Introducing FREME: Deploying Linguistic Linked Data

Felix Sasaki[1], Tatiana Gornostay[2], Milan Dojchinovski[3], Michele Osella[4], Erik Mannens[5], Giannis Stoitsis[6], Phil Ritchie[7], Kevin Koidl[8]

[1] DFKI / W3C fellow, felix.sasaki@dfki.de; [2] Tilde, tatiana.gornostay@tilde.lv; [3] InfAI, milan.dojchinovski@fit.cvut.cz; [4] ISMB, osella@ismb.it; [5] iMinds, erik.mannens@ugent.be; [6] Agro-Know, stoitsis@agroknow.gr; [7] VistaTEC, philr@vistatec.ie; [8] Wripl, kevin@wripl.com

Abstract. This paper introduces the FREME project, a new Horizon 2020 innovation action. It aims at building an open framework of e-Services for multilingual and semantic enrichment of digital content, based on a reusable set of open Application Programme Interfaces and Graphical User Interfaces to FREME enrichment services. In addition, the paper discusses how the project deploys Linguistic Linked Data (LLD), especially existing LLD resources, LLD best practices and the LLD reference architecture.

1 Introduction

FREME is a Horizon 2020 innovation action that started in February 2015 with the duration of two years.[1] Partners are DFKI (as coordinator), Tilde, iMinds, Agro-Know, Wripl, VistaTEC, InfAI and ISMB. The project aims at building an open framework of e-Services for multilingual and semantic enrichment of digital content.

2 Motivation for FREME

The growing amount of digital content across languages, sectors and borders leads to both challenges and business opportunities for many industries. Linked data (LD) and language technology (LT) solutions exist, providing e.g. machine translation, entity recognition, or multilingual linked data sets. These solutions face several issues, e.g.: a plethora of content formats to process; adaptability and "silo solution" dependency; and usability in an industry application scenario: the lack of adequate tooling for a given or new group of user types (authors, translators, data wranglers or scientists etc.) in selected business scenarios.

The FREME framework addresses these issues by providing a **reusable set of open Application Programme Interfaces and Graphical User Interfaces** to FREME **enrichment services**. In this way, the project improves the existing **processes of digital content management**. The improvement goes through the whole content value chain: content creation (or authoring), content translation/localization, publishing and access to content including cross-language sharing and personalised content recom-

[1] See the FREME homepage at http://www.freme-project.eu .

mendations. Thus, the motivation is to open new opportunities for all sectors that are involved in digital content management.

3 Basic Concepts of FREME

3.1 Overview

Figure 1 visualizes the basic concept of the project.

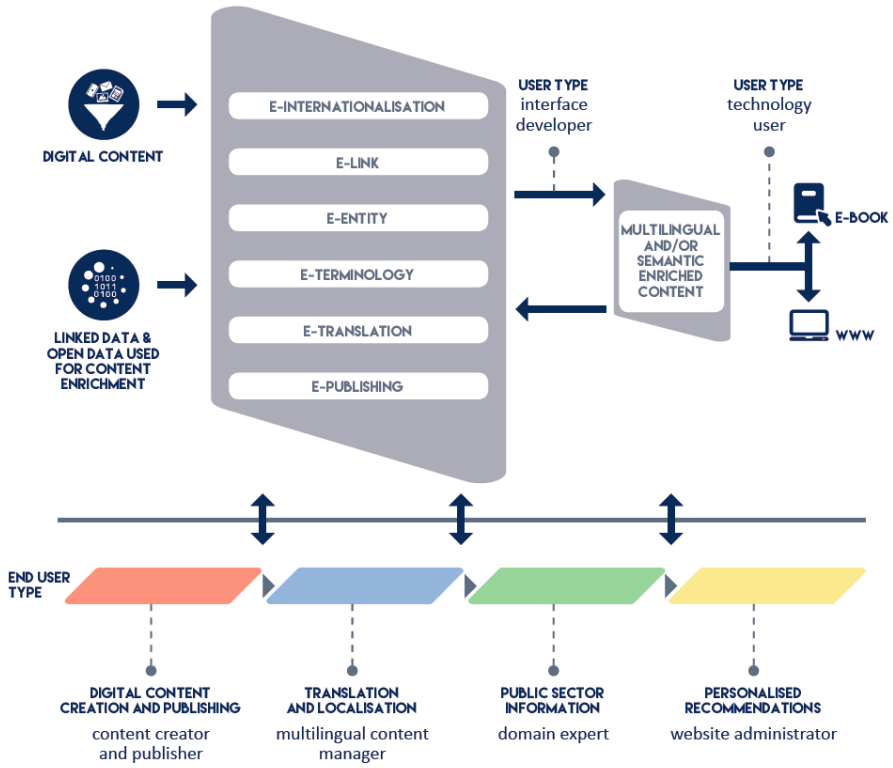

Figure 1: Overview of FREME

The main goal is to provide a set of interfaces for enrichment of **digital content**. We understand digital content as any type of content that exists in a digital form (text, video, audio, images, and others). Digital content is stored in various formats, e.g, textual content can be stored in structured formats (e.g. using the linked data technology stack) and unstructured formats (using e.g. PDF, PPTX etc.). The share of unstructured representation of digital content still prevails to a great extent. We work with the textual type of digital content in its structured and unstructured formats. We aim at transforming unstructured content into its structured representation.

