OPEN FRAMEWORK OF E-SERVICES FOR MULTILINGUAL AND SEMANTIC ENRICHMENT OF DIGITAL CONTENT

Data Innovation and Take up in FREME

Co-funded by the Horizon 2020
Framework Programme of the European Union

DELIVERABLE NUMBER	D1.7
DELIVERABLE TITLE	Data Innovation and Take up in FREME
RESPONSIBLE AUTHOR	InfAI

GRANT AGREEMENT N.	644771
PROJECT REF. NO	H2020-644771
PROJECT ACRONYM	FREME
PROJECT FULL NAME	Open Framework of E-Services for Multilingual and Semantic Enrichment of Digital Content
STARTING DATE (DUR.)	01/02/2015 (24 months)
ENDING DATE	31/01/2016
PROJECT WEBSITE	www.freme-project.eu
COORDINATOR	Felix Sasaki
ADDRESS	DFKI Alt-Moabit 91c 10559 Berlin
REPLY TO	felix.sasaki@dfki.de
PHONE	+49-30 23895 1807
FAX	+49-30 23895 1810
EU PROJECT OFFICER	Pierre Paul-Sondag

WORKPACKAGE N. | TITLE	WP1 | Data Innovation and take up and technology transfer
WORKPACKAGE LEADER	Tilde
DELIVERABLE N. | TITLE	D1.7 | Data Innovation and take up in FREME
RESPONSIBLE AUTHOR	Milan Dojchinovski, InfAI
REPLY TO	dojchinovski@informatik.uni-leipzig.de
DOCUMENT URL	http://www.freme-project.eu/resources/deliverables/
DATE OF DELIVERY (CONTRACTUAL)	M21
DATE OF DELIVERY (SUBMITTED)	15 November 2016 (after extension)
VERSION | STATUS	V5 | Final
NATURE	O (Other)
DISSEMINATION LEVEL	PU (Public)
AUTHORS (PARTNER)	InfAI

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
01	Initial draft of the deliverable.	27/10/2016	Milan Dojchinovski, INFAI
02	Review within consortium.	31/10/2016	Felix Sasaki, DFKI
03	Review comments have been integrated.	2/11/2016	Milan Dojchinovski, INFAI
04	Implemented comments received from Vistatec and prepared pre-final version.	9/11/2016	Milan Dojchinovski, INFAI
05	Reviewed by WP1 leader	11/11/2016	Tatjana Gornostaja, TILDE

PARTICIPANTS		CONTACT
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI, Germany)		Felix Sasaki Email: felix.sasaki@dfki.de
Tilde SIA (Tilde, Latvia)		Tatjana Gornostaja Email: tatjana.gornostaja@tilde.com
iMINDS VZW (IMINDS, Belgium)		Frank Salliau Email: frank.salliau@ugent.be
Agro-Know IKE (Agroknow, Greece)		Giannis Stoitsis Email: stoitsis@agroknow.com
Wripl Technologies Limited (Wripl, Ireland)		Kevin Koidl Email: kevin@wripl.com
VistaTEC EV (VTEC, Ireland)		Phil Ritchie Email: phil.ritchie@vistatec.com
Institut für Angewandte Informatik Ev (InfAI, Germany)		Sebastian Hellmann Email: hellmann@informatik.uni-leipzig.de
Istituto Superiore Mario Boella (ISMB, Italy)		Michele Osella Email: osella@ismb.

ACRONYMS LIST

RDF	Resource Description Framework
XML	Extensible Markup Language
NIF	NLP Interchange Format
ITS 2.0	Internationalization Tag Set (ITS) Version 2.0
NERD	Named Entity Recognition and Disambiguation ontology
XLIFF	XML Localisation Interchange File Format
TaaS	Terminology as a Service
HTML	HyperText Markup Language

EXECUTIVE SUMMARY

This deliverable summarises the work done in Task 1.2 on data innovation and take up in FREME (M4-M21). It primarily defines how interoperability at the syntactic and semantic level is handled in FREME and how provenance of information is ensured.

TABLE OF CONTENTS

1. INTRODUCTION

Dealing with different content and data formats is a crucial for many Web scenarios. Currently, the content can be structured in various data formats which differ in syntax and semantics. FREME ultimately addresses the heterogeneity of the content and data formats at the following levels:

- Syntactic interoperability: the need for interoperability between data formats at syntactic level. FREME considers the interoperability at the syntactic level of the following data formats:
  - HTML (Hypertext Markup Language) - a standardised markup language for creating Web pages and Web applications. FREME partners extensively exploit and rely on this format.
  - XML - a markup language which defines rules and syntax for encoding documents in machine and human readable format.
  - XLIFF (XML Localisation Interchange File Format) - a standard format based on XML used for exchange of localised data between tools during the localisation process.
  - OpenOffice - and open-source software collection of office products, based on the standard Open Document Format (ISO/IEC 26300-1:2015).
  - RDF (Resource Description Framework) - an RDF format defines a metadata data model as a standard model for data interchange on the Web.
  - NIF (Natural Language Processing (NLP) Interchange Format) - an RDF/OWL-based format that aims to achieve the interoperability between NLP tools, language resources, and annotations. NIF is the core interoperability format used within FREME.
  - Plain text - FREME can also process plain text, and make it compatible with the other data formats.
- Semantic interoperability: addresses the need for exchange of data with unambiguous, shared meaning. FREME addresses this challenge with help of *data formats* such as NIF, ITS, RDF, *ontologies* (DBpedia and NERD) and *classification systems* such as TaaS.
  - DBpedia Ontology[1] - core ontology defined by DBpedia, which is used for expressing fine-grained entity types with the FREME e-Entity service. It is a shallow, cross-domain ontology defined and created by the DBpedia community.
  - NERD[2] (Named Entity Recognition and Disambiguation) - a set of mappings established between different named entity taxonomies.
  - TaaS[3] (Terminology Domain Classification System) - defines taxonomy for domain classification. The classification system is based and mapped to Eurovoc[4], the EU's multilingual thesaurus.
  - RDF, NIF, and ITS - RDF defines a metadata data with a semantic aspect in the main focus. NIF builds on top of RDF and supports the interoperability between NLP tools, language resources, and annotations. ITS defines means which enhance the integration of automated processing of human language into core Web technologies. ITS complements other considered data formats in the process of semantic interoperability.
- Provenance: to keep the provenance of the data exchange and production within FREME. This is important in order to ensure the quality and interoperability of data and to allow software clients to make assessments about quality of enrichments. NIF has been extended to ensure lossless tracking of provenance across formats and systems.

