



Co-funded by the Horizon 2020  
Framework Programme of the European Union  
Grant Agreement Number 644771



# FROM XML TO RDF STEP BY STEP: APPROACHES FOR LEVERAGING XML WORKFLOWS WITH LINKED DATA

XML PRAGUE | 12 FEBRUARY 2016

[www.freme-project.eu](http://www.freme-project.eu)



Felix Sasaki, DFKI / W3C Fellow

On behalf of the FREME Consortium and Contributors

# THE CO-AUTHORS OF THIS EFFORT AND PAPER

- Marta Borriello, Vistatec
- Christian Dirschl, Wolters Kluwer
- Axel Polleres, Vienna University of Economics and Business (WU)
- Phil Ritchie, Vistatec
- Frank Salliau, iMinds
- Felix Sasaki, DFKI / W3C Fellow
- Giannis Stoitsis, Agro-Know

# MOTIVATION – THIS BREAKS XML PROCESSING!

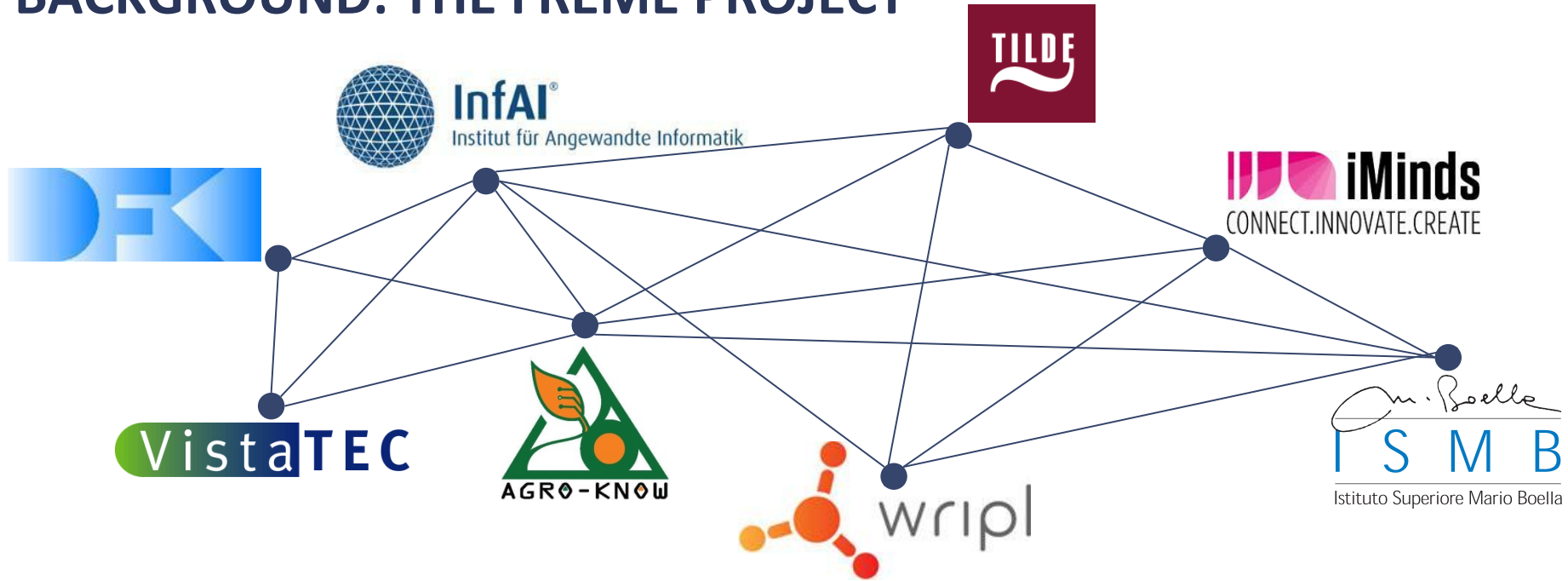
```
<myData> <head>...</head> <body>  
<linkedDataStorage>...</linkedDataStorage> ... </body> </myData>
```

- Validation
- Transformation
- Query
- ...
- Adaptation of schemas in real life scenarios often not possible

# IS THIS RDF CHIMERA AGAIN?

- No: RDF Chimera is about relation between formats
  - XML, HTML RDF, JSON
- Our Issue here is about integration of formats
  - RDF in XML workflows for multilingual and semantic enrichment of content

