .
            ?p coreon:value ?usagevalue .
        }
    }
```

Table 2
Sample SPARQL Query 5 Results: Returned Grape Varieties

[t]	termvalue	usagevalue
http://www.coreon.com/coreon-rdf#5f9ee3609323c01c4728b8aa	Riesling	
http://www.coreon.com/coreon-rdf#5f9ee3609323c01c4728b8bb	Cabernet Sauvignon	Preferred
http://www.coreon.com/coreon-rdf#5f9ee3609323c01c4728b8be	CS	Alternative
http://www.coreon.com/coreon-rdf#5f9ee3609323c01c4728b8c2	Merlot	

Listing 6: Querying Total Instances within Coreon Classes

```
SELECT ?k (COUNT(?k) AS ?count)
    {
        ?uri coreon:key ?k.
    }
GROUP BY ?k
ORDER BY DESC(?count)
```

Comments, or most used Properties the repository contains.

4. Conclusion

We have shared a pipeline on making MKS resources LLOD-compatible, mapping the Coreon-native data structure to RDF, conceiving the Coreon-RDF Vocabulary, and publishing MKS resources via the ELG hub. Besides making the SPARQL endpoint available to Coreon users through the state-of-the art ELG infrastructure, we implemented a productised piece of software, which provides multilingual TermBase eXchange-like terminology resources in the RDF data model and the Semantic Web context. To demonstrate the applicability and the result of our data structure mirroring activities and hosting of the resulting knowledge graphs, a selected set

Table 3
Sample SPARQL Query 6: Results

[k]	count
concept status	13806
usage status	10532
part of speech	10408
term type	10353
definition	5996

of demo MKS repositories will be accessible with the SPARQL endpoint through the ELG hub by summer 2021.

Beyond establishing structural interoperability, the implemented interface acts as a bridge between the MKS and other Semantic Web systems, enabling querying of elaborate multilingual terminologies stored in knowledge graphs. Our approach can act as a blueprint for similar conversion and integration activities, viewed as a starting point for an international standard. Deployed on the ELG platform, Coreon's SPARQL interface enables the Semantic Web community to query rich heterogeneous MKS data with a familiar, industry standard syntax, promoting data accessibility and contributing to the development of multilingual resources within the Semantic Web.

Acknowledgments

The research presented in this paper was partially supported by the European Language Grid project through its open call for pilot projects. The European Language Grid project has received funding from the European Union's Horizon 2020 Research and Innovation programme under the Grant Agreement no. 825627 (ELG).

References

[1] L. Guijarro, Semantic interoperability in egovernment initiatives, Computer Standards & Interfaces 31 (2009) 174–180. URL: https://doi.org/10.1016/j.csi.2007.11.011. doi:10.1016/j.csi.2007.11.011.

[2] G. Rehm, K. Marheinecke, S. Hegele, S. Piperidis, K. Bontcheva, J. Hajič, K. Choukri, A. Vasiļjevs, G. Backfried, C. Prinz, J. M. Gómez-Pérez, L. Meertens, P. Lukowicz, J. van Genabith, A. Lösch, P. Slusallek, M. Irgens, P. Gatellier, J. Köhler, L. Le Bars, D. Anastasiou, A. Auksoriūtė, N. Bel, A. Branco, G. Budin, W. Daelemans, K. De Smedt, R. Garabík, M. Gavriilidou, D. Gromann, S. Koeva, S. Krek, C. Krstev, K. Lindén, B. Magnini, J. Odijk, M. Ogrodniczuk, E. Rögnvaldsson, M. Rosner, B. Pedersen, I. Skadiņa, M. Tadić, D. Tufiș, T. Váradi, K. Vider, A. Way, F. Yvon, The European language technology landscape in 2020: Language-centric and human-centric AI for cross-cultural communication in multilingual Europe, in: Proceedings of the 12th Language Resources and Evaluation Conference,

European Language Resources Association, Marseille, France, 2020, pp. 3322–3332. URL: https://www.aclweb.org/anthology/2020.lrec-1.407.

[3] G. Rehm, M. Berger, E. Elsholz, S. Hegele, F. Kintzel, K. Marheinecke, S. Piperidis, M. Deligiannis, D. Galanis, K. Gkirtzou, P. Labropoulou, K. Bontcheva, D. Jones, I. Roberts, J. Hajič, J. Hamrlová, L. Kačena, K. Choukri, V. Arranz, A. Vasiljevs, O. Anvari, A. Lagzdiņš, J. Meļņika, G. Backfried, E. Dikici, M. Janosik, K. Prinz, C. Prinz, S. Stampler, D. Thomas-Aniola, J. M. Gómez-Pérez, A. Garcia Silva, C. Berrío, U. Germann, S. Renals, O. Klejch, European language grid: An overview, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 3366–3380. URL: https://www.aclweb.org/anthology/2020.lrec-1.413.

[4] G. Rehm, S. Piperidis, K. Bontcheva, J. Hajic, V. Arranz, A. Vasiljevs, G. Backfried, J. M. Gómez-Pérez, U. Germann, R. Calizzano, N. Feldhus, S. Hegele, F. Kintzel, K. Marheinecke, J. M. Schneider, D. Galanis, P. Labropoulou, M. Deligiannis, K. Gkirtzou, A. Kolovou, D. Gkoumas, L. Voukoutis, I. Roberts, J. Hamrlová, D. Varis, L. Kacena, K. Choukri, V. Mapelli, M. Rigault, J. Melnika, M. Janosik, K. Prinz, A. García-Silva, C. Berrio, O. Klejch, S. Renals, European language grid: A joint platform for the european language technology community, in: D. Gkatzia, D. Seddah (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, EACL 2021, Online, April 19-23, 2021, Association for Computational Linguistics, 2021, pp. 221–230. URL: https://www.aclweb.org/anthology/2021.eacl-demos.26/.

[5] R. Stanković, I. Obradović, M. Utvić, Developing termbases for expert terminology under the tbx standard, Editors Gordana Pavlović Lažetić Duško Vitas Cvetana Krstev (2014).

[6] J. M. Almendros-Jiménez, A. Becerra-Terón, Discovery and diagnosis of wrong SPARQL queries with ontology and constraint reasoning, Expert Systems with Applications 165 (2021) 113772. URL: https://doi.org/10.1016/j.eswa.2020.113772. doi:10.1016/j.eswa.2020.113772.

[7] M. P. di Buono, P. Cimiano, M. F. Elahi, F. Grimm, Terme-à-llod: Simplifying the conversion and hosting of terminological resources as linked data, in: M. Ionov, J. P. McCrae, C. Chiarcos, T. Declerck, J. Bosque-Gil, J. Gracia (Eds.), Proceedings of the 7th Workshop on Linked Data in Linguistics, LDL@LREC 2020, Marseille, France, May 2020, European Language Resources Association, 2020, pp. 28–35. URL: https://www.aclweb.org/anthology/2020.ldl-1.5/.

[8] C. Chiarcos, P. Cimiano, T. Declerck, J. P. McCrae, Linguistic linked open data (LLOD). introduction and overview, in: C. Chiarcos, P. Cimiano, T. Declerck, J. P. McCrae (Eds.), Proceedings of the 2nd Workshop on Linked Data in Linguistics, LDL 2013: Representing and linking lexicons, terminologies and other language data, Pisa, Italy, September 23, 2013, Association for Computational Linguistics, 2013, pp. i–xi. URL: https://www.aclweb.org/anthology/W13-5501/.