FREME: Multilingual Semantic Enrichment with Linked Data and Language Technologies

[1,2]Milan Dojchinovski, [3]Felix Sasaki, [4]Tatjana Gornostaja, [1]Sebastian Hellmann,
[5]Erik Mannens,[5]Frank Salliau, [6]Michele Osella, [7]Phil Ritchie, [8]Giannis Stoitsis,
[9]Kevin Koidl,[1]Markus Ackermann,[1]Nilesh Chakraborty
[1]InfAI, Germany; [2]FIT CTU in Prague; [3]DFKI, Germany; [4]Tilde, Latvia;
[5]iMinds, Belgium; [6]ISMB, Italy; [7]Vistatec, Ireland; [8]Agro-Know, Greece; [9]Wripl, Ireland
[1]surname@informatik.uni-leipzig.de; [2]milan.dojchinovski@fit.cvut.cz; [3]felix.sasaki@dfki.de; [4]tatjana.gornostaja@tilde.lv;
[5]name.surname@ugent.be; [6]osella@ismb.it; [7]philr@vistatec.ie; [8]stoitsis@agroknow.gr; [9]kevin@wripl.com

Abstract

In the recent years, Linked Data and Language Technology solutions gained popularity. Nevertheless, their coupling in real-world business is limited due to several issues. Existing products and services are developed for a particular domain, can be used only in combination with already integrated datasets or their language coverage is limited. In this paper, we present an innovative solution FREME - an open framework of e-Services for multilingual and semantic enrichment of digital content. The framework integrates six interoperable e-Services. We describe the core features of each e-Service and illustrate their usage in the context of four business cases: i) authoring and publishing; ii) translation and localisation; iii) cross-lingual access to data; and iv) personalised Web content recommendations. Business cases drive the design and development of the framework.

Keywords: Linked Data, NLP, multilinguality, enrichment

1. Introduction

In the last few years Linked Data and Language Technologies have reached a state of maturity. From only 294 Linked Datasets in September 2011, the Linked Open Data (LOD) cloud has grown to 1,091 datasets in April 2014 (Schmachtenberg et al., 2014). These developments triggered number of research in assuring certain level of quality making the data ready for consumption. Language Technologies, such as machine translation systems, entity recognition and terminology systems have also reached a certain level of maturity. Nevertheless, there is still a gap between the actual business needs and the language and linked data technologies, which needs to be bridged; usage **domains** usually differ and the tools need to be appropriately adopted, **multilinguality** and language coverage is not well supported, the tools are offered as ad-hoc solutions without **standard interfaces**, and existing solutions for **enrichment** of content is limited only to specific knowledge bases. In this paper, we present FREME, an open framework with e-Services for multilingual and semantic enrichment of digital content, which aims at developing APIs and GUIs in order to fill this gap and offer industry ready solution.

2. FREME e-Services

The FREME framework offers a set of e-Services for multilingual and semantic enrichment of content. The services are available as RESTful APIs. The design of the services was driven by the requirement of having interoperable APIs by using the NLP Interchange Format (NIF) (Hellmann et al., 2013) format and the Internationalization Tag Set (ITS) Version 2.0[1].

The framework defines following set of e-Services:

- *e-Entity:* perform named entity recognition (NER), classification and linking of entities in multilingual texts. The e-Entity is independent of the back end engine and currently it can be utilized with FREME NER and the DBpedia Spotlight[2] NER engines. While DBpedia Spotlight performs entity linking only against the DBpedia[3] dataset, FREME NER can be easily adapted to any freely available or proprietary dataset. A user can upload its datasets, preferably in the SKOS format[4], and then use the dataset for entity linking. To this end, FREME NER provides an API for managing datasets. Currently, FREME NER can perform entity recognition in English, German, Dutch, French, Italian and Spanish texts.

- *e-Link:* enrich content with additional information from available Linked Data sources. The enrichment is supported via a query template mechanism. Therefore, users do not need to deal with the complexity in writing Linked Data queries to retrieve enrichment information.

- *e-Translation:* offers cloud based translation from a given source language to a target language. Currently, 63 language pairs are covered. Same as for the e-Entity service, the e-Translation service is independent from the back end machine translation system, so any other engine can be easily plugged in.