Enrichment services. By enrichment we understand annotation of content with additional information. Content can be enriched on any step of its value chain that makes it intelligent, i.e., discoverable, interoperable, and aggregatable further on.
Multilingual enrichment. Under multilingual enrichment we understand annotation of content with additional linguistic information in a language or languages other than the language of content itself. The following languages of the project partners and / or their customers are in focus: English, German, Dutch, French, Italian, Spanish, Greek, Latvian, Lithuanian, and Estonian.
Semantic enrichment. Under semantic enrichment we understand annotation of content with additional information that transforms unstructured content into its structured representation. Semantic enrichment deploys the semantic richness provided by linked open vocabularies, as well as data sources that are not (yet) available as linked data.
In FREME, enrichment information is stored in two ways: first, as information using the Natural Language Processing Interchange Format (NIF), see Hellmann et al. (2013). This storage is independent of the underlying format: the enrichment information is stored in a stand-off manner, with pointers to the original location of content items. Second, FREME allows storing information inside the enriched content itself. The approach for storing the information then depends on the format in question. E.g. in the case of HTML5 and XML, the project relies on the Internationalization Tag Set (ITS) 2.0.

3.2 e-Services

The FREME framework develops and integrates **e-Services** based on existing and mature technologies. By e-Services we understand in most cases RESTful web services and graphical user interfaces. Details for each of the six e-Services are provided below. The framework is designed in an extensible manner, so that both project partners as well as external partners are able to add more services.
e-Translation is based on cloud machine translation services for building custom machine translation systems. The service takes content to be translated and the source and target language parameters as input.
e-Terminology: is based on cloud terminology services for terminology management and terminology annotation. The service takes content as input and enriches the content with information from terminology data bases.
e-Entity is based on entity recognition and existing linked entity data sets. The service takes content as input and enriches it with information related to entities (e.g. names of persons, places, etc.).
e-Internationalization supports other e-services to handle Internationalization Tag Set (ITS) 2.0 so called “data categories”: these are metadata items for handling the multilingual content production cycle. For example, the “Translate” data category specifies whether a given piece of content should be translated or not. This is of relevance e.g. for the e-Translation service. The “Terminology” data category provides a standardised way to store the output of e-Terminology as part of a NIF representation or in content formats. The “Text Analysis” data category allows storing the output of e-Entity.

e-Link is based on NIF and various (linked open) data sets. e-Link receives NIF documents (with annotated entities) and performs enrichment relying on the data sets. An example usage would be: have a NIF document with the entity "Berlin" annotated and enrich it with information from DBpedia about the current population of Berlin. The difference to querying linked data directly via SPARQL is that the e-Link query approach uses a query template mechanism. It hides the details of the actual query. The user of e-Link just has to provide parameters to the template. In the above example we assume a template called "query the population of a given place". The identifier for the place (e.g. http://dbpedia.org/resource/Berlin) is a parameter. The template based approach replies to the needs of business case partners. It enables them to enrich digital content with information from data sources without having to become linked data experts. During configuration of e-Link, these experts set up the templates for a given application.
e-Publishing has two aspects. First, it is a cloud based content authoring environment, allowing content authors to deal with e-Services in a WYSIWYG manner. Second, e-Publishing is a web service. Input is digital content in various forms (e.g. plain text, HTML). Output is content made available in the EPUB 3 format. We use EPUB 3 since it is the standardised format for representing digital book content.

e-Services can be deployed independently of each other. In addition, some e-Services can benefit from processing the enrichment information created via other e-Services. This is made possible by using NIF as the format for storing and pipelining enrichment information. Example **chains of e-Services** that are of interest for business case partners (as of writing) are e.g.:

- e-Entity, e-Link
- e-Entity, e-Link, e-Translate
- e-Entity, e-Link, e-Translation, e-Publishing
- e-Entity, e-Link, e-Terminology

The framework does not hard wire these chains. Processing and then storing the enrichment information using NIF enables these and other chains. NIF is both input and output of e-Services.