[1] DBpedia Ontology - http://wiki.dbpedia.org/services-resources/ontology

[2] NERD - http://nerd.eurecom.fr/ontology

[3] TaaS - http://www.taas-project.eu/

[4] Eurovoc - http://eurovoc.europa.eu/

2. NLP INTERCHANGE FORMAT (NIF)

Natural Language Processing Interchange Format (NIF) is an RDF/OWL-based format that aims to achieve interoperability between Natural Language Processing (NLP) tools, language resources and annotations. Within the FREME framework NIF is used to i) manipulate with data and enrichment information and ii) provide the basis for interoperability among the FREME e-services at the syntactic as well as semantic level.

Support for NIF in the e-Entity service. The E-entity service contributes with named entity recognition (NER) functionalities. It performs the spotting of entity mentions, classification of named entities and linking entity mentions with their representation in a given knowledge base. The spotting phase identifies the exact location of entities in the given source document by providing their exact begin and end index. Next, the classification phase assigns entity types (ontological classes) to each entity mention. The granularity of the entity types can vary from few coarse grained types (from the NERD ontology) to very fine-grained types from the DBpedia ontology or another provided source. Along with these information, the e-Entity also provides confidence information for enrichments. In FREME, we encode this information using the NIF format. In Listing 1, we present an example of NER results stored in NIF.

```
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
@prefix nif:   <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .

<http://freme-project.eu/#collection>
      a             nif:ContextCollection ;
      nif:hasContext  <http://freme-project.eu/#offset_0_53> ;
      <http://purl.org/dc/terms/conformsTo>
            <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/2.1> .

<http://freme-project.eu/#offset_0_53>
      a             nif:Context , nif:OffsetBasedString ;
      nif:beginIndex  "0"^^xsd:nonNegativeInteger ;
      nif:endIndex    "53"^^xsd:nonNegativeInteger ;
      nif:isString    "Tokyo is the capital and most populous city of Japan."^^xsd:string .

<http://freme-project.eu/#offset_47_52>
      a               nif:OffsetBasedString , nif:Phrase ;
      nif:anchorOf        "Japan"^^xsd:string ;
      nif:annotationUnit    [ a                 nif:EntityOccurrence ;
                        itsrdf:taAnnotatorsRef  <http://freme-project.eu/tools/freme-ner> ;
                        itsrdf:taClassRef       <http://nerd.eurecom.fr/ontology#Location> ,
<http://dbpedia.org/ontology/Place> , <http://dbpedia.org/ontology/Location> ,
<http://dbpedia.org/ontology/Country> , <http://dbpedia.org/ontology/PopulatedPlace> ;
                        itsrdf:taConfidence     "0.9992633350787467"^^xsd:double ;
                        itsrdf:taIdentRef       <http://dbpedia.org/resource/Japan>
```

```
                    ] ;
    nif:beginIndex      "47"^^xsd:nonNegativeInteger ;
    nif:endIndex        "52"^^xsd:nonNegativeInteger ;
    nif:referenceContext  <http://freme-project.eu/#offset_0_53> .

<http://freme-project.eu/#offset_0_5>
    a                nif:OffsetBasedString , nif:Phrase ;
    nif:anchorOf        "Tokyo"^^xsd:string ;
    nif:annotationUnit    [ a                  nif:EntityOccurrence ;
                    itsrdf:taAnnotatorsRef  <http://freme-project.eu/tools/freme-ner> ;
                    itsrdf:taClassRef      <http://dbpedia.org/ontology/City> ,
<http://dbpedia.org/ontology/PopulatedPlace> , <http://dbpedia.org/ontology/Location> ,
<http://dbpedia.org/ontology/Settlement> , <http://dbpedia.org/ontology/Place> ,
<http://nerd.eurecom.fr/ontology#Location> ;
                    itsrdf:taConfidence     "0.9808069169938456"^^xsd:double ;
                    itsrdf:taIdentRef      <http://dbpedia.org/resource/Tokyo>
                    ] ;
    nif:beginIndex      "0"^^xsd:nonNegativeInteger ;
    nif:endIndex        "5"^^xsd:nonNegativeInteger ;
    nif:referenceContext  <http://freme-project.eu/#offset_0_53> .
```

Listing 1. Output of e-Entity in NIF.

Support for NIF in the e-Link service. The NIF format provides the foundation for the e-Link service. The e-Link service provides the support for enrichment of content with additional information from any available Linked Data dataset. It consumes NIF, which is then enriched with information from a specified knowledge base (e.g. DBpedia). Additional information is retrieved via predefined templates via dedicated SPARQL endpoints.

NIF serves as the foundation for the e-Link service and it can be used to enrich data with additional information from a Linked Open Data dataset. In Listing 2, we show an example of additional information retrieved via the e-Link service.

```
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
@prefix nif:   <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .

<http://freme-project.eu/#collection>
    a            nif:ContextCollection ;
    nif:hasContext  <http://freme-project.eu/#offset_0_14> ;
    <http://purl.org/dc/terms/conformsTo>
          <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/2.1> .
```

```
<http://freme-project.eu/#offset_0_14>
      a            nif:OffsetBasedString , nif:Context ;
      nif:beginIndex  "0"^^xsd:nonNegativeInteger ;
      nif:endIndex    "15"^^xsd:nonNegativeInteger ;
      nif:isString    "This is Tokyo."^^xsd:string .

<http://freme-project.eu/#offset_8_13>
      a                nif:Phrase , nif:OffsetBasedString ;
      nif:anchorOf        "Tokyo"^^xsd:string ;
      nif:annotationUnit    [ a                  nif:EntityOccurrence ;
                        itsrdf:taAnnotatorsRef  <http://freme-project.eu/tools/freme-ner> ;
                        itsrdf:taClassRef      <http://dbpedia.org/ontology/PopulatedPlace> ,
<http://dbpedia.org/ontology/Country> , <http://dbpedia.org/ontology/Location> ,
<http://dbpedia.org/ontology/Place> , <http://nerd.eurecom.fr/ontology#Location> ;
                        itsrdf:taConfidence     "0.9992633350787467"^^xsd:double ;
                        itsrdf:taIdentRef      <http://dbpedia.org/resource/Tokyo>
                       ] ;
      nif:beginIndex       "8"^^xsd:nonNegativeInteger ;
      nif:endIndex        "13"^^xsd:nonNegativeInteger ;
      nif:referenceContext  <http://freme-project.eu/#offset_0_14> .

<http://dbpedia.org/resource/Tokyo_Metropolitan_Teien_Art_Museum>
      <http://xmlns.com/foaf/0.1/based_near>
            <http://dbpedia.org/resource/Tokyo> .

<http://dbpedia.org/resource/Tokyo_Metropolitan_Museum_of_Photography>
      <http://xmlns.com/foaf/0.1/based_near>
            <http://dbpedia.org/resource/Tokyo> .

<http://dbpedia.org/resource/National_Showa_Memorial_Museum>
      <http://xmlns.com/foaf/0.1/based_near>
            <http://dbpedia.org/resource/Tokyo> .
```

Listing 2. Output of the e-Link service in NIF. The highlighted triples are provided via the e-Link service.