- *e-Publishing:* create digital publishing content in the open and standardised ePub[5] format. An HTML5 docu-

[1] http://www.w3.org/TR/its20/

[2] http://spotlight.dbpedia.org

[3] http://wiki.dbpedia.org/

[4] http://www.w3.org/TR/skos-reference/

[5] http://idpf.org/epub

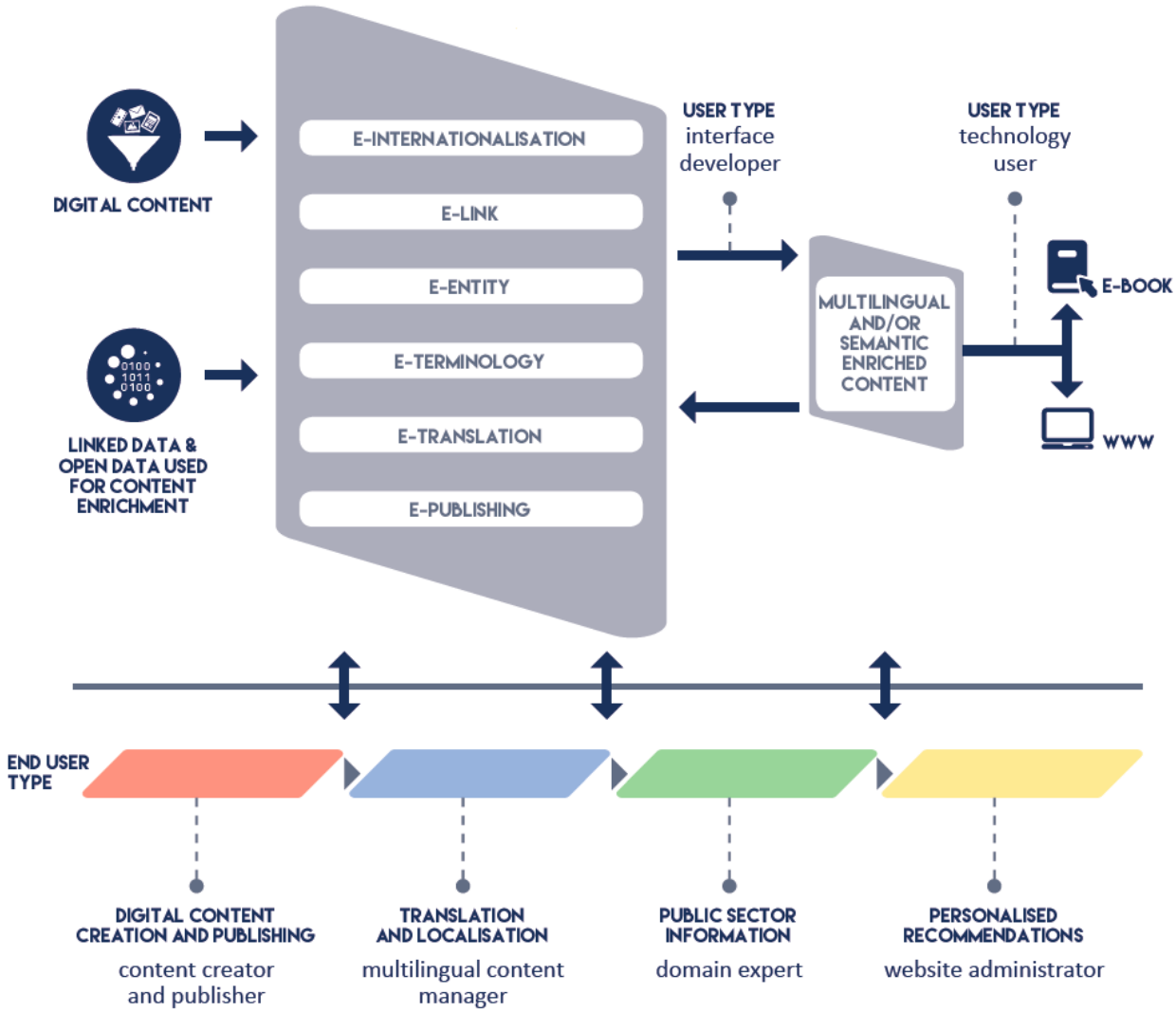

Figure 1: Overview of FREME.

ments with accompanying multimedia content are used to generate ready eBooks.

- *e-Terminology:* provides terminology mechanism to annotate content with terminology information. It can help out to automatically identify domain specific terms in documents and lookup for their translations. The terminology terms are organized in a controlled vocabulary. AGROVOC[6] is an example of such organized vocabulary from the food and agriculture domain.
- *e-Internationalisation:* supports other e-Services to make use of the ITS 2.0 which provides foundation to integrate automated creation and processing of multilingual Web content. For example, ITS defines "translate" data category to express information about whether a content should be translated or not, or the "terminology" data category to mark terms and associate them with their definitions. Another useful data category is the "text analysis" which is used to annotate content with lexical and conceptual information, for example to link an entity mention with its resource representation in a Linked Data dataset.

