Session 2:
The NLP Interchange Format

Dr. Milan Dojchinovski
http://dojcinovski.mk

CTU in Prague, Czech Republic
Institute for Applied Informatics (AKSW), Germany

Nexus Linguarum

Outline

1. Motivation
2. NIF in the Nutshell
3. NIF aware Web Services
4. Exercises
5. Q&A

Motivation

The “Developers Nightmare”

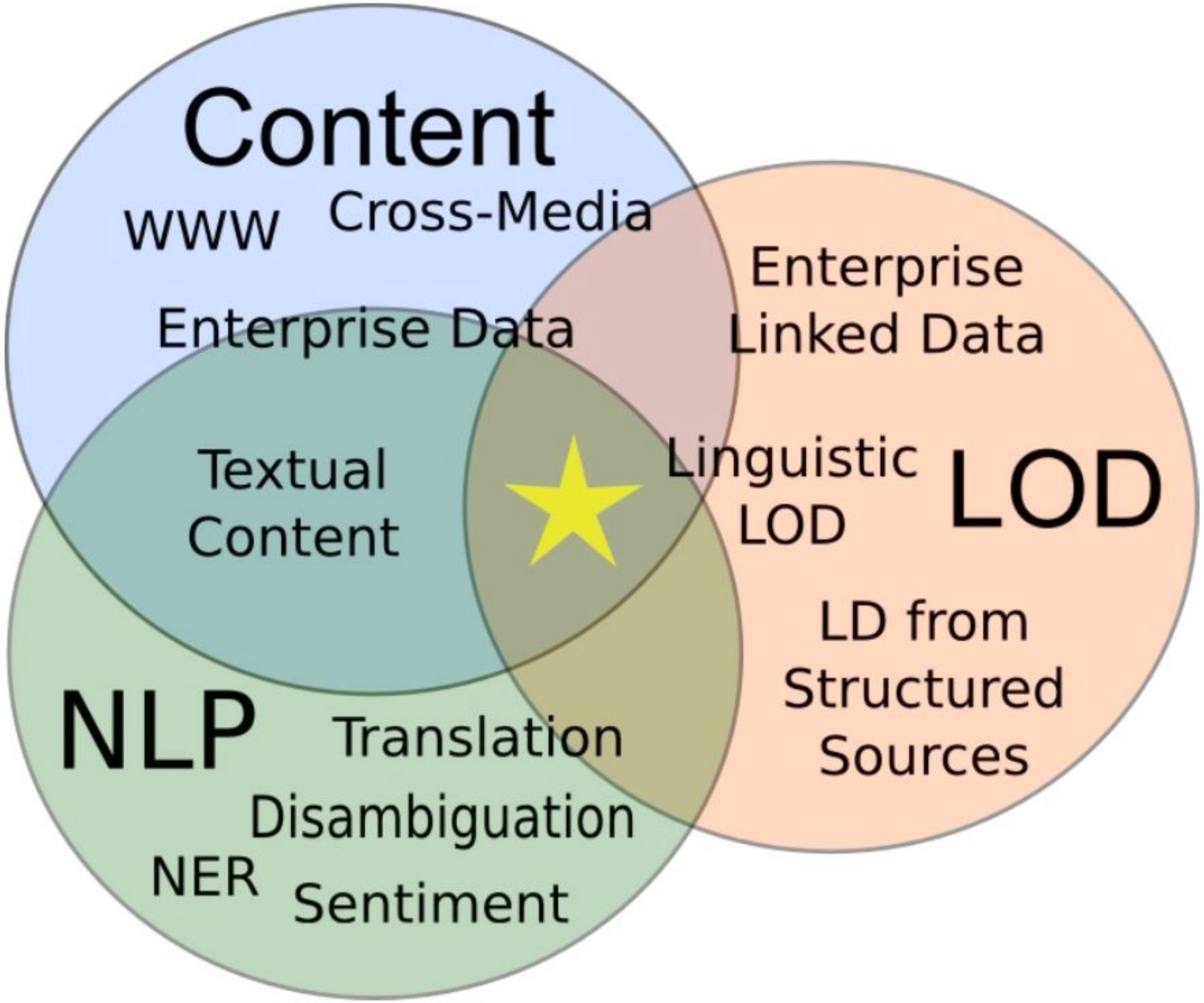
○ Many NLP tools fulfill similar functions but are **not interoperable**

○ **Heterogeneous** output formats (JSON, XML)

○ NLP Web services with **heterogeneous API parameters**

○ **Heterogeneous ways of annotating text**

Introduction – Bird's View



Outline

1. Motivation
2. NIF in the Nutshell
3. NLP (NIF) aware Web Services
4. Exercises
5. Q&A

NLP Interchange Format

The ***NLP Interchange Format (NIF)*** *is an RDF/OWL-based format that aims to achieve interoperability between NLP tools, language resources and annotations.*

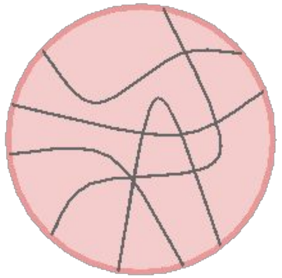
NIF in the Nutshell

- Way to **mint URIs** for arbitrary strings on the Web
- Logical **formalisation of strings** and **annotations** via an ontology
- Easy and human **understandable format**
- Builds on **existing standards** (RDF, LAF/GrAF, RFC 5147)
- **Reuses existing RDF tools** and implementations
- **Decreases development costs** for integration of tools and resources