3.3 Business Cases

The innovation, robustness and usability of the FREME framework of e-Services is shaped by the four FREME real world business cases.
BC 1: Authoring and publishing multilingually and semantically enriched eBooks. For publishing companies, digital content itself is exploding and is loosing value. Via the project partner iMinds, we build the e-Services so that they provide additional value to publishers, going beyond digital publications. Initial discussions hint that enhanced search engine optimisation via FREME e-Services could be an attractive FREME application scenario for many publishing companies.
BC 2: Integrating semantic enrichment into multilingual content in translation and localisation. In the translation and localisation industry, demand for translation and the need for speed and quality are increasing. At the same time prices being paid are going down. Via the project partner VistaTEC, FREME allows for integrating

enrichment functionalities in localisation workflows. The outcome enables localisation companies to provide value beyond translation, e.g. by adding information from linked (open) data sources via e-Link to translated content.

BC 3 Enhancing the cross-language access and sharing of open agriculture and food data. In the area of public sector information, the discovery of data often is difficult due to missing multilingual metadata. E.g. many metadata items in the agriculture area are not in the language of the person who wants to use the metadata for search. Via the project partner Agro-know, a key player in the agriculture and food data domain, FREME tackles this challenge. The e-Translation service allows for (metadata) automated translation and in this way fosters cross-language, public information access.

BC 4 Personalisation of Web content. Web site personalisation is a field with many emerging solutions and companies. Currently many solutions focus on English speaking markets. Via the project partner Wripl, FREME demonstrates how to deploy underlying technologies e.g. via e-Entity in a larger number of languages, enabling SMEs and start-ups to reach out to global markets.

4 Linguistic Linked Data and FREME

The data value chains that are built with FREME rely heavily on linguistic linked data sets (LLD). The LIDER project is crucial in providing the basis for a linguistic linked data cloud[2]. LIDER fosters LLD as a basis for content analytics tasks of unstructured multilingual cross-media content. FREME e-Services, especially e-Entity, e-Link and E-Terminology, can be seen as prototypical examples of content analytics tasks. By providing these services together with e-Translation and several metadata items relevant for translation workflow information (via e-Internationalisation), FREME provides a technology stack that spans across content analytics and machine translation technologies.

The relevance of LIDER work on LLD can be seen in three areas: creation of LLD data sets, best practices on multilingual linguistic linked data, and deployment of the LIDER reference architecture.

4.1 Linguistic Linked Data Sets

Data sets are relevant for FREME in two ways. First, as content to be enriched via FREME. These data sets mostly come from business case partners and are specific to their needs and customers. Second, data sets to be used in enrichment e-Services. Here, figure 2 provides an overview of relevant data sets.

[2] See the LIDER homepage at http://lider-project.eu/ and an overview of the LLD at http://linguistic-lod.org/llod-cloud .

Data set	Data type	Data volume	Sector	Language	e-Service
DBpedia	Linked Data (RDF)	500GB RDF data	Multi-domain	119	e-Entity, e-Terminology, e-Translation
TaaS database	TBX	About 3.2 M terms	Multi-domain	24	e-Terminology
EuroVoc	Linked Open Data (SKOS)	About 7 K concepts	Multi-domain	23	e-Terminology
AgroVoc	Linked Open Data (RDF)	32 K concepts	Agriculture and food safety	20	e-Terminology
LinkedGeoData	Linked Open Data (RDF)	2 billion triples	Geography	20	e-Entity
NIF DBpedia Corpus	NIF (RDF)	Wikipedia article text - size?	Multi-domain	171	e-Entity
GeoNames	TSV	Above 8 M place names	Geography	mostly EN	e-Entity
Joint Research Centre Names	CSV	0.5 million name variants	NLP	30	e-Entity
BabelNet	Lemon (RDF)	13GB 1.1 billion triples	General	50	e-Translation
EU Open Data Portal	Mostly CSV and XLS(X)	6500 datasets	Multi-domain	23	e-Entity
DGT-Translation Memory	TMX	38 million translation units	Multi-domain	23	e-Translation
LetsMT! parallel corpora	Various	2.5 B bilingual 5.6 B monolingual sentences	Multi-domain	109	e-Translation

Figure 2: Data sets relevant for FREME

Not all data sets are linguistic linked data sets. E.g. the LetsMT! parallel corpora are not represented in RDF. Some of the data sets are moving towards the linguistic linked data cloud and will be made available via FREME as linked data in the technical sense, e.g. part of terminological data in the TaaS database.
Initial discussions in the project have shown that in some cases a non-linked data representation of linguistic resources with a clear path towards linked data (e.g. by providing URIs for all data items) is the preferable approach for technical reasons. E.g., tooling for machine translation training or for training of statistical named entity recognition currently is far more efficient relying on non-linked data representations. On the other hand, for exchanging data sets and for enriching them with additional information, linked data representations are the more adequate approach.
Similar lessons have been learned in the FALCON project[3], with a focus on using LLD in translation and localisation workflows. In FREME we take a similar approach, driven by tooling available in the four business cases. An additional goal then is to make this tooling linguistic linked data aware, e.g. providing linked data enabled machine translation systems.