FREME Datasets. FREME also provides supports in creation and consumption of multilingual datasets offered as Linked Data. It provides access to such datasets via the e-Link and e-Entity services. Datasets from various domains such as DBpedia, the Geopolitical ontology[7], AGROVOC, Organization Name Linked Data (ONLD)[8], the Virtual International Autority File datasets (VIAF)[9] and ORCID[10] are considered and offered.

Availability and License. A detailed API documentation for the FREME e-Services is online available[11]. Latest news about the framework and other FREME related information

[6] http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus

[7] http://www.fao.org/countryprofiles/geoinfo/en/

[8] http://www.lib.ncsu.edu/ld/onld/

[9] http://viaf.org/

[10] http://orcid.org/

[11] http://api.freme-project.eu/doc/0.5/

can be also found at http://www.freme-project.eu/. The framework itsels if available under the Apache 2.0 license. For the underlying services various licenses apply.

3. Use Cases

The development of the FREME e-Services is driven by four real-world business cases.

BC 1: Authoring and publishing multilingually and semantically enriched eBooks. The FREME e-Services provides opportunity for publishers in several views: decrease production costs, increase discoverability of their published content, richer and more valuable content, and efficient localisation of their content. FREME can suport this list of needs using the e-Translate, e-Entity and e-Link services.

BC 2: Semantic enrichment into multilingual content in translation and localisation. Vistatec is a company offering localizaton and translation services. As part of the highly competitive translation and localisation industry Vistatec faces the challenge of differentiating their services and highlighting the unique value of their solutions. In this direction, the FREME e-Services can help Vistatec to: i) improve the accuracy and relevance of translated content; and ii) add customer, and ultimately, end user value to content that we translate. FREME e-Services can aid translators and editors in crafting contextually and linguistically relevant and accurate translations and, providing content consumers with semantically enriched and more interactive experiences. To this end, Vistatec is integrating its open source translation editor, Ocelot[12] with FREME by enabling Ocelot as a client of the e-services, particularly e-Entity, e-Link and e-Internationalization.

BC 3: Cross-language access, enrichment and sharing of open agriculture and food data. Currently, there are large organizations working on agriculture and food topics producing and sharing scientific and educational digital content. Nevertheless, the content is offered as monolingual and accompanied with poor metadata information. The Agro-Know company tackles these issues and with help from FREME e-Services aims at solving them. Using e-Entity and e-Link services, Agro-Know will provide semantic enrichment, using e-Translation offer the content to the multilingual marker and with help of e-Internationalisation enable automated creation and processing of the content.

BC 4: Personalised content recommendations. Personalised content-based recommendation services such as Wripl[13] are based on efficient understanding of the aboutness of the documents in order to generate relevant recommendations. It is a recent trend that many systems are moving towards understanding documents in terms of entities. In the context of FREME, Wripl can benefit from performing entity recognition in multilingual documents (using e-Entity), enrich the documents with additional information (using e-Link), and identify terms in the content and lookup for their translations (using e-Terminology).

[12] http://open.vistatec.com/ocelot
[13] http://wripl.com/

4. Details on e-Services

In this section, we provide details on e-Services and toy examples of their application.

4.1. E-Entity

Resource: /e-entity/{ner-engine}/documents
Input: Berlin is the capital of Germany.

```
:doc1#char=0,33
   a nif:RFC5147String, nif:Context ;
   nif:beginIndex "0" ;
   nif:endIndex "33" ;
   nif:isString "Berlin is the capital of
       Germany." .

:doc1#char=0,6
   a nif:RFC5147String ;
   nif:beginIndex "0" ;
   nif:endIndex "6" ;
   nif:anchorOf "Berlin" ;
   itsrdf:taIdentRef dbpedia:Berlin ;
   itsrdf:taClassRef nerd:Location ;
   nif:referenceContext :doc1#char=0,33 .
```

Listing 1: Results from the e-Entity service.

Listing 1 shows an example of results returned by the e-Entity service. Each recognized entity is identified with its position in the source text, its type class from a controlled classification system and its link to a knowledge base (e.g., DBpedia). The e-Entity service can be also configured to perform linking to a specific knowledge base. For more information about the offered features we refer the reader to the online documentation of the FREME APIs[14].

4.2. E-Link

Resource: /e-link/documents/?templateid={id}
Input: text with recognized entities in the NIF format (i.e., output from the e-Entity service).

```
@prefix dbpedia: <http://dbpedia.org/
    resource/>
@prefix foaf: <http://xmlns.com/foaf/0.1/>

dbpedia:Ethnological_Museum_of_Berlin
   foaf:based_near   dbpedia:Berlin .
dbpedia:Museum_of_Asian_Art
   foaf:based_near   dbpedia:Berlin .
```

Listing 2: Results from the e-Links service.