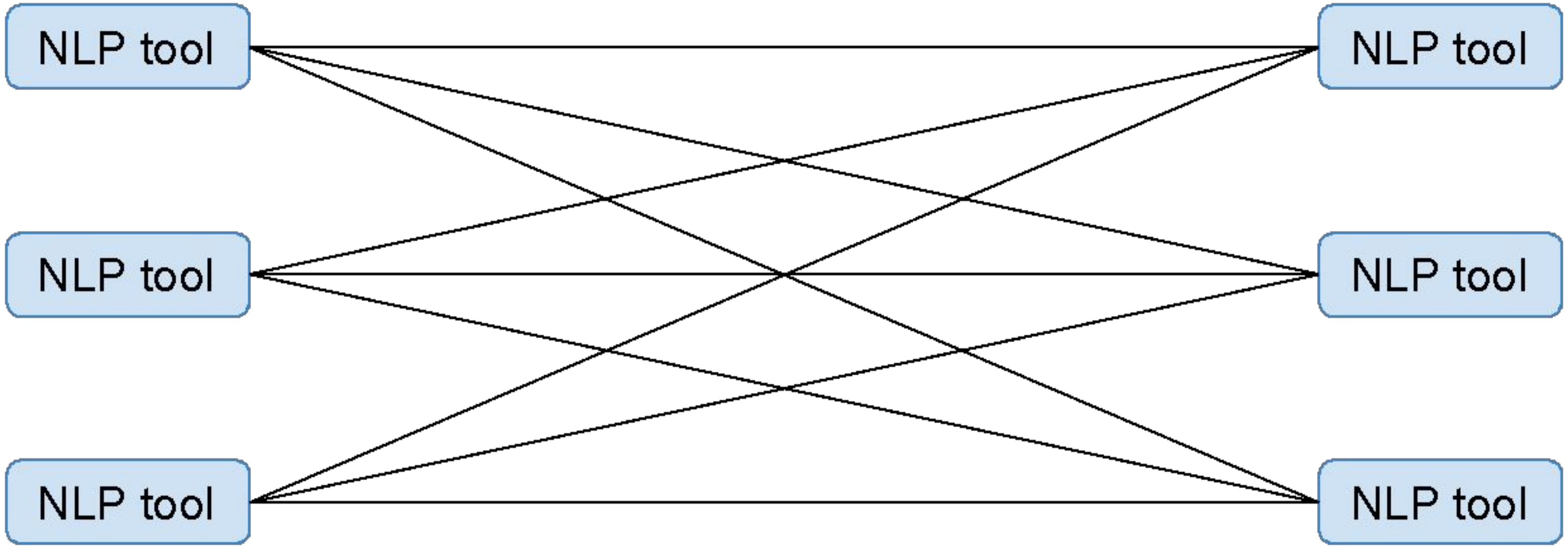
Pre-NIF Spaghetti Architecture

Need for integration

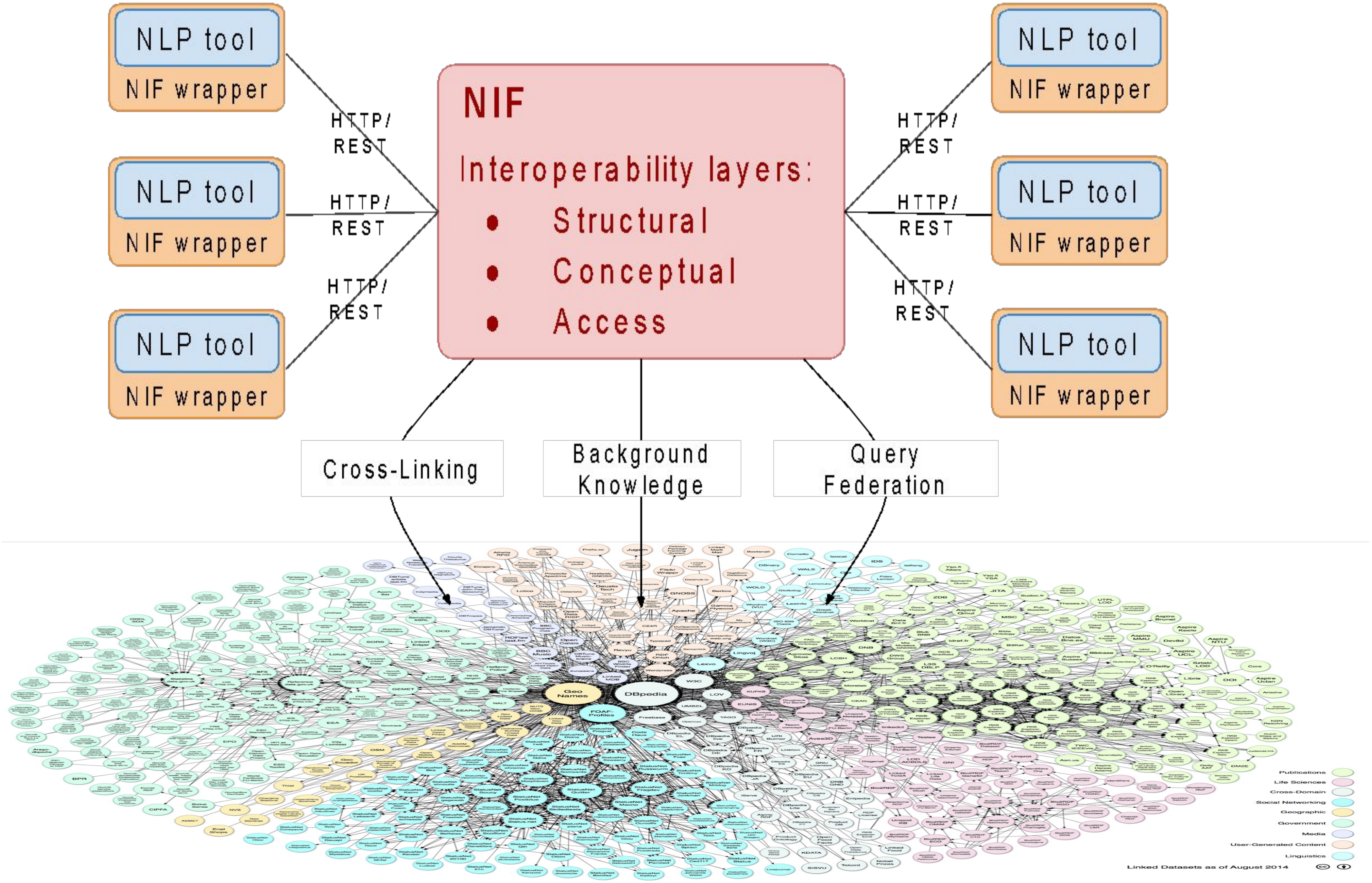
- One-to-one integration
- Hard to maintain



WTF! Spaghetti ?!!



NIF Architecture



NLP tool

NIF wrapper

HTTP/ REST

NIF

Interoperability layers:

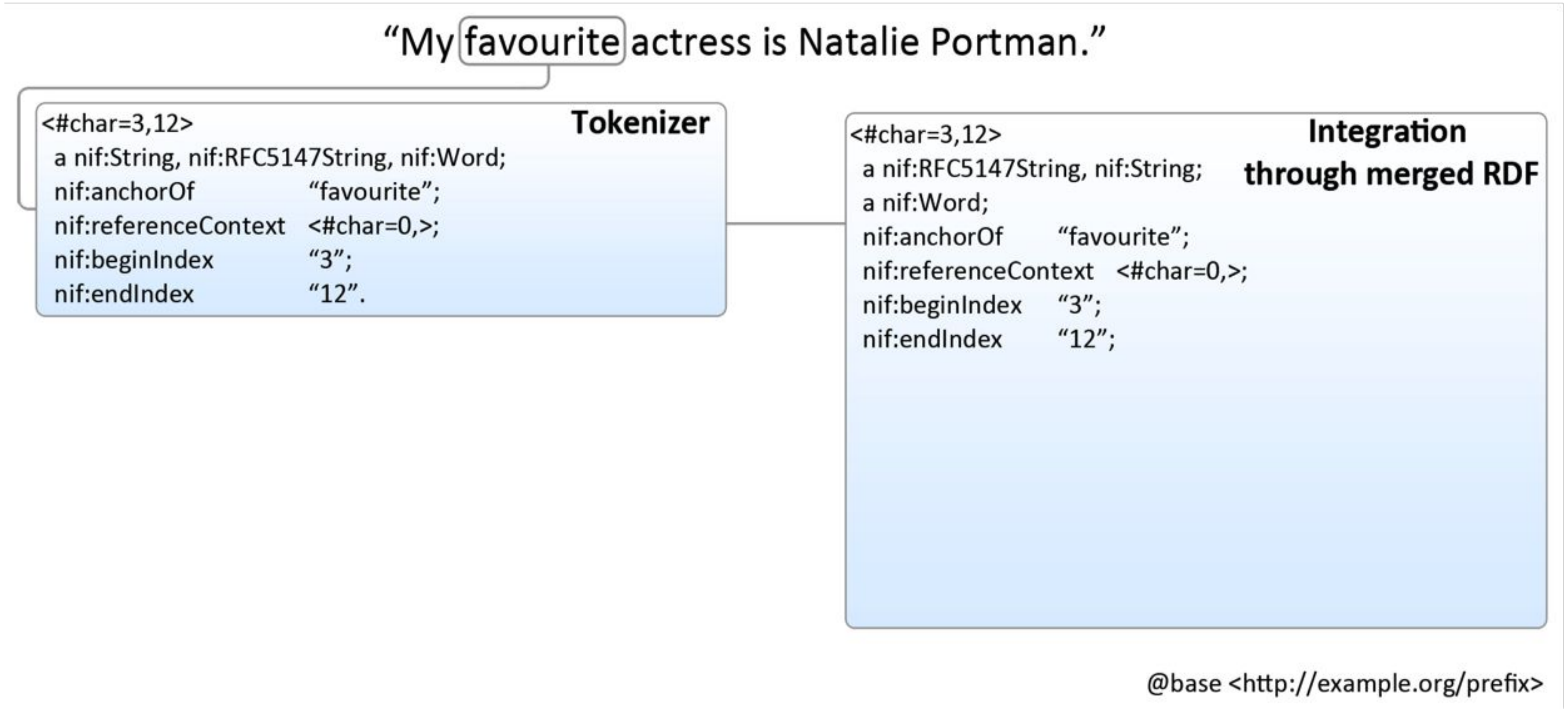
- Structural
- Conceptual
- Access

Cross-Linking

Background Knowledge

Query Federation

Simple tokenization



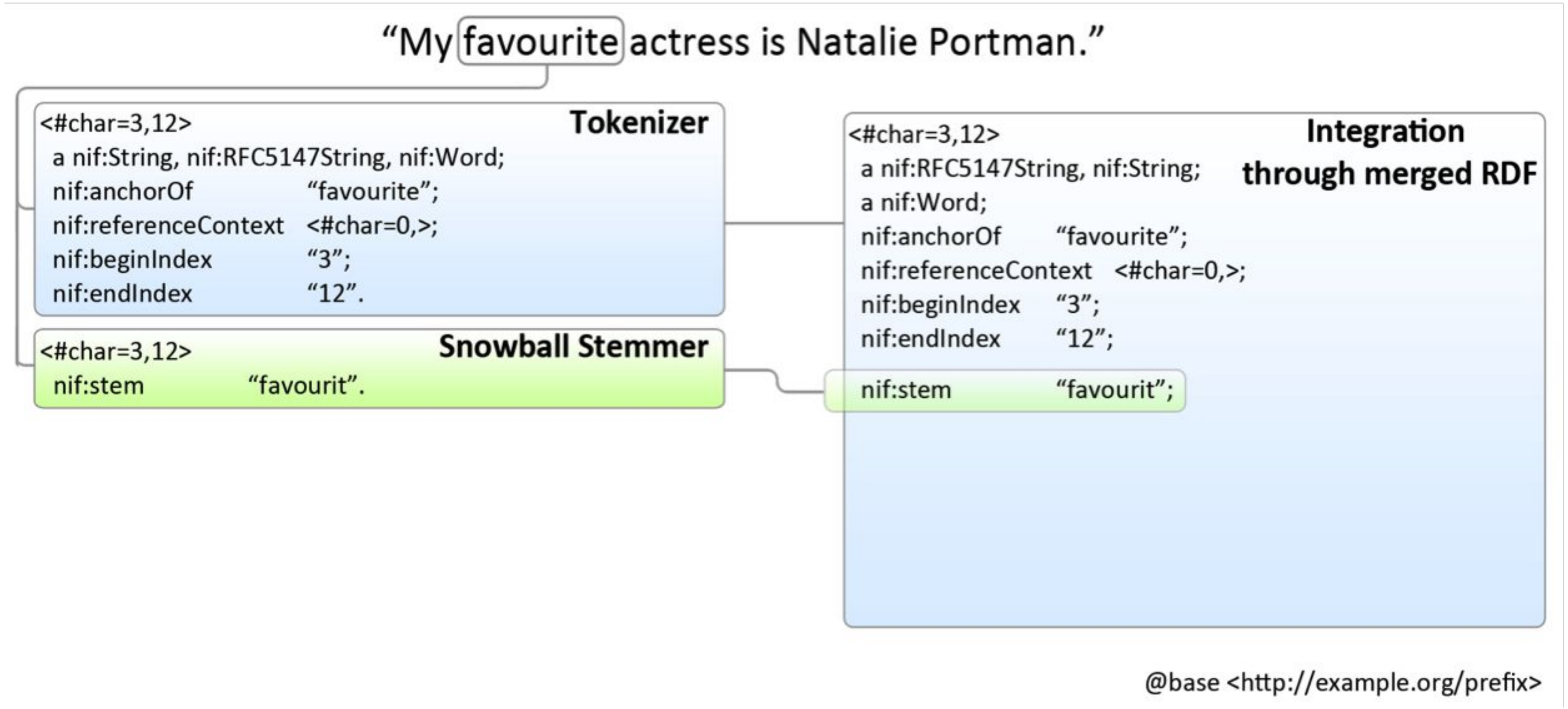
“My favourite actress is Natalie Portman.”

```
<#char=3,12>
  a nif:String, nif:RFC5147String, nif:Word;
  nif:anchorOf              “favourite”;
  nif:referenceContext   <#char=0,>;
  nif:beginIndex            “3”;
  nif:endIndex              “12”.
```

```
<#char=3,12>
  a nif:RFC5147String, nif:String;
  a nif:Word;
  nif:anchorOf        “favourite”;
  nif:referenceContext   <#char=0,>;
  nif:beginIndex      “3”;
  nif:endIndex        “12”;
```

@base <http://example.org/prefix>

… plus stemming



"My favourite actress is Natalie Portman."