The content from Listing 2 has been enriched with a template which retrieves a list of museums (maximum 10) within a 50 km radius around each location entity. The template is executed against the DBpedia dataset. In the source document (see Listing 2), there is one entity representing the Tokyo city, and using the template we have retrieved 3 museums in the radius of 50 km around Tokyo (see the highlighted sections in Listing 2). The e-Link service is fully configurable via the templating mechanism and users can have a full control over the information they retrieve.

Support for NIF in the e-Internationalisation service. The e-Internationalisation service provides the support for wide range of data formats such as HTML, XML, NIF, etc. It also supports round tripping scenarios and the conversion from one format to another. The core standard behind the e-

Internationalisation service is the ITS 2.0 format, which allows the integration of automated processing of human language into Web content such as HTML. NIF extensively makes use of ITS 2.0 and the properties it defines. Section 3 and Section 4 provide more information on how e-Internationalisation contributes to the interoperability and provenance in FREME.

Support for NIF in the e-Terminology service. The e-Terminology service performs term identification, annotation, extraction and translation lookup. Same as the e-Entity service, the e-Terminology service encodes the term related information in the NIF format. In addition, it exploits the Lemon ontology to encode related lexical information. More information regarding the e-Terminology service can be found in the D1.8 deliverable and in a dedicated online tutorial[5].

Support for NIF in the e-Translation service. The e-Translation service provides automated machine translation of content. e-Translation encodes the source text together with its translation in a single NIF document. e-Translation also provides the support for ITS 2.0 which is used to instruct the translation engine whether or not to translate specific parts of the content. e-Translation, with support of NIF, can also process segmented translations. More information about the e-Translation service can be found in the deliverable D1.8.

[5] E-terminology tutorial - https://freme-project.github.io/tutorials/Getting-started-with-e-Terminology.html

3. INTEROPERABILITY IN FREME

One of the key requirements of the FREME framework is to define a set of interoperable services which can generate, exchange, and use the exchanged information. The following technologies and standards form the basis for creating interoperability within the FREME framework: URIs (web-scale data integration), RDF (directed, acyclic graphs, excellent tool support), SPARQL (scalable and mature graph database implementations), OWL (formal modelling language and rich data validation paradigms), and best practices around the NIF format [HLA+13] defined within the NLP2RDF project.

3.1 SYNTACTIC INTEROPERABILITY

The FREME framework consists of several e-services which can process content in various formats, such as HTML, XML, plain text, NIF, or XLIFF[6]. Each of these formats have own standard syntax and representation. Thus, in FREME, there is a need to align these formats and enable lossless conversion from one format to another.

The syntactic interoperability in FREME is enabled via reuse of the following data formats and standards. **Internationalisation Tag Set (ITS) Version 2.0[7].** ITS 2.0 is used to convey localisation information between different formats such as NIF, HTML and XML. In Listing 3, below we provide an example of an HTML content processed by the e-Entity service. The results from the processing is the same HTML document along with the enrichments embedded in its content. The output of the process contains ITS 2.0 "Text Analysis Markup" (e.g. "its-ta-class-ref") properties which are encoded in the NIF document.

```
<!DOCTYPE html>
<html>
  <body>
    <h1>Info about Diego Maradona.</h1>
    <p>Diego Maradona is from Argentina.</p>
  </body>
</html>
```

Listing 3. Input HTML for processing.

```
<!DOCTYPE html>
<html>
  <body>
    <h1>Info about <span data-its-ta-class-
refs="http://dbpedia.org/ontology/Person
http://dbpedia.org/ontology/SoccerManager" its-ta-class-
ref="http://dbpedia.org/ontology/Agent" its-ta-confidence="0.9763824695354" its-
ta-ident-ref="http://dbpedia.org/resource/Diego_Maradona">Diego
Maradona</span>.</h1>
    <p><span data-its-ta-class-refs="http://dbpedia.org/ontology/SoccerManager
http://dbpedia.org/ontology/Person" its-ta-class-
```

[6] http://docs.oasis-open.org/xliff/xliff-core/v2.0/os/xliff-core-v2.0-os.html

[7] https://www.w3.org/TR/its20/

```
ref="http://dbpedia.org/ontology/SportsManager" its-ta-
confidence="0.9869992701528016" its-ta-ident-
ref="http://dbpedia.org/resource/Diego_Maradona">Diego Maradona</span> is from
<span data-its-ta-class-refs="http://dbpedia.org/ontology/Country
http://dbpedia.org/ontology/Place http://dbpedia.org/ontology/Location" its-ta-
class-ref="http://dbpedia.org/ontology/Country" its-ta-
confidence="0.9804963628413852" its-ta-ident-
ref="http://dbpedia.org/resource/Argentina">Argentina</span>.</p>
  </body>
</html>
```

Listing 4. Enriched HTML.

3.1.1 A round tripping scenario: an HTML enrichment use case

The FREME round-tripping functionality allows users to enrich documents structured in specific formats and retrieve the results from the processing inline with the source documents. HTML is a widely used format and huge amount of content on the Web is generated in the HTML format. FREME can accept documents in HTML, process and enrich the content, and return the enrichments inline with the source content.

In order to realise this workflow, there is a need to: 1) extract the content for processing, 2) enrich the content using some of the available services, and 3) integrate the results in the source content, while keeping the structure of the content unchanged. In order to achieve this, the first step is a conversion of the HTML/XML to NIF. In this step, two NIF files are created. The first is a NIF file containing just the textual content extracted from the HTML/XML document. The second file contains a context that includes the markup. We call the latter “skeleton”. The skeleton file is locally saved while the first NIF file (without markup) is submitted to enriching FREME e-services. Next, the enriched NIF file and the skeleton file are submitted to the e-Internationalisation service which returns the original HTML/XML document with the addition of enrichment annotations.

Listing 5 presents a simple example of HTML which is submitted to the e-Entity service for processing. In this example, the e-Entity service is invoked by setting the input and output format parameter to HTML, which means that we want to retrieve the enrichments integrated in the source HTML document.

```
<html>
    <head>
        <title>Roundtripping</title>
    </head>
    <body>
        <p>Welcome to Dublin</p>
    </body>
</html>
```

Listing 5. Input HTML document sent for processing.