4.2 Best Practices for the Creation of LLD

As discussed in the previous section, many data sets are not yet available as linguistic linked data. The LIDER project is working on best practises for creating LLD. This endeavour is undertaken under the helm of the W3C “Best Practises for Multilingual

[3] See http://falcon-project.eu/ for more information.

Linked Open Data" (BPMLOD) community group. As of writing, three best practises have been drafted:[4]

- General Patterns: a set of common practices and patterns that can be applied to publish linked data in a multilingual context.
- Guidelines for creating bilingual dictionaries.
- Guidelines for creating multilingual dictionaries.

All of these best practices are relevant for FREME business cases. The business case partners have their own data sets that they want to deploy in e-Services. In e-Translation, data sets can be used to provide translations for given lexical items. Currently there is a plethora of formats for such data sets. As part of deploying the best practices, we will rely on LEMON to representing bilingual and multilingual dictionaries.

BabelNet, see Navigli and Ponzetto (2012), is a resource that demonstrates the approach towards multilingual dictionaries. BabelNet is crucial for building general, domain independent multilingual and semantic enrichment applications. The tool Babelfy shows how to deploy BabelNet for such applications. Babelfy also demonstrates the approach (see section 4.1) of relying on LLD resources (here BabelNet) not in a native RDF representation but using them in as part of other tooling, i.e. for statistical training of named entity recognition.

For the conversion of LLD resources, off-the-shelf tooling is crucial. In the realm of the BPMLOD group, a TBX2RDF converter has been created. This implementation will help FREME to tackle conversion tasks, e.g. for the forehand mentioned TaaS database. In addition, it demonstrates the best practice of using the LEMON model for representing terminological resources as linguistic linked data.

4.3 FREME and the LIDER Reference Architecture

Within the LIDER project, a reference architecture for working with linguistic linked data has been created, cf. Koidl et al. (2014), esp. section 4.2. FREME instantiates several parts of the architecture.

e-Services as LLD aware services. By using NIF as the interchange format between e-Services, FREME provides e-Services as LLD aware services. In the terminology of the reference architecture the e-Services allow to constitute LLD based workflows.

Service composition via using NIF as interchange format. The combination of e-Services via FREME is an example of linked data service composition. The e-Services are LLD aware: both service workflow input/output and the actual interfaces comply to linked data standards and best practices. The current approach in FREME does not foresee a declarative description for composing services. This is left to the software client using the e-Services.

We connect to the LIDER reference architecture for two reasons. First, it eases the task of knowledge and technology transfer. Via the architecture, several FREME partners learn more easily how to build linguistic linked data enabled applications. Second, the reference architecture can also be seen as providing input to standardisation activities within W3C or other organisations. The W3C "Linked Data for Lan-

[4] See http://bpmlod.github.io/report/ for details.

guage Technologies" (LD4LT) community group serves as a forum also for LIDER and now also FREME to discuss this and other potential standardisation tasks. In a long term the e-Services may become the basis for standardised processing of both data and language technologies on the Web. However, this is not a main work item of FREME.

5 Conclusions and Next Steps

This paper introduced the FREME project: its motivation and goals, the outline of e-Services, and the four business cases. We then discussed the role of linguistic linked data for FREME, including existing data sets, best practices and tooling for new data sets, and the LIDER reference architecture for working with linguistic linked data.

As of writing, early prototypes of e-Services are available. The e-Services are being developed in an agile manner, taking feedback from the four business cases into account. Next steps will be including this feedback. A special focus that relates to the topic of this paper is linguistic linked data sources. FREME is looking for working with data set providers who could make their data set available via the e-Services, data set users who want to use LLD for multilingual and semantic enrichment, and providers of multilingual and semantic technologies. The last group could benefit from FREME by making components available for a larger audience, crossing the realms of data and language technologies as well as several industry sectors.

Acknowledgements

The FREME project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644771.

References

Hellmann, S., J. Lehmann, S. Auer and M. Brümmer (2013). *Integrating NLP using Linked Data.* In: Proceedings of the 12th International Semantic Web Conference, Sydney, Australia, (2013).

Kodil, K., D. Lewis, P. Cimiano, M. Hartung, J. McCrae, C. Unger, V.R. Doncel, A. Gómez-Pérez, J. Gracia, M. Brümmer, S. Hellmann, B. Klimek, S. Auer, T. Flati, R. Navigli and P. Buitelaar (2014). *LIDER Deliverable D3.1.1: Linguistic Linked Data Reference Architecture – Phase I.* Available at http://lider-project.eu/sites/default/files/D3.1.1-v1.0.pdf

Navigli, R. and S. Ponzetto (2012). *BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network.* Artificial Intelligence, 193, Elsevier, 2012, pp. 217-250.