Tokenizer

```
<#char=3,12>
 a nif:String, nif:RFC5147String, nif:Word;
 nif:anchorOf              "favourite";
 nif:referenceContext   <#char=0,>;
 nif:beginIndex            "3";
 nif:endIndex              "12".
```

Snowball Stemmer

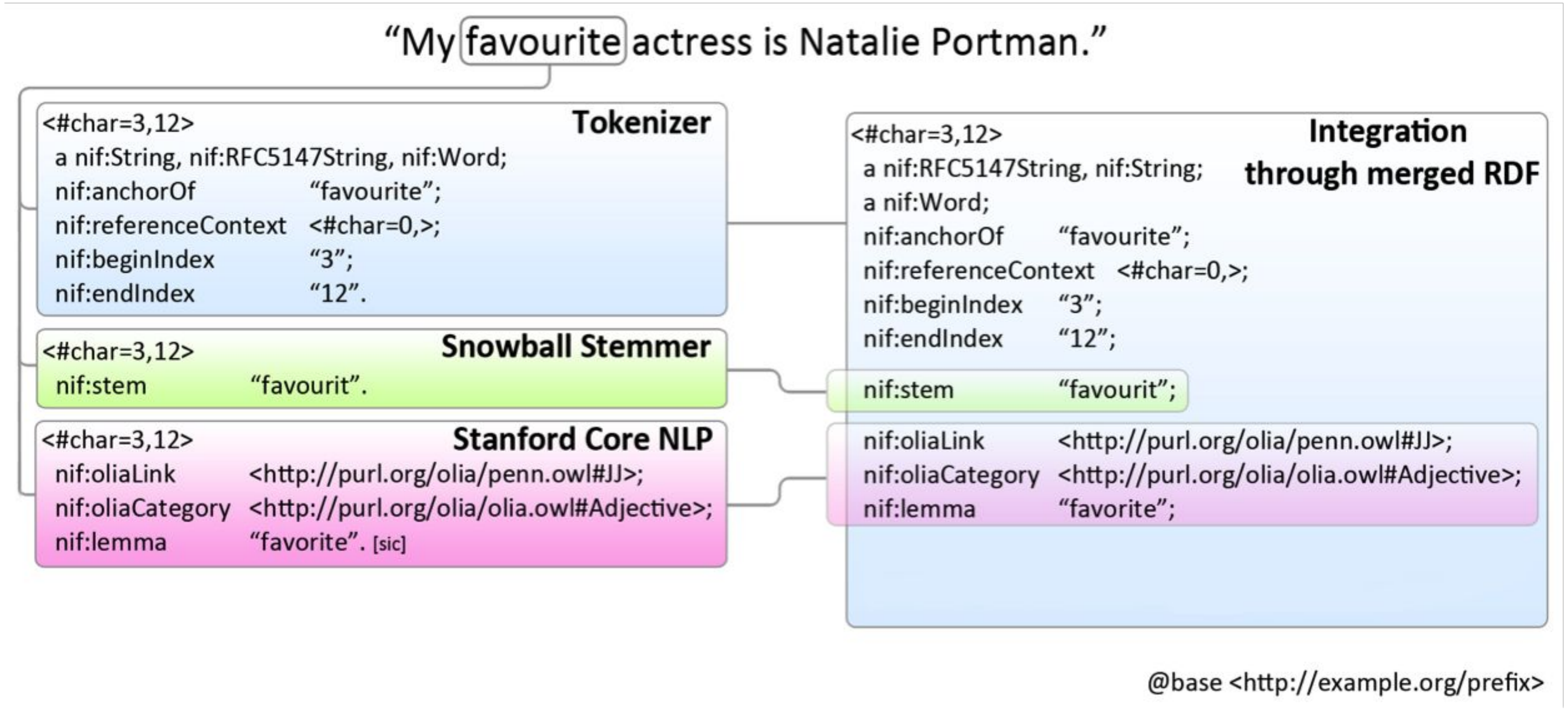
```
<#char=3,12>
 nif:stem              "favourit".
```

Integration through merged RDF

```
<#char=3,12>
 a nif:RFC5147String, nif:String;
 a nif:Word;
 nif:anchorOf        "favourite";
 nif:referenceContext   <#char=0,>;
 nif:beginIndex      "3";
 nif:endIndex        "12";
 nif:stem            "favourit";
```

@base <http://example.org/prefix>

… plus POS tagging



"My favourite actress is Natalie Portman."

Tokenizer

```
<#char=3,12>
  a nif:String, nif:RFC5147String, nif:Word;
  nif:anchorOf            "favourite";
  nif:referenceContext    <#char=0,>;
  nif:beginIndex          "3";
  nif:endIndex            "12".
```