The HTML from Listing 5 is then processed by e-Internationalisation and two supporting NIF documents are created: one that represents the content (cf. Listing 6), and one that represents the "skeleton" of the HTML document (Listing 7).

```
<http://freme-project.eu/#offset_0_31>
      a             nif:OffsetBasedString , nif:Context ;
      nif:beginIndex  "0"^^xsd:nonNegativeInteger ;
      nif:endIndex    "31"^^xsd:nonNegativeInteger ;
      nif:isString    "Roundtripping Welcome to Dublin"@en .

<http://freme-project.eu/#offset_14_31>
      a                 nif:Phrase , nif:OffsetBasedString , nif:String ;
      nif:anchorOf        "Welcome to Dublin"@en ;
      nif:beginIndex       "14"^^xsd:nonNegativeInteger ;
      nif:endIndex        "31"^^xsd:nonNegativeInteger ;
      nif:referenceContext  <http://freme-project.eu/#offset_0_31> ;
      dc:identifier       "2" .

<http://freme-project.eu/#offset_0_13>
      a                 nif:Phrase , nif:OffsetBasedString ;
      nif:anchorOf        "Roundtripping"@en ;
      nif:beginIndex       "0"^^xsd:nonNegativeInteger ;
      nif:endIndex        "13"^^xsd:nonNegativeInteger ;
      nif:referenceContext  <http://freme-project.eu/#offset_0_31> ;
      dc:identifier       "1" .
```

Listing 6. Extracted content in NIF.

Listing 7 shows the skeleton representation of the HTML document which holds representation of the HTML structure.

```
<http://freme-project.eu/doc1/#offset_0_121>
      a             nif:OffsetBasedString , nif:Context , nif:String ;
      nif:beginIndex  "0"^^xsd:nonNegativeInteger ;
      nif:endIndex    "121"^^xsd:nonNegativeInteger ;
      nif:isString    "<!DOCTYPE
html>\r\n<html><head>\r\n\t<title>Roundtripping</title>\r\n</head>\r\n<body>\r\n<p>Welco
me to Dublin</p>\r\n\r\n</body></html>"@en .

<http://freme-project.eu/#offset_14_31>
      a                 nif:OffsetBasedString , nif:String ;
      nif:anchorOf        "Welcome to Dublin"@en ;
      nif:beginIndex       "14"^^xsd:nonNegativeInteger ;
```

```
    nif:endIndex        "31"^^xsd:nonNegativeInteger ;
    nif:referenceContext  <http://freme-project.eu/#offset_0_31> ;
    nif:wasConvertedFrom  <http://freme-project.eu/doc1/#offset_82_99> ;
    dc:identifier       "2" .

<http://freme-project.eu/#offset_0_13>
    a                nif:OffsetBasedString , nif:String ;
    nif:anchorOf        "Roundtripping"@en ;
    nif:beginIndex       "0"^^xsd:nonNegativeInteger ;
    nif:endIndex        "13"^^xsd:nonNegativeInteger ;
    nif:referenceContext  <http://freme-project.eu/#offset_0_31> ;
    nif:wasConvertedFrom  <http://freme-project.eu/doc1/#offset_39_52> ;
    dc:identifier       "1" .

<http://freme-project.eu/#offset_0_31>
    a            nif:OffsetBasedString , nif:Context , nif:String ;
    nif:beginIndex  "0"^^xsd:nonNegativeInteger ;
    nif:endIndex    "31"^^xsd:nonNegativeInteger ;
    nif:isString    "Roundtripping Welcome to Dublin"@en .
```

Listing 7. Skeleton representation.

Listing 8 shows the NIF document which holds the content sent and processed by the e-Entity service. The entities are spotted, classified and linked with their representation in the specified knowledge base, which is DBpedia in this example.

```
<http://freme-project.eu/#offset_0_31>
    a            nif:OffsetBasedString , nif:Context , nif:String ;
    nif:beginIndex  "0"^^xsd:nonNegativeInteger ;
    nif:endIndex    "31"^^xsd:nonNegativeInteger ;
    nif:isString    "Roundtripping Welcome to Dublin"@en .

<http://freme-project.eu/#offset_14_31>
    a                nif:Phrase , nif:OffsetBasedString , nif:String ;
    nif:referenceContext  <http://freme-project.eu/#offset_0_31> ;
    nif:anchorOf        "Welcome to Dublin"@en ;
    nif:beginIndex       "14"^^xsd:nonNegativeInteger ;
    nif:endIndex        "31"^^xsd:nonNegativeInteger ;
    dc:identifier       "2" .

<http://freme-project.eu/#offset_0_13>
    a                nif:Phrase , nif:OffsetBasedString , nif:String ;
    nif:referenceContext  <http://freme-project.eu/#offset_0_31> ;
    nif:anchorOf        "Roundtripping"@en ;
```

```
    nif:beginIndex      "0"^^xsd:nonNegativeInteger ;
    nif:endIndex        "13"^^xsd:nonNegativeInteger ;
    dc:identifier       "1" .

<http://freme-project.eu/#offset_25_31>
    a                   nif:Phrase , nif:OffsetBasedString , nif:String ;
    nif:referenceContext  <http://freme-project.eu/#offset_0_31> ;
    nif:anchorOf        "Dublin"@en ;
    nif:beginIndex      "25"^^xsd:nonNegativeInteger ;
    nif:endIndex        "31"^^xsd:nonNegativeInteger ;
    itsrdf:taIdentRef <http://http://dbpedia.org/resource/Dublin> .
```

Listing 8. Enriched NIF.

Finally, the e-Internationalisation service is merging the enrichments and the source document. This step is done by the utilisation of the "NIF skeleton" and the enriched content document. Note that some new lines might be lost while conversion, however, this is not relevant since these new lines does not affect the appearance of the HTML file in the browser. Listing 9 shows the final results from the round-tripping scenario - an enriched HTML document with embedded annotations.

```
<html>
    <head>
        <title>Roundtripping</title>
    </head>
    <body>
        <p>Welcome to <span its-ta-class-
ref="http://dbpedia.org/ontology/Location" its-ta-
confidence="0.9247384578843453" its-ta-ident-
ref="http://dbpedia.org/resource/Dublin">Dublin</span></p>
    </body>
</html>
```

Listing 9. Enriched HTML document.

3.1.2 Processing XML documents

For years, companies have heavily invested in adoption of XML based formats in their enterprise application workflows. Having the XML content integrated with Linked Data is of high importance since it will provided added value to the content itself. Nevertheless, companies are not willing to trade established XML workflows for the benefits of Linked Data and RDF. In FREME, we tackle this issue, and we present several options for the integration of Linked Data information into XML based formats.

In this section, we present how a user can process XML content with FREME. We below present two examples: one is on processing XML based on the DocBook 5.0 format, the second is on processing TEI

P5 documents. DocBook and TEI P5 are one of the many XML based formats heavily used in XML workflows.

DocBook is an XML based markup language defined for compiling technical documentation. The content created in the DocBook format are structured in a presentation-neutral form, thus the content can be published in variety formats, including HTML, ePub and PDF. Listing 10 shows an example of a simple DocBook document.

```
<article xmlns="http://docbook.org/ns/docbook"
xmlns:xlink="http://www.w3.org/1999/xlink" version="5.0">
    <info>
        <title>From XML to RDF step by step: Approaches for Leveraging XML
Workflows with Linked
Data</title>
    </info>
    <sect1 xml:id="s1">
        <title>Introduction</title>
        <para>We very much welcome you in the city of Prague, a home of
XML!</para>
    </sect1>
</article>
```

Listing 10. Simple DocBook 5.0 document.