Listing 2 shows the results of execution of the e-Link service. The results from the execution is an enriched document where each city in the source document is enriched with a list of nearest museums. Currently, the e-Link service accepts a text in the NIF format, e.g., with a list of recognized entities, and enriches this content with additional information. The scope of enrichment is defined using a pre-defined templates, designed as SPARQL queries. The templates are used to hide the details of the actual query. An example of a template is one that enriches each city occurring in the text with a list of closest museums. It is a requirement,

[14] http://api.freme-project.eu/doc/0.3/

that a template needs to be defined before the enrichment operation is executed.

4.3. E-Translation

Resource: `/e-translation/{engine}`
Input: text or a NIF document containing text for translation.

```
:doc1#char=0,33
  a nif:RFC5147String , nif:Context ;
  nif:beginIndex "0";
  nif:endIndex "33" ;
  nif:isString "Berlin is the capital of
      Germany."@en ;
  itsrdf:target "Berlin ist die
      Bundeshauptstadt der Bundesrepublik
      Deutschland."@de .
```

Listing 3: Results from the e-Translation service.

Listing 3 shows the results from translation text using the e-Translation service from English to German. The output document contains both, the translation and the source text.

4.4. E-Terminology

Resource: `/e-terminology/{engine}`
Input: text or a NIF document containing text for enrichment with terminology information.
The output of e-Terminology contains for each term information using the lemon ontolex model, including e.g. a unique term identifiers, association with (a set of) terminological, language agnostic domains and concepts, and language specific lexical representations. These can be interpreted by the forehand described e-Translation service to improve translation output.

4.5. E-Publishing

Resource: `/e-publishing/html`
Input: a set of HTML documents and a metadata file defining eBook characteristics.
The output of ePublishing is an eBook following the ePub 3 standard. If the service takes HTML as input that has been enriched via other e-Services, this can produce a semantically enriched eBook.

4.6. E-Internationalisation and Pipelining

FREME allows via e-Internationalisation to process various digital content formats, like HTML5, general XML, XLIFF or other markup formats that are relevant for certain business cases. For some formats also so-called roundtripping is possible. That is, the outcome of certain e-Services can be stored in the original format again. This is e.g. important for the forehand described semantically enriched eBook.
FREME provides also a pipelining service. This allows the user to chain several e-Services without making separate web services requests. The re-use of pipelines can benefit from such chains.
The pipelining service shows that FREME is not a set of unrelated services but a framework. The use of NIF as the enrichment format and API specification eases the extension of the framework with services that use these standards. No change in the FREME backend is needed. As of writing, access to FREME is not only via web services requests, but also a growing set of graphical user interfaces has been created.

4.7. Example usage: FREME in Ocelot

Ocelot is an open source translation editor for localisation quality assurance. It integrates FREME e-Services in a non-invasive way for the translator. As the translator is working, Ocelot sends the document content in background to the FREME e-Services. The enriched data is then displayed to the translator in pop-up windows as suggested translations in much the same way as Translation Memory fuzzy matches and Machine Translation suggestions are currently.
In this way translators can benefit from automatic named entity and terminology recognition making the morphological and lexical aspect of their task easier. Additionally, they are presented with automatic target language translations which contain semantic links to related topics which can aid their understanding of the material requiring translation.
The translator has full control of what enrichment is saved along with their final translation. In this way their job is elevated further in that she is making decisions which will impact the end user experience of the content consumer. Ultimately, this enhances the experience of content end users which will therefore be valued by Vistatec customers thus driving Vistatec revenue whilst making a positive contribution to the task of translators.

5. Conclusions and Future Work

Although Linked Data and Language Technologies have attracted attention there is still gap between the available tooling and actual business needs. In this paper, we described the motivations behind the FREME framework. We describe the current state of the development and describe the core functionalities of the six e-Services offered by the framework. We presented the four business cases which guide the development and discussed how FREME can support their needs. We also present the most recent integration of the FREME e-Services in the Ocelot translation editor.
The FREME framework is developed in an agile manner and currently its third release has been announced. Since the quality of enrichments and the scalability of the e-Services are important factors for a real-world scenarios, in the following period we plan to focus on this two aspects. In our future work, we will continue to work on creation of Linked Data datasets and their integration in the framework.

Acknowledgements This work was supported via the FREME project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644771.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating nlp using linked data. In Harith Alani, et al., editors, *The Semantic Web – ISWC 2013*, volume 8219 of *Lecture Notes in Computer Science*, pages 98–113. Springer Berlin Heidelberg.

Schmachtenberg, M., Paulheim, H., and Bizer, C. (2014). Adoption of linked data best practices in different topical domains. In *The Semantic Web – ISWC 2014*, Lecture Notes in Computer Science. Springer Berlin Heidelberg.