Snowball Stemmer

```
<#char=3,12>
  nif:stem          "favourit".
```

Stanford Core NLP

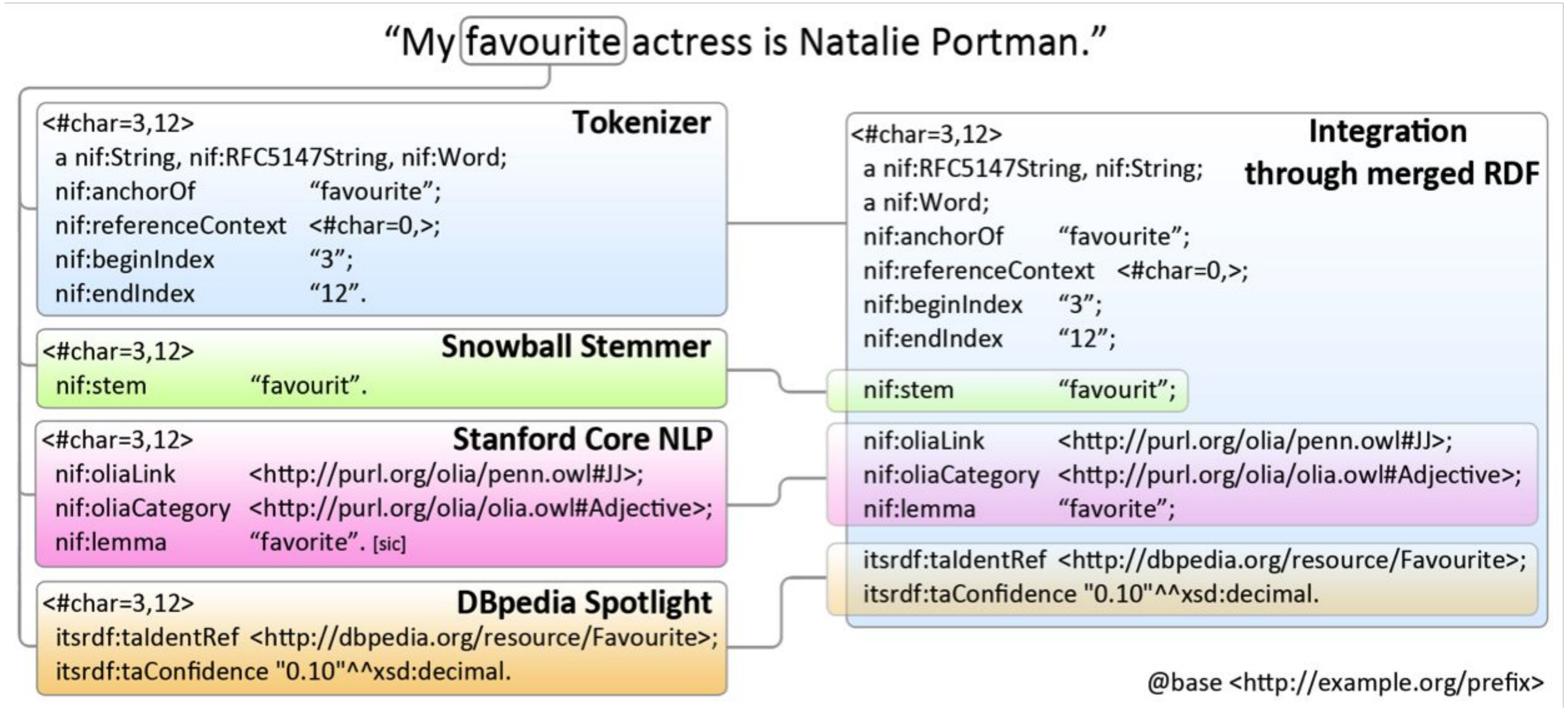
```
<#char=3,12>
  nif:oliaLink       <http://purl.org/olia/penn.owl#JJ>;
  nif:oliaCategory   <http://purl.org/olia/olia.owl#Adjective>;
  nif:lemma          "favorite". [sic]
```

Integration through merged RDF

```
<#char=3,12>
  a nif:RFC5147String, nif:String;
  a nif:Word;
  nif:anchorOf        "favourite";
  nif:referenceContext   <#char=0,>;
  nif:beginIndex      "3";
  nif:endIndex        "12";

  nif:stem            "favourit";

  nif:oliaLink        <http://purl.org/olia/penn.owl#JJ>;
  nif:oliaCategory    <http://purl.org/olia/olia.owl#Adjective>;
  nif:lemma           "favorite";
```

@base <http://example.org/prefix>

… plus Entity Linking



“My favourite actress is Natalie Portman.”

Tokenizer

```
<#char=3,12>
  a nif:String, nif:RFC5147String, nif:Word;
  nif:anchorOf           “favourite”;
  nif:referenceContext   <#char=0,>;
  nif:beginIndex         “3”;
  nif:endIndex           “12”.
```

Snowball Stemmer

```
<#char=3,12>
  nif:stem          “favourit”.
```

Stanford Core NLP

```
<#char=3,12>
  nif:oliaLink       <http://purl.org/olia/penn.owl#JJ>;
  nif:oliaCategory   <http://purl.org/olia/olia.owl#Adjective>;
  nif:lemma          “favorite”. [sic]
```

DBpedia Spotlight

```
<#char=3,12>
  itsrdf:taIdentRef <http://dbpedia.org/resource/Favourite>;
  itsrdf:taConfidence "0.10"^^xsd:decimal.
```

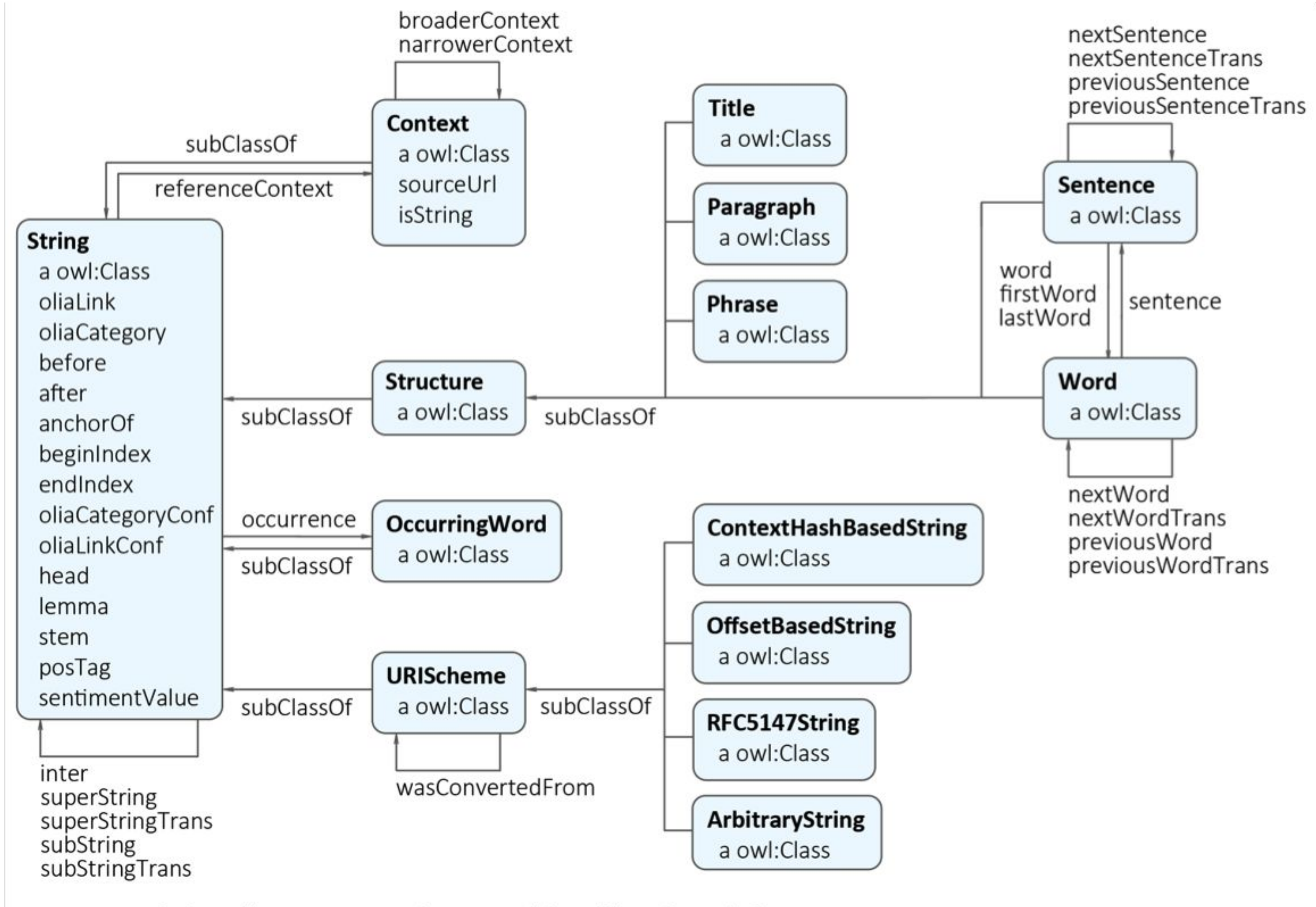
Integration through merged RDF

```
<#char=3,12>
  a nif:RFC5147String, nif:String;
  a nif:Word;
  nif:anchorOf        “favourite”;
  nif:referenceContext   <#char=0,>;
  nif:beginIndex      “3”;
  nif:endIndex        “12”;

  nif:stem            “favourit”;

  nif:oliaLink        <http://purl.org/olia/penn.owl#JJ>;
  nif:oliaCategory    <http://purl.org/olia/olia.owl#Adjective>;
  nif:lemma           “favorite”;

  itsrdf:taIdentRef <http://dbpedia.org/resource/Favourite>;
  itsrdf:taConfidence "0.10"^^xsd:decimal.
```