In FREME, we identified several options for integrating Linked Data information into XML documents. In the following example we present one of that approaches which is embedding the information via structured markup and with help of microdata[8] and schema.org[9].

[8] https://www.w3.org/TR/microdata/

[9] http://schema.org/

```
<article xmlns="http://docbook.org/ns/docbook"
xmlns:xlink="http://www.w3.org/1999/xlink" version="5.0">
    <info>
        <title>From XML to RDF step by step: Approaches for Leveraging XML
Workflows with Linked
Data</title>
    </info>
    <sect1 xml:id="s1">
        <title>Introduction</title>
        <para>We very much welcome you in the city of <emphasis
vocab="http://schema.org/"  typeof="Place" property="name"
resource="http://dbpedia.org/resource/Prague">Prague</
emphasis>, a home of <emphasis vocab="http://schema.org/" typeof="Thing"
property="name"
resource="http://dbpedia.org/resource/XML">XML</emphasis>!</para>
    </sect1>
</article>
```

Listing 11. Linked Data in XML via structured markup.

The DocBook example from Listing 10 has been processed via the e-Entity service, and the entity enrichments have been embedded in the source XML document via structured markup. Listing 11 shows the enriched DocBook document. As shown, two entities have been identified, “Prague” and “XML”, and their related information (type and URI identifier) has been embedded in the XML document via the microdata mechanism and schema.org vocabulary.

Another relevant approach for the integration of Linked Data in XML is by anchoring Linked Data in XML attributes. In Listing 12, we show an example of the integration of Linked Data within XML attributes. In the example ITS 2.0 is used to encode the information. In particular, entity identifiers are provided via the its-ta-ident-ref attribute, and entity types via the its-ta-class-ref attribute.

The drawback of this approach is that it relies on the expressivity of the ITS 2.0 vocabulary, thus in some cases there might be a need to integrate some very specific information, e.g. population of a city, but due to the lack of such properties in ITS, the integration of this information might not be possible.

```
<article xmlns="http://docbook.org/ns/docbook" version="5.0">
    <info>
        <title>From XML to RDF step by step: Approaches for Leveraging XML
Workflows with Linked
Data</title>
    </info>
    <sect1 xml:id="s1">
        <title>Introduction</title>
        <para>We very much welcome you in the city of <span its-ta-ident-
ref="http://dbpedia.org/resource/Prague" its-ta-class-
ref="http://nerd.eurecom.fr/ontology#Location">Prague</span>, a home of <span its-
ta-ident-ref="http://dbpedia.org/resource/XML" its-ta-class-
ref="http://www.w3.org/2002/07/owl#Thing">XML</span>!</para>
    </sect1>
</article>
```

Listing 12. Linked Data enrichments in XML attributes.

In FREME, we also investigated the processing of XML documents structured according to the TEI P5 schema. TEI P5 consist of a set of guidelines for structuring XML documents. The TEI guidelines specify a descriptive encoding scheme in XML format. The Text Encoding Initiative (TEI) is a text-centric community of practice in the field of digital humanities, which is actively collaborating for more than three decades. Below we present an example of an XML document formatted according to the TEI P5 guidelines.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0"
xmlns:exist="http://exist.sourceforge.net/NS/exist" version="5.0">
    <teiHeader xmlns="http://www.tei-c.org/ns/1.0">
        <fileDesc>
            <titleStmt>
                <title>From XML to RDF step by step: Approaches for
Leveraging XML Workflows with Linked Data</title>
            </titleStmt>
            <publicationStmt>
                <p>Published for XML Prague</p>
            </publicationStmt>
            <sourceDesc>
                <p>Made by the FREME project and collaborators</p>
             </sourceDesc>
        </fileDesc>
    </teiHeader>
    <text>
        <body>
            <p>We very much welcome you in the city of Prague, a home of
XML!</p>
        </body>
    </text>
</TEI>
```

Listing 13. TEI P5 example document.

The example in Listing 13 has been enriched via FREME. In Listing 14, we show the same TEI P5 document with the integrated Linked Data information. The approaches presented here for embedding Linked Data in XML based documents have been documented in [BDP+16].

```
<TEI xmlns="http://www.tei-c.org/ns/1.0"
xmlns:exist="http://exist.sourceforge.net/NS/exist" version="5.0">
    <teiHeader xmlns="http://www.tei-c.org/ns/1.0">
        <fileDesc>
            <titleStmt>
                <title>From <span its-ta-ident-
ref="http://dbpedia.org/resource/XML" its-ta-class-
ref="http://www.w3.org/2002/07/owl#Thing">XML</span> to RDF step by step:
Approaches for Leveraging <span its-ta-ident-
ref="http://dbpedia.org/resource/XML" its-ta-class-
ref="http://www.w3.org/2002/07/owl#Thing">XML</span> Workflows with Linked
Data</title>
            </titleStmt>
            <publicationStmt>
                <p>Published for <span its-ta-ident-
ref="http://dbpedia.org/resource/XML" its-ta-class-
ref="http://www.w3.org/2002/07/owl#Thing">XML</span> <span its-ta-ident-
ref="http://dbpedia.org/resource/Prague" its-ta-class-
ref="http://nerd.eurecom.fr/ontology#Location">Prague</span></p>
            </publicationStmt>
            <sourceDesc>
                <p>Made by the FREME project and collaborators</p>
             </sourceDesc>
        </fileDesc>
    </teiHeader>
    <text>
        <body>
            <p>We very much welcome you in the city of <span its-ta-ident-
ref="http://dbpedia.org/resource/Prague" its-ta-class-
ref="http://nerd.eurecom.fr/ontology#Location">Prague</span>, a home of
<span its-ta-ident-ref="http://dbpedia.org/resource/XML" its-ta-class-
ref="http://www.w3.org/2002/07/owl#Thing">XML</span>!</p>
        </body>
    </text>
</TEI>
```

Listing 14. Integrated Linked Data enrichments in TEI P5 document.

3.2 SEMANTIC INTEROPERABILITY

The semantic interoperability addresses the need for the exchange of data with unambiguous, shared meaning. FREME addresses this challenge with help of data formats such as RDF, NIF, ITS and selected RDF ontologies.

In FREME, we have integrated the **DBpedia Ontology**, a mid-size ontology developed by the DBpedia community. The ontology is usually updated bi-annually along with the official DBpedia releases. The

latest version has been published on April 2016[10] and it consists of 754 classes, 1,103 object properties, 1,608 datatype properties, as well as 410 equivalent class assertions to several other ontologies such as FOAF, Wikidata and schema.org. The DBpedia Ontology is exclusively used for expressing fine grained entity types with the FREME e-Entity service. Currently, DBpedia is one of the core central datasets and hub for links in the Linked Open Data cloud. Many other datasets are linking and reusing DBpedia identifiers both, at the instance and at the ontology level. The DBpedia ontology is also mapped to several ontologies including schema.org, the DUL ontology[11], GeoNames[12] and LinkedGeoData ontologies. In Listing 15, we provide some of the available mappings for the http://dbpedia.org/ontology/Airport.

