TELIX

an RDF-based model for linguistic annotation

Emilio Rubiera[1], Luis Polo[1], Diego Berrueta[1], Adil El Ghali[2]
(1) Fundación CTIC, Spain;
(2) IBM France

what's TELIX?

linguistic knowledge and text annotations as RDF graphs

some features:

- multi-layered annotations
- syntax trees and feature structures
- ...

Outline

1. linguistic annotations & RDF
2. introducing TELIX
3. annotating text with RDFa
4. related work and conclusions

what are text annotations?

- texts are marked with **annotations** to improve shared understanding

picture from sismel.it

text annotations nowadays

- annotations are currently produced not any longer by monks but by software (NLP)
  - segmentation, word disambiguation, syntactic and discourse analysis, etc.

- ... but there are still problems to solve:
  - integration of annotation layers
  - interoperability between tools

RDF & text annotations

- RDF can integrate multiple annotations layers
- RDF annotations can be freely organised as needed
- provenance information can be encoded as named graphs
- URIs make it possible to link linguistic resources across multiple RDF graphs
- leverage the web of data for annotations (using DBpedia for concept disambiguation)

text segments as RDF resources

- reusing Dublin Core for corpora description
  - dctype:Text and dctype:Collection
- segment (inspired in LAF), a fragment of choice in a NL production
  - Text, audio, etc.
- textual units
  - Token, Word, Sentence, Paragraph, Section

example: segments as resources

"This factory produces superb maraging steel"

```
ex:t1 a telix:Token ; telix:value "maraging"@en.
ex:t2 a telix:Token ; telix:value "steel"@en.
ex:s1 a telix:Sentence ; telix:value "This..."@en.
```

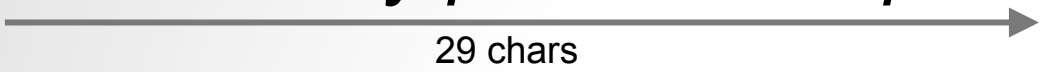
example: related segments

"This factory produces superb maraging steel"

```
ex:t1 telix:precedes ex:t2 .
ex:s1 telix:contains ex:t1 .
```

example: optional segment location

"This factory produces superb maraging steel"




```
ex:t1 telix:offset "29"^^xsd:int .
```