@base <http://example.org/prefix>

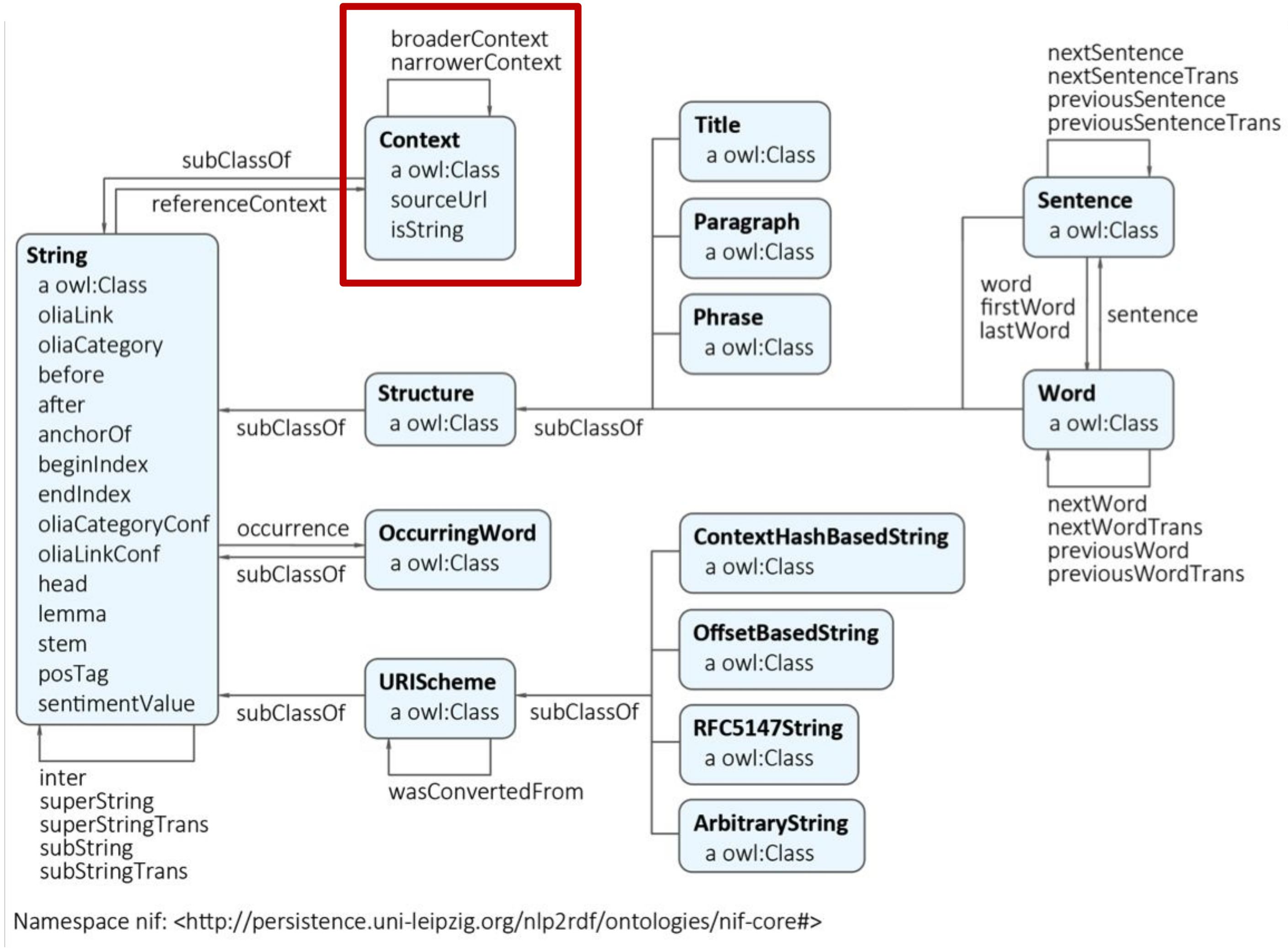
What You Need Is What You Get!