```
// mappings to the schema.org vocabulary
<dbo:Airport>   <owl:equivalentClass>   <http://schema.org/Airport> .

// mappings to LinkedGeoData ontology
<dbo:Airport>   <owl:equivalentClass>   <http://linkedgeodata.org/ontology/Airport> .

// mappings to Wikidata
<dbo:Airport>   <owl:equivalentClass>   <https://www.wikidata.org/wiki/Q1248784> .
```

Listing 15. Mappings between DBpedia airport class with classes from the Wikipedia and Linked GeoData.

Use case scenario. Consider a client submits the following content to the e-Entity service - "I will depart from the Charles de Gaulle Airport.". The content is processed and the airport entity Charles de Gaulle has been identified and classified with the DBpedia dbo:Airport class. Next, the user might want to retrieve a description of this type of entity in Wikidata in his/her language (e.g. German). However, the client might not have a link to the Wikidata resource which describes the airport. By looking into the mappings the client can retrieve the mapping between DBpedia class dbo:Airport and the wikidata class for the same type of entity (https://www.wikidata.org/wiki/Q1248784). Finally, the client can retrieve the localised description for this type of entity from Wikidata. Overall, by relying on DBpedia we assure a high support for a semantic interoperability and a shared meaning of enrichments exchanged within the FREME framework.

In addition, in FREME, we also exploit the **NERD Ontology**[13] (Named Entity Recognition and Disambiguation). The NERD ontology is a set of mappings established manually between the different taxonomies defining named entity types. Concepts included in the NERD ontology are collected from different schema types including the DBpedia ontology (for DBpedia Spotlight and Lupedia), lightweight taxonomies (for AlchemyAPI, Yahoo!, Wikimeta, and Zemanta) or simple flat type lists (for Extractiv, OpenCalais, Saplo, Semitags). The main benefit of relying on the NERD ontology is that it assures semantic interoperability for the entity recognition systems.

[10] The DBpedia Ontology - http://wiki.dbpedia.org/dbpedia-version-2016-04

[11] The DUL ontology - http://ontologydesignpatterns.org/wiki/Ontology:DOLCE+DnS_Ultralite

[12] The GeoNames ontology - http://www.geonames.org/ontology/documentation.html

[13] The NERD Ontology - http://nerd.eurecom.fr/ontology

Use case scenario. Consider a client submitting content for processing to several NER system. The classification schema will usually vary. For example, one system might assign the nerd:Location type to the entity, while another might assign the dbo:Location type. This might be a problem for the client since he/she cannot compare and/or aggregate the results retrieved from different systems. In order to solve this issue, a client can utilise the NERD ontology and based on provided mappings make reasonable assessment of the results.

In FREME, we integrate both, the NERD and the DBpedia ontology. While NERD ontology aims at solving semantic interoperability issues for coarse-grained entity types, the DBpedia ontology provides more fine-grained classification scheme.

FREME also exploits the **Terminology Domain Classification System** (TaaS), which defines the taxonomy for domain classification. The classification system is based and mapped to Eurovoc, the EU's multilingual thesaurus. In FREME, we use the taxonomy as basis for the definition of domains, which are further used for domain specific adaptation of FREME technologies.

In long texts, the list of recognised entities can be very large containing also entities which are not relevant to the domain of the document. Thus, there is a need for the retrieval of entities only from a specific domain. FREME enables users to filter out such irrelevant entities by specifying the domain of interest (i.e. politics and administration). The implementation of this feature is realised by populating the list of TaaS domains with corresponding entity types[14], e.g. the types dbo:PoliticalConcept and dbo:PublicService belong to the domain of politics and administration. More information on the domain specific adaptation using TaaS can be found in [SDN16].

3.3 PIPELINING

A key benefit of FREME is its ability to combine e-Services, which is enabled by a high interoperability support at syntactic and semantic levels. FREME offers a pipelining feature which allows an easy combination of two or more services in a pipelining fashion. The pipeline consists of one or more steps. Each step can take various input formats. If no format is specified, the step assumes NIF. In Listing 16, we show an example of a pipelining consisting of four services. The first step in the example pipeline evokes the e-Entity service. The second step uses the e-Link service to gather information with a selected query template. The third step calls the e-Terminology service to enrich the content with terminology related information. This needs source and target languages (English and Dutch in the example). The last step calls the e-Translation service, with the same language pairs.

```
{
    "id": 55,
    "description": "Example pipeline",
    "serializedRequests": [
    {
        "endpoint": "http://api-dev.freme-project.eu/current/e-entity/freme-
ner/documents",
        "parameters": {"language": "en"}
    } , {
        "endpoint": "http://api.freme-project.eu/current/e-link/documents/",
```

[14] See the list of domains and related entity types at https://github.com/freme-project/freme-ner/blob/master/src/main/resources/domains.csv