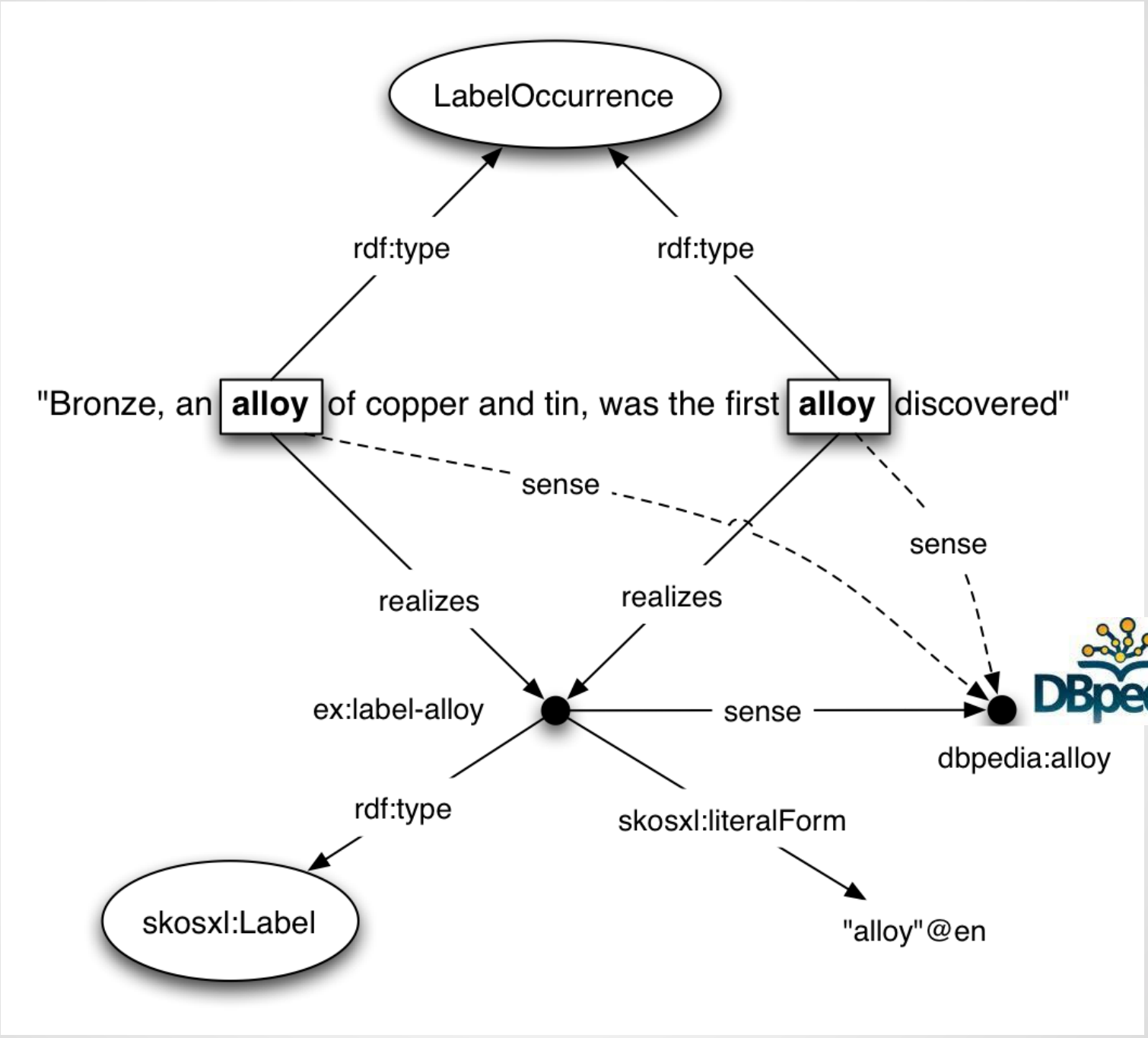
words and senses in RDF

*"Bronze, an [**alloy**] of copper and tin, was the first [**alloy**] discovered"*

linguistic entity	**OWL class**	**Example**	**language dependence**	**text dependence**
Concept	skos:Concept	dbpedia:Alloy	-	-
Lexeme	skosxl:Label	Alloy: “a metal made by combining two…” (just once)	+	-
Word (occurrence)	telix:Label Occurrence	"alloy"-1 "alloy"-2	+	+

from SKOS to TELIX

- In SKOS, words are just literals
  - ...but RDF literals cannot be subjects, i.e., ~~"hound"@en ex:synonym "dog"@en .~~
- W3C SKOS-XL reifies lexical entities as RDF resources (*skosxl:Label*)
- TELIX fixes the interpretation of skosxl:Label as lexemes
- TELIX introduces specific properties to capture lexical relations
  - Synonyms, homonyms, hypernyms and hyponyms ex:hound telix:synonym ex:dog .


LabelOccurrence
rdf:type
rdf:type
"Bronze, an alloy of copper and tin, was the first alloy discovered"
sense
sense
realizes
realizes
ex:label-alloy
sense
dbpedia:alloy
rdf:type
skosxl:literalForm
skosxl:Label
"alloy"@en