The NIF Ontology



Namespace nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>

NIF Context



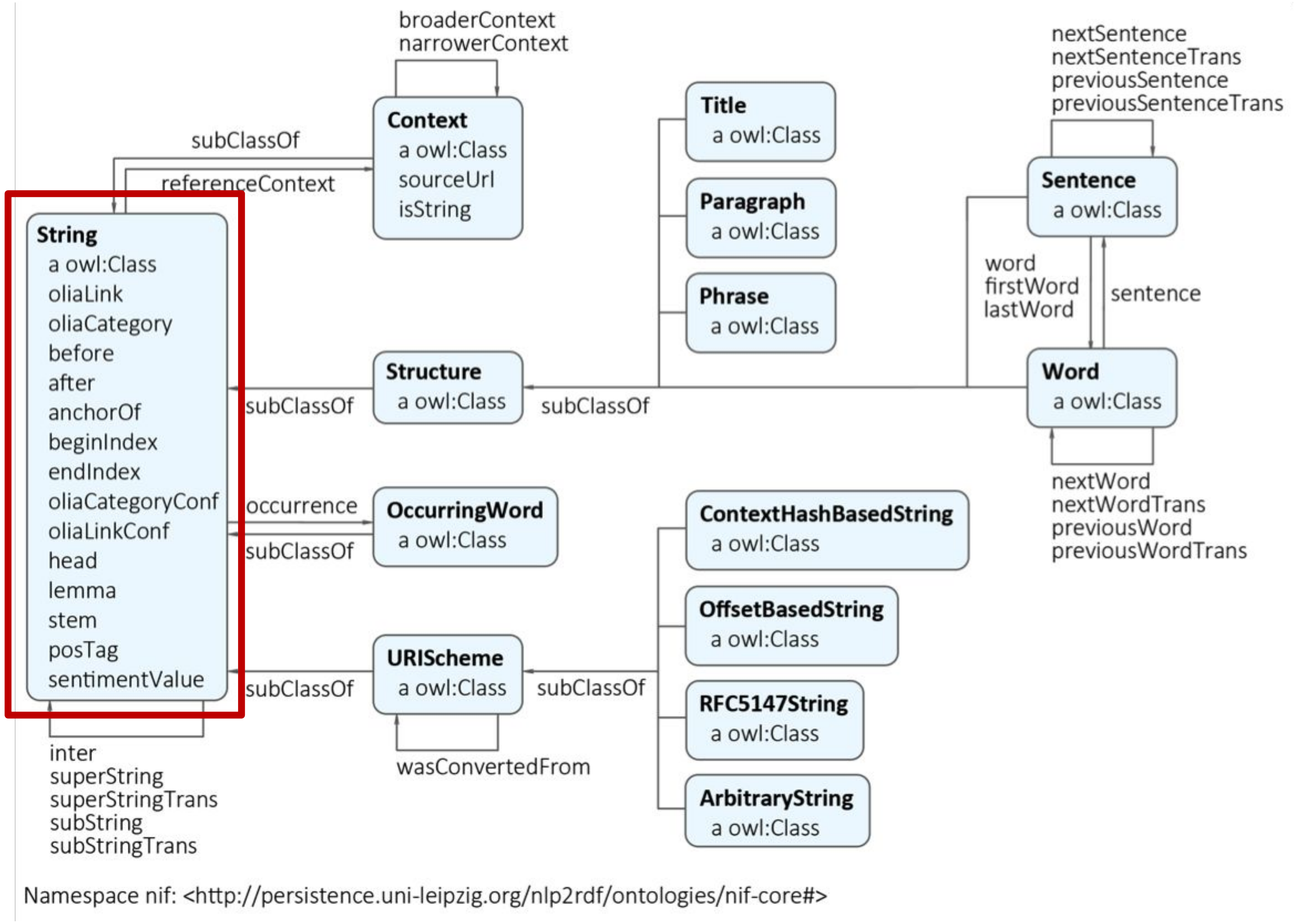
Namespace nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>

Context

```
@prefix nif:   <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .

<http://cli.nlp2rdf.org/snowball#char=0,39>
      a                   nif:RFC5147String , nif:Context , nif:Sentence ;
      nif:beginIndex      "0" ;
      nif:endIndex        "39" ;
      nif:isString        "My favorite actress is Natalie Portman." .
```

- nif:Context - the content of the document
- nif:isString contains document content
- In NIF the document != content of the document
- Two documents can have the same content, BUT must not have the same URI

NIF Strings



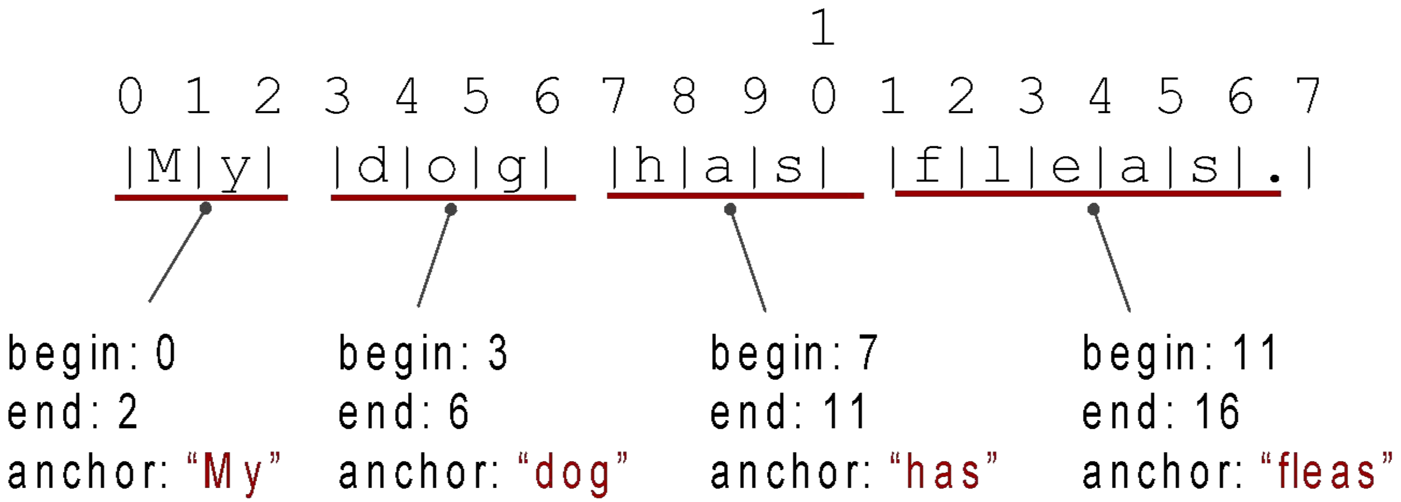
Namespace nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>

NIF Strings

```
@prefix nif:   <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .

<http://cli.nlp2rdf.org/snowball#char=23,30>
       a                  nif:RFC5147String , nif:Word ;
       nif:anchorOf        "Natalie" ;
       nif:beginIndex       "23" ;
       nif:endIndex        "30" ;
       nif:referenceContext  <http://cli.nlp2rdf.org/snowball#char=0,39> .
```

- Address arbitrary strings in the document
- To address use **string offsets in relation to the context**
- nif:anchorOf holds the string

Counting Offsets



```
                    1
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7
|M|y| |d|o|g| |h|a|s| |f|l|e|a|s|.|
```

begin: 0
end: 2
anchor: “My”

begin: 3
end: 6
anchor: “dog”

begin: 7
end: 11
anchor: “has”

begin: 11
end: 16
anchor: “fleas”