```
        "parameters": {"templateid": "3"}
    } , {
        "endpoint": "http://api-dev.freme-project.eu/current/e-
terminology/tilde",
        "parameters": {
        "source-lang": "en", "target-lang": "nl" }
    } , {
        "endpoint": "http://api-dev.freme-project.eu/current/e-
translation/tilde",
        "parameters": {
        "source-lang": "en", "target-lang": "nl" }
    } ]
}
```

Listing 16. Pipeline example.

The example pipeline shows several benefits. First, one can compare the outcome of several e-Services. In the example named entity recognition and terminology annotation are used to enrich the same content. This combination has the potential to improve both services via data based comparisons. Second, there is no need to hardwire the combination of services, as long as the services adhere to the linguistic linked data stack. This can be seen in the second step invoking the e-Links service. The e-Link service is installed at a different server (with the domain api.freme-project.eu) than the other e-services. The combination of services does not need a hardwired integration. Third, the pipeline and in this way the e-Services are agnostic to given input and output formats. Format coverage is realised with the e-Internationalisation service. Separating the actual services and the formats to be processed has the advantage that other services easily can be integrated and benefit from the growing set of formats being supported. Fourth, the pipelining greatly allows for automation of repetitive processes and for making the content itself intelligent. For example, a client application could analyse the content with regards to the language of content and use this information for adapting the pipeline automatically. The pipelining feature is documented in [SDN16].

4. PROVENANCE IN FREME

In order to ensure quality assessment, interoperability of the content and to allow software agents to make assessment about the quality of the enrichment we provide extensive support for provenance in FREME. Following set of requirements guided the development of provenance support in FREME.

- Identification of the annotator - unique identification of the agent who produced the annotation.
- Identification of the type of the generated information - what type of enrichment has been produced.
- Unambiguous encoding of annotations - each annotation should be uniquely encoded in the output format. In other words, each enrichment should be provided as stand-alone annotation.
- Expressive encoding of confidence information - confidence information should be able to express in different granularity. For example, specific confidence information for the entity spotting or entity classification step.

In FREME, provenance information is primarily handled in NIF along with support of the ITS 2.0. In Section 4.1 describes how provenance information is supported in NIF, while in Section 4.2 we explain how ITS supports provenance. In Section 4.2 we also provide examples on how provenance information is expressed in other content formats such as HTML and XML.

4.1 NIF 2.1

The latest NIF revision introduces additional vocabulary to express provenance and confidence information for annotations. In addition, it introduces several classes for entity, term and translation annotations.

Provenance is added by providing information on which annotator (person, service, algorithm) produced given piece of annotation information. Thus, for each annotation provided by the e-entity, e-terminology or e-translation a provenance information is available. An extended support for confidence information has been also provided which can be used to express the degree of certainty for the annotation. The provenance and confidence information can be added via two methods:

- **companion properties** - properties, specifically declared for a given property or text span annotation class.
- **generic properties** - a simpler, but more verbose representation of the provenance and confidence.

In relation to the provenance and confidence related extensions, NIF 2.1 allows expression of several alternative annotations for a same nif:String. This is enabled using a nif:AnnotationUnit resource, which can be created and linked to the annotation target (a nif:String) using the nif:annotationUnit object property. This allows unanimous tracking of provenance and confidence information across annotations from several systems in the same NIF document. For example, same content can be processed by multiple systems and a store the results in a single RDF graph. Further, the retrieved results can be fused according to some pre-defined criteria.

In the context of FREME, this enables well-defined combination of results from multiple e-services. For example, same content can be processed by a pipeline consisting of the entity, e-terminology and e-translation services and resulting in a single NIF document serialised as one RDF graph. Note that although same sequences of words can be spotted as different named entities and terms for different e-service configurations, differing individual annotation results can still be clearly associated to the e-service or e-service variant they originated from by the provenance information.

```
01 <http://freme-project.eu/#offset_0_33>
02    a nif:Context , nif:OffsetBasedString ;
03    nif:beginIndex "0"^^xsd:nonNegativeInteger ;
04    nif:endIndex "33"^^xsd:nonNegativeInteger ;
05    nif:isString "Diego Maradona is from Argentina."^^xsd:string .
06
07 <http://freme-project.eu/#offset_23_32>
08    a nif:OffsetBasedString ;
09    nif:beginIndex  "23"^^xsd:nonNegativeInteger ;
10    nif:endIndex "32"^^xsd:nonNegativeInteger ;
11    nif:anchorOf "Argentina"@en ;
12    nif:referenceContext <http://freme-project.eu/#offset_0_33> ;
13    nif:annotationUnit [
14       a nif:TermOccurrence ;
15       itsrdf:term "yes" ;
16       rdfs:label "Argentina"@en ;
17       itsrdf:taAnnotatorsRef  <https://services.tilde.com/terminology> ;
18       itsrdf:taConfidence  "1"^^xsd:double
19    ] ;
20    nif:annotationUnit [
21       a nif:EntityOccurrence ;
22       itsrdf:taIdentRef      <http://dbpedia.org/resource/Argentina> ;
23       itsrdf:taAnnotatorsRef  <http://freme-project.eu/tools/freme-ner> ;
24       itsrdf:taConfidence     "0.9804963628413852"^^xsd:double
25    ] .
```

Listing 17. Fused results from e-entity and e-terminology.

Listing 17 provides an example of a result of a pipeline of two e-services. E-terminology spotted "Argentina" as a term. Its annotations are grouped within the nif:AnnotationUnit at line 13-19. The itsrdf:taAnnotatorsRef assertion at line 17 informs consumers of the data that these pieces of annotation information originate from e-terminology[15]. In the same vein, entity spotting and linking results for the same token from e-Entity are grouped in lines 20 to 25 and can be attributed to that service easily. Both results also provide confidence estimates for their annotation decisions (lines 18 and 24).

A further addition in the NIF 2.1 release allows to formulate NIF documents that express annotation for given primary text data without containing the primary data itself (stand-off) annotation. In NIF 2.0 each nif:Context instance was required to carry a data property assertions via nif:isString that would contain the textual context annotated. However, circumstances are conceivable under which re-distribution or

[15] The provenance identifiers in this example are coarse grained. Linking to specific Linked Data resources that specify which service release was used and which parameters (e.g. used dataset, requested domains/types for named entities) is equally possible. Using the PROV-O vocabulary, information can either be provided just pertaining to a service/service configuration on it's own (prov:Agent) or combined with information about the time of the service invocation and completion of the request (prov:Activity).

publication of primary data is not possible, e.g. due to license restrictions or confidentiality/privacy concerns. If annotations for this data can and should still be provided in an interoperable manner as Linked Data, these annotations can now be expressed as a NIF 2.1 documents with its nif:Context instances utilising the nif:contextStringRef property to specify a web resource to obtain the string data their annotations pertain to. This web resource can then be subject to different licensing and access restrictions, but it's payload is subject to the same requirements that also applies to the string literals for nif:isString to ensure well-defined, unequivocal semantics for the string index coordinates of index based nif:String variants (e.g. nif:OffsetBasedString).

To ensure that results from FREME e-services are interoperable and composable with precedent NIF data generated earlier, NIF 2.1 is fully backward-compatible with NIF 2.0. Nonetheless, minor aspects of the recommended practice were revised and one scheme for nif:Strings, nif:RFC5147String, was deprecated, among other reasons due to lack of specificity for index semantics in the underlying RFC. All other minor changes on OWL restriction in the 2.1 revision are compatible with NIF data according to the 2.0 revision under OWL satisfiability, i.e. NIF 2.0 documents will still be consistent according to OWL2 reasoning using the NIF 2.1 ontology.

NIF 2.1 incorporates the community feedback and cater some specific requirements from FREME and other ongoing projects.

4.2 ITS 2.0

ITS 2.0 specification defines data categories which can be used to integrate automated processing of human language into core Web technologies. Its main focus is HTML and XML based formats, but it can be also leveraged in combination with NIF.

Some of the core concepts for expressing provenance information in FREME are based on ITS. In FREME, we implement following ITS properties:

- itsrdf:taAnnotatorsRef - the attribute provides a way to associate all generated information within a given nif:annotationUnit with information about the agent that generated the information. This property is extensively used across all FREME services.
- itsrdf:taIdentRef - a property used to provide an identifier for the entity annotation.
- itsrdf:termInfoRef - a property used to provide an identifier to a term annotation.
- itsrdf:taConfidence - a property holding the confidence of the agent (that produced the annotation) in its own computation.
- itsrdf:termConfidence - the confidence of the agents producing the annotation that the annotated unit is a term or not.
- itsrdf:mtConfidence - translation confidence score as a rational number in the interval 0 to 1 (inclusive).
- Itsrdf:term - a property with the value "yes" or "no" indicating whether or not is the annotation a term.
- Itsrdf:target - a property that holds the translation.

4.2.1 Provenance information in HTML

Apart of providing provenance information in NIF, FREME can also encode provenance and confidence information in HTML. This is supported via the ITS 2.0 properties which are used to encode this

information. In the following listing we provide an example on provenance/confidence information encoded in HTML via FREME.