# BACKGROUND: THE FREME PROJECT



- Two year H2020 Innovation action; start February 2020
- Industry partners leading four business cases around digital content and (linked) data
- FREME = A framework for multilingual and semantic enrichment of digital content
- Is there a real need for this? Oh yes! See the following business cases

# BUSINESS CASE “LINKED DATA IN PUBLISHING WORKFLOWS”

- Wolters Kluwer, Agroknow
- Enrichment of academic publication metadata

Before FREME	Result of deploying FREME
<pre>&lt;dc:creator&gt; &lt;ags:creatorPersonal&gt; Stoitsis, Giannis, Agroknow &lt;/ags:creatorPersonal&gt; &lt;/dc:creator&gt;</pre>	<pre>&lt;dc:creator&gt; &lt;ags:creatorPersonal&gt;Stoitsis, Giannis&lt;/ags:creatorPersonal&gt; &lt;nameIdentifier schemeURI= "http://orcid.org/" nameIdentifierScheme= "ORCID"&gt;0000-0003-3347-8265 &lt;/nameIdentifier&gt; &lt;affiliation&gt;Agroknow&lt;/affiliation&gt; &lt;/dc:creator&gt;</pre>
<pre>&lt;dc:subject&gt; &lt;ags:subjectClassification scheme="ags:ASC"&gt; &lt;![CDATA[J10]]&gt; &lt;/ags:subjectClassification&gt; &lt;/dc:subject&gt;</pre>	<pre>&lt;dc:subject freme-enrichment= "http://aims.fao.org/aos/agrovoc/c_426 http://aims.fao.org/aos/agrovoc/c_24135 http://aims.fao.org/aos/agrovoc/c_4644 http://aims.fao.org/aos/agrovoc/c_7178"&gt; &lt;ags:subjectClassification scheme= "ags:ASC"&gt;&lt;![CDATA[J10]]&gt; &lt;/ags:subjectClassification&gt; &lt;/dc:subject&gt;</pre>

## BUSINESS CASE

### “LINKED DATA IN XML LOCALIZATION WORKFLOWS”

- Vistatec – workflows integrating localization XML formats XLIFF, ITS 2.0 and linked data, in the Ocelot editor for translation editing and review – see GUI screenshot next slide

<b>Process Step</b>	<b>FREME e-service</b>
Conversion of native document to Extensible Localization Interchange File Format	e-Internationalization
Translation	e-Terminology and e-Entity
Semantic enrichment	e-Link
Content publication	e-Pub

Ocelot - fremexliff

File View Filter Segment Extensions Help

Doc Stats LQI Prov Other ITS

Data Category	Type	Value	Count
LQI	mstranstabon	10.0-10.0	1

Translations Concordance Search

Translation Results

&lt;p>Bray is a great place to base yourself while you are in Wicklow as many of the city's most popular attractions are within easy driving distance. The Bray Heritage Centre and the National Sealife Centre are two of the interesting place to visit here. Also, be sure to check out Bray Beach and the Killruddery House and Gardens.&lt;br>&lt;br>	100%	Bray > &lt;p est l'endroit parfait pour vous, lorsque vous avez fini de la ville de l'église auant d' abouts les plus populaires se trouvent à une distance. Le Centre National Heritage Bray sealle et le centre de l' une ou l' autre des deux intéressant qu' auparavant. De même, n' oubliez pas de retirer la plage et Bray Killruddery Gardens.&lt;br /> > > &lt;	FREME e-Translation
--	------	---	---------------------

Label Segments

Source: &lt;p>Bray is a great place to base yourself while you are in Wicklow as many of the city's most popular attractions are within easy driving distance. The Bray Heritage Centre and the National Sealife Centre are two of the interesting place to visit here. Also, be sure to check out Bray Beach and the Killruddery House and Gardens.&lt;br>&lt;br>

Target: &lt;p>Bray is a great place to base yourself while you are in Wicklow as many of the city's most popular attractions are within easy driving distance. The Bray Heritage Centre and the National Sealife Centre are two of the interesting place to visit here. Also, be sure to check out Bray Beach and the Killruddery House and Gardens.&lt;br>&lt;br>

Original Target:

Edit Distance: 0

Enrichment

&lt;p>Bray is a great place to base yourself while you are in Wicklow as many of the city's most popular attractions are within easy driving distance. The Bray Heritage Centre and the National Sealife Centre are two of the interesting place to visit here. Also, be sure to check out Bray Beach and the Killruddery House and Gardens.&lt;br>&lt;br>

http://wikipedia.org/resource/Bray

View info about Bray

e-Link Service Results

Bray

Abstract Info Image Links

Bray (Irish: Bré, meaning "hill", formerly Bri Chualann) is a town in north County Wicklow, Ireland. It is a busy urban centre and seaside resort, with a population of 31,872 making it the ninth largest urban area in Ireland at the 2011 census. It is situated about 20 km (12 mi) south of Dublin on the east coast. The town straddles the Dublin-Wicklow border, with a portion of the northern suburbs situated in County Dublin. Bray's scenic location and proximity to Dublin make it a popular destination for tourists and day-trippers from the capital. Bray is home to Ireland's only film studios, Ardmore Studios, hosting Irish and international productions for film, television and advertising. Some light industry is located in the town, with business and retail parks concentrated largely on its southern periphery. Bray town centre has a range of shops serving the consumer needs of the surrounding area. Commuter links between Bray and Dublin are provided by rail, Dublin Bus and the M11 and M50 motorways.

Close

SPARQL queries are executed to retrieve desired related information



# BUSINESS CASE “LINKED DATA IN BOOK METADATA”

- iMinds – linked data in book metadata
- A potential approach for embedding linked data in ONIX

```
<Contributor>
  <NameIdentifier>
    <NameIDType>
      <IDTypeName>Meta4Books ContributorID</IDT
      <IDValue>65097</IDValue>
    </NameIDType>
  </NameIdentifier>
  <ContributorRole>A01</ContributorRole>
  <SequenceNumber>1</SequenceNumber>
  <NamesBeforeKey>Jonathan</NamesBeforeKey>
  <KeyNames>Franzen</KeyNames>
  <Entity>
    <URI>http://viaf.org/viaf/84489381/</URI>
  </Entity>
</Contributor>
```

# APPROACHES FOR LINKED DATA INTEGRATION

1. Convert XML to linked data
2. Embed linked data into XML via structured markup
3. Anchor Linked data in XML attributes
4. Embed linked data in metadata sections of XML files
5. Anchor linked data via annotations in XML content

Try them out with DocBook or TEI content at

<http://api-dev.freme-project.eu/doc/freme-showcase/xml-to-rdf.html>

Implementation uses FREME, the Okapi framework and Saxon-CE, the Swiss army knife of XML in the browser processing

# SCREENSHOT FROM DEMO

Call FREME e-Entity with XML content

Refresh output

Set the document type of the input: DocBook

Set the language of the input: English

Set the data set to be used for enrichment: DBpedia

Set the output type: 1. Convert XML to linked data

```
<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>
```

# 1. CONVERT XML TO LINKED DATA

```
<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>
```

# 1. CONVERT XML TO LINKED DATA

```
<http://freme-project.eu/#char=140,146>
  a          nif:Phrase , nif:RFC5147String , nif:String , nif:Word ;
  nif:anchorOf    "Prague"^^xsd:string ;
  nif:beginIndex  "140"^^xsd:int ;
  nif:endIndex    "146"^^xsd:int ;
  nif:referenceContext <http://freme-project.eu/#char=0,162> ;
  itsrdf:taClassRef  <http://dbpedia.org/ontology/City> ,
<http://dbpedia.org/ontology/Location> , <http://dbpedia.org/ontology
/Settlement> , <http://nerd.eurecom.fr/ontology#Location> ,
<http://dbpedia.org/ontology/PopulatedPlace> , <http://dbpedia.org/ontology
/Place> ;
  itsrdf:taConfidence "0.9990763283024261"^^xsd:double ;
  itsrdf:talentRef   dbpedia:Prague .
```

# 1. CONVERT XML TO LINKED DATA

## Benefits

- No need to change XML workflow
- Similar to RDF Chimera approach
- Difference: here focus on adding new (linked) information

## Drawback

- New tool chain needed
- No useful representation of mixed content

## 2. EMBED LINKED DATA INTO XML VIA STRUCTURED MARKUP

```
<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>
```

## 2. EMBED LINKED DATA INTO XML VIA STRUCTURED MARKUP

### Benefits

- Relying on hooks for data integration, e.g. RDFa 1.1 lite
- Common for search engine optimization, cf. schema.org
- May use other syntaxes like json-ld

### Drawback

- May break XML validation
- May need at least adapted tool chains to understand RDFa / json-ld