feature structures

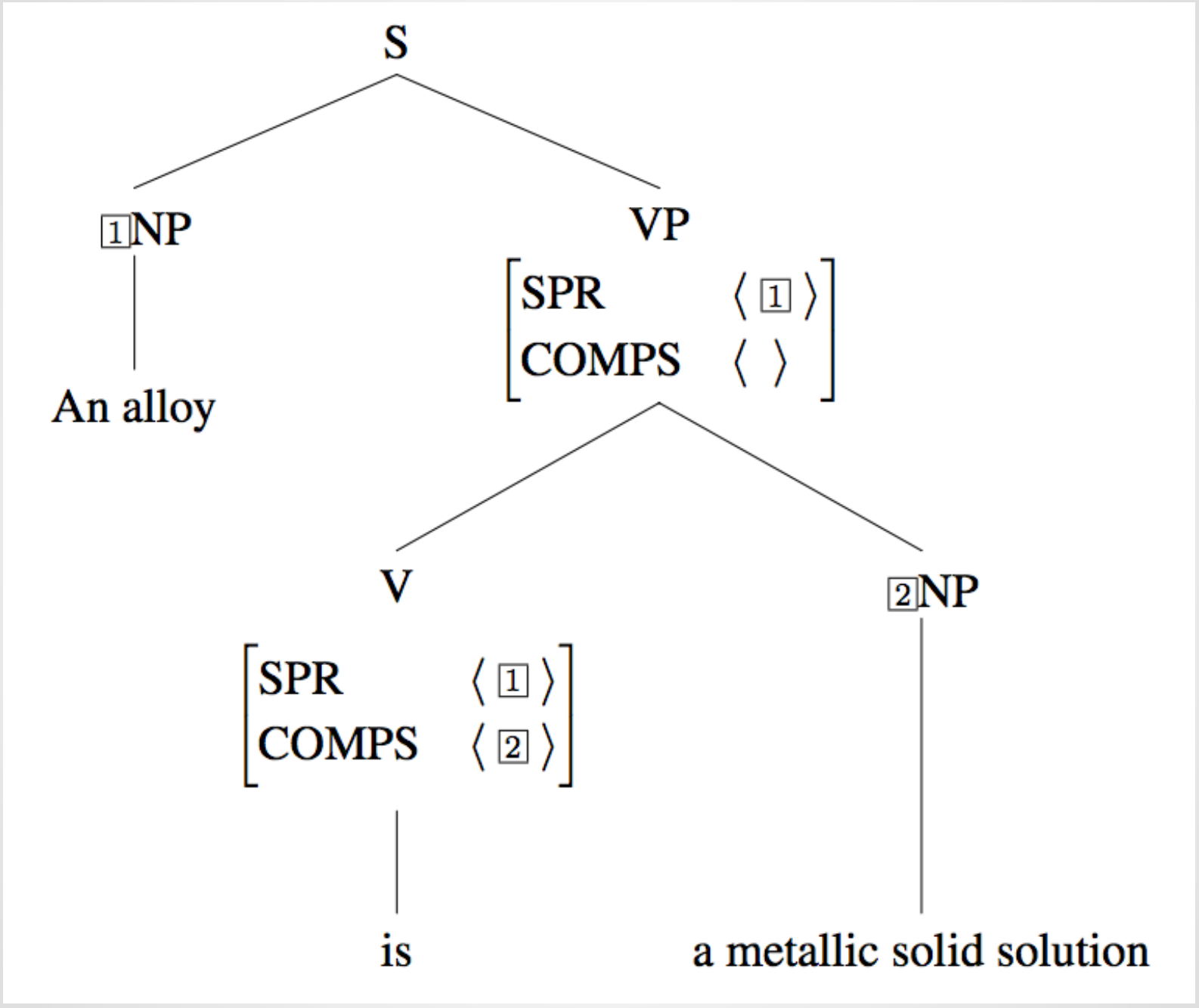
FS are pair-value recursive structures that describe linguistic information

$$\begin{bmatrix} \textit{word} & \\ \text{POS} & \text{noun} \\ \text{AGR} & \begin{bmatrix} \text{GEND} & \text{masc} \\ \text{NUM} & \text{sg} \end{bmatrix} \end{bmatrix}$$

(a) AVM representation

```
[ rdf:type telix:LabelOccurrence ;
  telix:value "alloy"@en ;
  telix:pos telix:NNS ;
  telix:agreement [ telix:number telix:Singular ;
                    telix:gender telix:Masculine ] ] .
```

(b) RDF-based TELIX representation.