```
<html its-ta-annotators-ref="https://services.tilde.com/terminology">
    <head>
        <title>dummy title</title>
    </head>
    <body>
        <p id="n7"><span class="test1"></span>Welcome to the city of <span
class="test2"></span><span its-term-info-ref="http://freme-
project.eu/#offset_10_16" its-term="yes">Berlin</span>!</p>
    </body>
</html>
```

Listing 18. Provenance information encoded in HTML via FREME e-Terminology.

The source HTML document that was submitted to the FREME e-Terminology service did not include any annotations or enrichments. The source document was processed by the e-Terminology service, enrichments were generated and further via the e-Internationalisation service they have been incorporated in the source HTML document.

4.2.2 Processing XML and handling provenance

FREME can also process XML and encode the provenance information, if any present. In this section we present a scenario where an XML based document based on the XLIFF 2.0 specification is first converted to HTML. Then, using the e-Entity service the HTML document is processed and enrichment information is embedded in the HTML document. Finally, the HTML document is converted back to XLIF. Listing 19 shows the input XLIFF 2.0 document.

```
<xliff xmlns='urn:oasis:names:tc:xliff:document:2.0' version='2.0'  srcLang='en'
trgLang='fr'>
    <file id='f1'>
        <unit id='u1'>
            <segment id='s1'>
        <source>See you in the city of Prague!</source>
    </segment>
        </unit>
    </file>
</xliff>
```

Listing 19. Simple XLIFF 2.0 document.

The XLIFF document is first converted to HTML using a dedicated XSLT template, which performs conversion of XLIFF 2.0 to HTML, for further processing with e-Services. The FREME framework provides a management API for the manipulation with XSLT conversion templates. A user can create, retrieve information about a specific template, update, remove or execute a template.

Upon the execution of the template[16], the XLIFF document is converted to HTML and the results are shown in the following listing.

```
<html xmlns="http://www.w3.org/1999/xhtml">
    <head>
        <title>@@@</title>
        <script type="application/xml">
            <xliff xmlns="urn:oasis:names:tc:xliff:document:2.0"
                version="2.0"
                srcLang="en"
                trgLang="fr">
                <file id="f1">
                    <unit id="u1">
                        <segment id="s1">
                            <anchor xmlns="http://www.w3.org/1999/xhtml"
id="n1"/>
                        </segment>
                    </unit>
                </file>
            </xliff>
        </script>
    </head>
    <body>
        <div id="xyz1xyz">
            <p id="n1">See you in the city of Prague!</p>
        </div>
    </body>
</html>
```

Listing 20. Results from the XLIFF to HTML conversion.

Next, the HTML is being processed with the e-Entity service and the enrichments are embedded into the source HTML document. The results from this step are presented in Listing 21.

```
<html>
    <head>
        <title>@@@</title>
        <script type="application/xml">
            <xliff xmlns="urn:oasis:names:tc:xliff:document:2.0"
                version="2.0"
                srcLang="en"
                trgLang="fr">
                <file id="f1">
                    <unit id="u1">
                        <segment id="s1">
                            <anchor xmlns="http://www.w3.org/1999/xhtml"
id="n1"/>
                        </segment>
```

[16] The conversion template used in the example has the ID “xliff20-to-html” and it can be retrieved at: https://api.freme-project.eu/current/toolbox/xslt-converter/manage/xliff20-to-html

```
                    </unit>
                </file>
            </xliff>
        </script>
    </head>
    <body>
        <div id="xyz1xyz">
            <p id="n1">See you in the city of
                <span data-its-ta-class-refs="http://dbpedia.org/ontology/City
http://dbpedia.org/ontology/Settlement http://dbpedia.org/ontology/PopulatedPlace
http://dbpedia.org/ontology/Place http://nerd.eurecom.fr/ontology#Location
http://dbpedia.org/ontology/Location" its-ta-class-
ref="http://dbpedia.org/ontology/City" its-ta-confidence="0.8277752999056328"
its-ta-ident-ref="http://dbpedia.org/resource/Prague">Prague</span>!
            </p>
        </div>
    </body>
</html>
```

Listing 21. Enriched HTML document converted from XLIFF.

Finally, we can convert the enriched HTML document back to the XLIFF content using an XSLT template[17] for an HTML to XLIFF conversion. The results are presented in Listing 22.

```
<?xml version="1.0" encoding="UTF-8"?>
<xliff xmlns="urn:oasis:names:tc:xliff:document:2.0"
       version="2.0"
       srcLang="en"
       trgLang="fr">
    <file id="f1">
        <unit id="u1">
            <segment id="s1">
                <source xmlns:itsm="urn:oasis:names:tc:xliff:itsm:2.1">See you in
the city of
                    <mrk id="m1"
                    type="itsm:generic"
itsm:taIdentRef="http://dbpedia.org/resource/Prague">Prague</mrk>!
                </source>
            </segment>
        </unit>
    </file>
</xliff>
```

Listing 22. Results from the HTML to XLIFF conversion.

[17] The XSLT template use for conversion of HTML to XLIFF has the ID “html-to-xliff20” and it can be retrieved at: https://api.freme-project.eu/current/toolbox/xslt-converter/manage/html-to-xliff20

5. CONCLUSION

This deliverable summarises the work done in WP1 Task 1.2 Data innovation and take up in FREME (M4-M21). It provides details on how the interoperability at syntactic and semantic levels is ensured in FREME and how the provenance of information is supported. In order to support the interoperability and lossless conversion between different formats (e.g., HTML, XLIFF and XML), technologies such as NIF, ITS, URI and RDF, have been adopted within the FREME framework. In addition, NIF has been extended to ensure a lossless tracking of provenance across formats and systems.

REFERENCES

[SDN16] F. Sasaki and M. Dojchinovski and Jan Nehring: Chainable and Extendable Knowledge Integration Web Services (to appear), First Workshop on Knowledge Extraction and Knowledge Integration - ISWC 2016, Kobe, Japan, October 2016.

[HLA+13] Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer: Integrating NLP using Linked Data. 12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia, 2013.

[BDP+16] Marta Borriello, Christian Dirschl, Axel Polleres, Phil Ritchie, Frank Salliau and Felix Sasaki and Giannis Stoitsis: From XML to RDF step by step: approaches for leveraging XML workflows with linked data, XML Prague Conference, 11-13 February 2016, Prague, Czech Republic, 2016.