### 3. ANCHOR LINKED DATA IN XML ATTRIBUTES

Example: Embedding anchors in XLIFF via ITS 2.0 text analytics markup

```
<source ...>
```

```
<mrk ...its:talentRef="http://dbpedia.org/resource/Berlin">
```

```
Berlin</mrk> is the capital of Germany!</source>
```

### 3. ANCHOR LINKED DATA IN XML ATTRIBUTES

#### Benefits

- Using existing XML attributes = no new markup is needed
- Toolchain can be kept as is

#### Drawback

- Actual data integration is just postponed
- Data integration does not leave a trace – missing provenance

## 4. EMBED LINKED DATA IN METADATA SECTIONS OF XML FILES

```
<article xmlns="http://docbook.org/ns/docbook" ...>
<info>
<title>From XML to RDF step by step: Approaches for Leveraging XML
Workflows with Linked
Data</title><annotation><programlisting>@prefix dbpedia-fr:
&lt;http://fr.dbpedia.org/resource/> .
...
&lt;http://freme-project.eu/#char=0,86>
  a          nif:Phrase , nif:RFC5147String , nif:String ;
  nif:anchorOf      "From XML to RDF step by step: Approaches for
Leveraging XML Workflows with Linked Data"@en ;
  nif:beginIndex    "0"^^xsd:nonNegativeInteger ;
  nif:endIndex      "86"^^xsd:nonNegativeInteger ;
  nif:referenceContext &lt;http://freme-project.eu/#char=0,162> ;
  &lt;http://purl.org/dc/elements/1.1/identifier>
```

## 4. EMBED LINKED DATA IN METADATA SECTIONS OF XML FILES

### Benefits

- Metadata section does not influence size of main content
- Clear separation of concerns and processing

### Drawback

- No per se relation to actual content
- Character offset pointers to content are fragile

## 5. ANCHOR LINKED DATA VIA ANNOTATIONS IN XML CONTENT – HERE USING W3C ANNOTATION MODEL

```
{ "id": "http://example.com/myannotations/a1", "type":  
"Annotation" ...
```

```
"selector": { ...
```

```
"/xlf:unit[1]/xlf:segment[1]/xlf:source/xlf:mrk[1]"
```

```
"itsrdf:talentRef": "http://dbpedia.org/resource/Berlin",
```

```
"itsrdf:taClassRef": "http://schema.org/Place", ... } }
```

## 5. ANCHOR LINKED DATA VIA ANNOTATIONS IN XML CONTENT

### Benefits

- Same as approach 4
- In addition, more robust anchoring via path expressions

### Drawback

- Resolution of path expressions can be computationally expensive

# APPROACHES FOR LINKED DATA INTEGRATION ...

1. Convert XML to linked data
2. Embed linked data into XML via structured markup
3. Anchor Linked data in XML attributes
4. Embed linked data in metadata sections of XML files
5. Anchor linked data via annotations in XML content

**... Or: Routes to Bridge between RDF and XML**

# ROUTES TO BRIDGE BETWEEN RDF AND XML

- XSPARQL – W3C Member submission
- Compilation of SPARQL queries into XQuery

```
prefix foaf: <http://xmlns.com/foaf/0.1/>
prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>

<kml xmlns="http://www.opengis.net/kml/2.2">{
  for $name $long $lat from <http://nunolopes.org/foaf.rdf>
  where { $person a foaf:Person; foaf:name $name;
          foaf:based_near [ a geo:Point;
                             geo:long $long;
                             geo:lat $lat ] }
  return <Placemark>
    <name>{fn:concat("Location of ", $name)}</name>
    <Point>
      <coordinates>{fn:concat($long, ",", $lat, ",0")}</coordinates>
    </Point>
  </Placemark> }
</kml>
```



## NEXT STEPS – WHAT DO YOU THINK?

- Conclusion with my co-authors:

“We believe that joint efforts in standardization bodies to bridge the gaps between RDF and XML in order to enable such transformations and integrated tooling in a standard way should be further pursued.”

- What do you think – is it worth documenting this further
  - Have a W3C community group on the topic?
  - Document approaches and best practices?
  - Provide of-the-shelf tooling?

# CONTACTS

FELIX SASAKI

Senior Researcher DFKI / W3C Fellow

On behalf of the FREME consortium and collaborators

E-mail: [felix.sasaki@dfki.de](mailto:felix.sasaki@dfki.de)