S
1 NP
VP
An alloy
SPR ⟨ 1 ⟩
COMPS ⟨ ⟩
V
2 NP
SPR ⟨ 1 ⟩
COMPS ⟨ 2 ⟩
is
a metallic solid solution

annotation support

annotations are reified as RDF resources so:

- they can be described (authorship, source...)
- they can be amalgamated (multi-layered)
- multiple (alternative) annotations are possible over the same text

annotations can be partitioned in named graphs

TELIX and RDFa: simple example

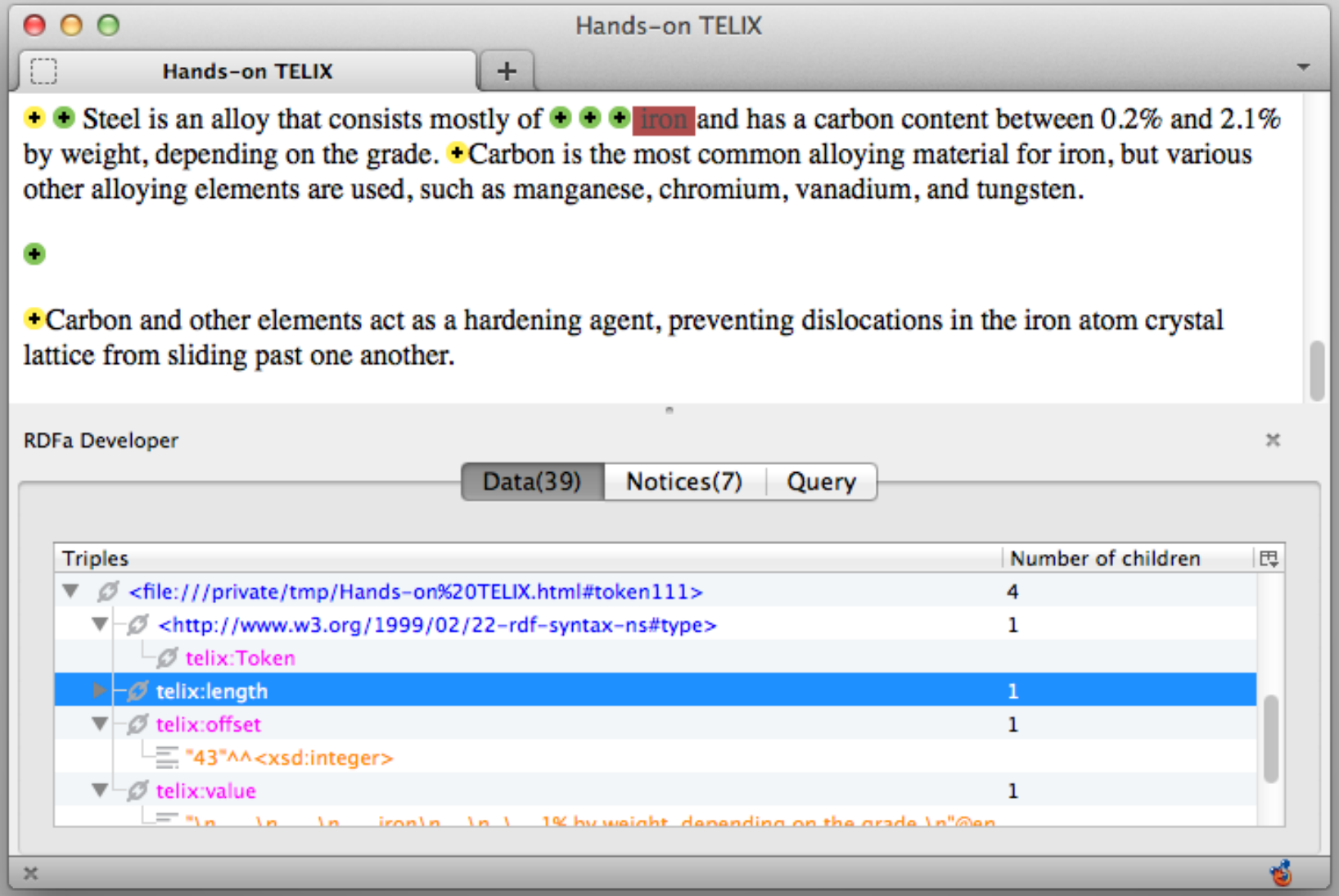
Carbon and other elements act as a hardening agent, preventing dislocations in the iron atom crystal lattice from sliding past one another.

TELIX and RDFa: simple example

```
<span about="#sent21"
typeof="telix:Sentence"  datatype="xsd:string"
property="telix:value"  id="sent21">
    Carbon and other elements act as a
    hardening agent, preventing dislocations in
    the iron atom crystal lattice from sliding past
    one another.
</span>
```

TELIX and RDFa: complex example

```
<div about="#document1" typeof="dctype:Text" datatype="" property="telix:value"
      rel="dct:hasPart" id="document1">
<p about="#para1" typeof="telix:Paragraph" datatype="" property="telix:value"
      rel="dct:hasPart" id="para1">
        <span about="#sent11" typeof="telix:Sentence" datatype="" property="telix:value"
              id="sent11">
          <span property="telix:position" content="1" datatype="xsd:integer" />
          Steel is an alloy that consists mostly of
          <span rel="dct:hasPart">
            <span about="#token111" typeof="telix:Token" datatype="" property="telix:value" id="lo111">
              <span property="telix:offset" content="43" datatype="xsd:integer" />
              <span property="telix:length" content="4" datatype="xsd:integer" />
              iron
            </span>
          </span>
          and has a carbon content between 0.2% and 2.1% by weight, depending on the grade.
        </span>
    <span about="#sent12" typeof="telix:Sentence" datatype="" property="telix:value"
  id="sent12">Carbon is the most common alloying material for iron, but various
  other alloying elements are used, such as manganese, chromium, vanadium, and
  tungsten.</span>
</p>
<p about="#para2" typeof="telix:Paragraph" datatype="" property="telix:value"
      rel="dct:hasPart" id="para2">
  <span about="#sent21" typeof="telix:Sentence" datatype="" property="telix:value"
        id="sent21">Carbon and other elements act as a hardening agent, preventing
  dislocations in the iron atom crystal lattice from sliding past one another.</span>
</p>
</div>
```


Hands-on TELIX
Hands-on TELIX
Steel is an alloy that consists mostly of iron and has a carbon content between 0.2% and 2.1% by weight, depending on the grade. Carbon is the most common alloying material for iron, but various other alloying elements are used, such as manganese, chromium, vanadium, and tungsten.
Carbon and other elements act as a hardening agent, preventing dislocations in the iron atom crystal lattice from sliding past one another.
RDFa Developer
Data(39)
Notices(7)
Query
Triples
Number of children
<file:///private/tmp/Hands-on%20TELIX.html#token111>
4
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
1
telix:Token
telix:length
1
telix:offset
1
"43"^^<xsd:integer>
telix:value
1

related work (I): annotation frameworks

- TEI, PAULA: XML-based
- LAF: ancestor of TELIX, we add multi-layered ann.
- GrAF: ad hoc XML syntax for graphs, not RDF
- LMF: XML-based, lexical-focused, limited for syntactic ann.
- NIF: same principles (RDF, URIs), different underlying ontologies (OLiA)

related work (II): linguistic models

(all of them available in OWL)

- GOLD: design to describe natural languages in general, not designed for annotations
- OLiA: linguistic knowledge distributed in a number of ontologies
- Lemon model: only focused on concepts/words, multi-linguism not SKOS-based
- WordNet: only focused on concepts/words, not SKOS-based
- POWLA: stay tuned for the next presentation

ongoing work

- W3C member submission
- participation in the W3C Group Ontology-Lexica.
- extending TELIX to capture other linguistic materializations

last words

experience it yourself!

- the OWL ontology:

  http://purl.org/telix/

- the full technical report:

  http://ontorule-project.eu/deliverables-deliverable/d14.html

thank you for your attention

Emilio Rubiera, Luis Polo, Diego Berrueta, Adil El Ghali
emilio.rubiera@fundacionctic.org

CTIC Centro Tecnológico

ONTORULE
Ontologies meet Business Rules