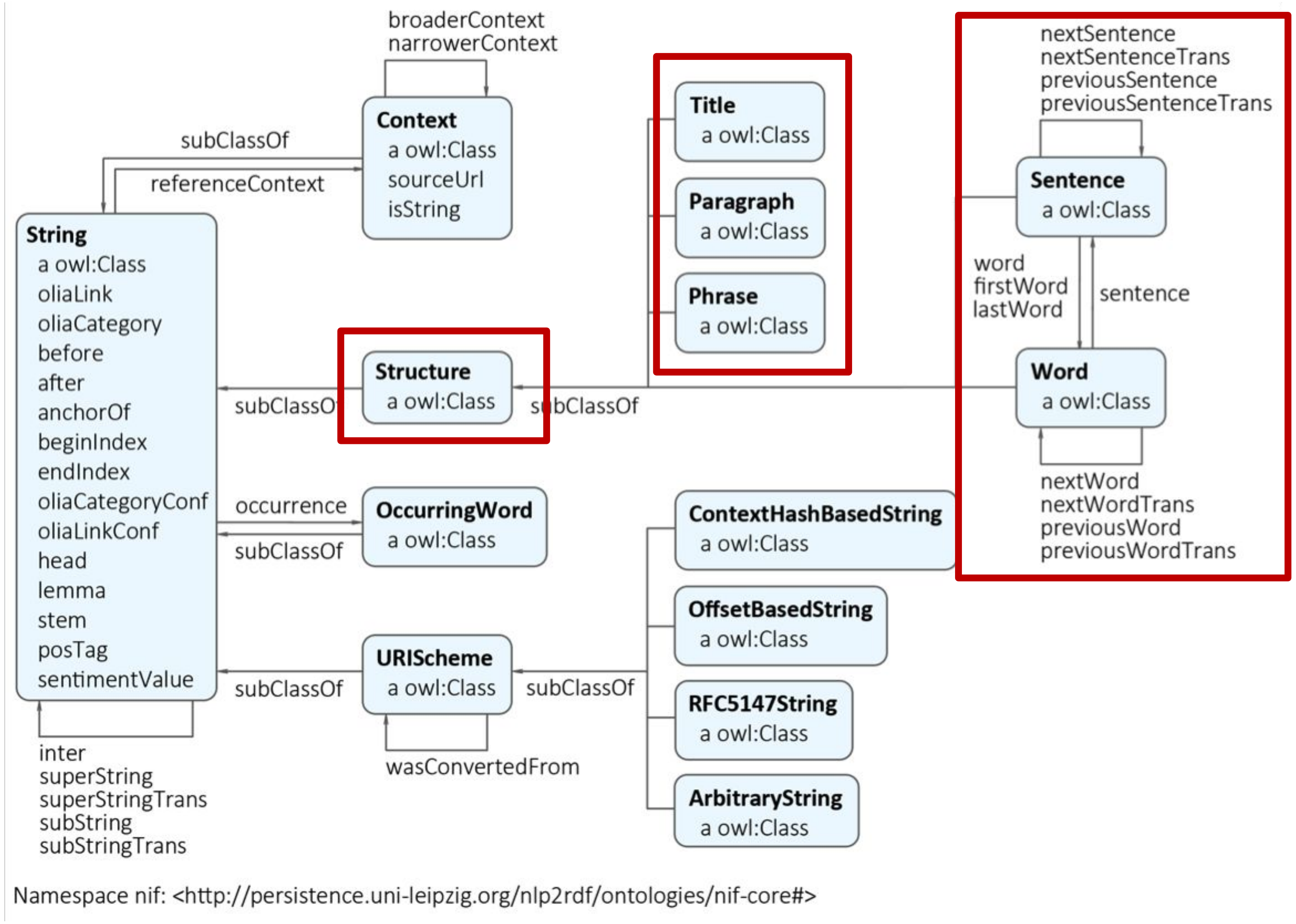
- Counting the gaps between the characters starting from 0 as specified in RFC 5147
- Exception: encoding Unicode Normal Form C (NFC) and counting is fixed on Unicode Code Units

Referencing Strings with the Context

```
@prefix nif:   <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .

<http://cli.nlp2rdf.org/snowball#char=23,30>
      a                    nif:RFC5147String , nif:Word ;
      nif:anchorOf         "Natalie" ;
      nif:beginIndex        "23" ;
      nif:endIndex         "30" ;
      nif:referenceContext  <http://cli.nlp2rdf.org/snowball#char=0,39> .
```

- nif:referenceContext property
  - a link between the string (annotation) and the context

NIF Structural Concepts



Namespace nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>

Words and Phrases

```
@prefix nif:   <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .

<http://cli.nlp2rdf.org/snowball#char=23,30>
      a                    nif:RFC5147String , nif:Word ;
      nif:anchorOf         "Natalie" ;
      nif:beginIndex        "23" ;
      nif:endIndex         "30" ;
      nif:referenceContext  <http://cli.nlp2rdf.org/snowball#char=0,39> .

<http://cli.nlp2rdf.org/snowball#char=23,38>
      a                    nif:RFC5147String , nif:Phrase ;
      nif:anchorOf         "Natalie Portman" ;
      nif:beginIndex        "23" ;
      nif:endIndex         "38" ;
      nif:referenceContext  <http://cli.nlp2rdf.org/snowball#char=0,39> .
```

- **nif:Word**, **nif:Phrase**

Sentences and Paragraphs

```
@prefix nif:   <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .

<http://cli.nlp2rdf.org/snowball#char=0,39>
        a                     nif:RFC5147String , nif:Context , nif:Sentence ;
        nif:anchorOf          "My favorite actress is Natalie Portman." ;
        nif:beginIndex        "0" ;
        nif:endIndex          "39" ;
        nif:firstWord         <http://cli.nlp2rdf.org/snowball#char=0,2> ;
        nif:isString          "My favorite actress is Natalie Portman." ;
        nif:referenceContext  <http://cli.nlp2rdf.org/snowball#char=0,39> .
```

- **nif:Sentence**, **nif:Paragraph**

Support for traversing

```
@prefix nif:   <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .

<http://cli.nlp2rdf.org/snowball#char=12,19>
        a                    nif:Word , nif:RFC5147String ;
        nif:anchorOf         "actress" ;
        nif:beginIndex        "12" ;
        nif:endIndex         "19" ;
        nif:nextWord         <http://cli.nlp2rdf.org/snowball#char=20,22> ;
        nif:previousWord     <http://cli.nlp2rdf.org/snowball#char=3,11> ;
        nif:referenceContext  <http://cli.nlp2rdf.org/snowball#char=0,39> ;
        nif:sentence         <http://cli.nlp2rdf.org/snowball#char=0,39> .
```

- **nif:previousWord, nif:nextWord**
- **nif:previousSentance, nif: nextSentence,**

Attachment of additional info to Strings

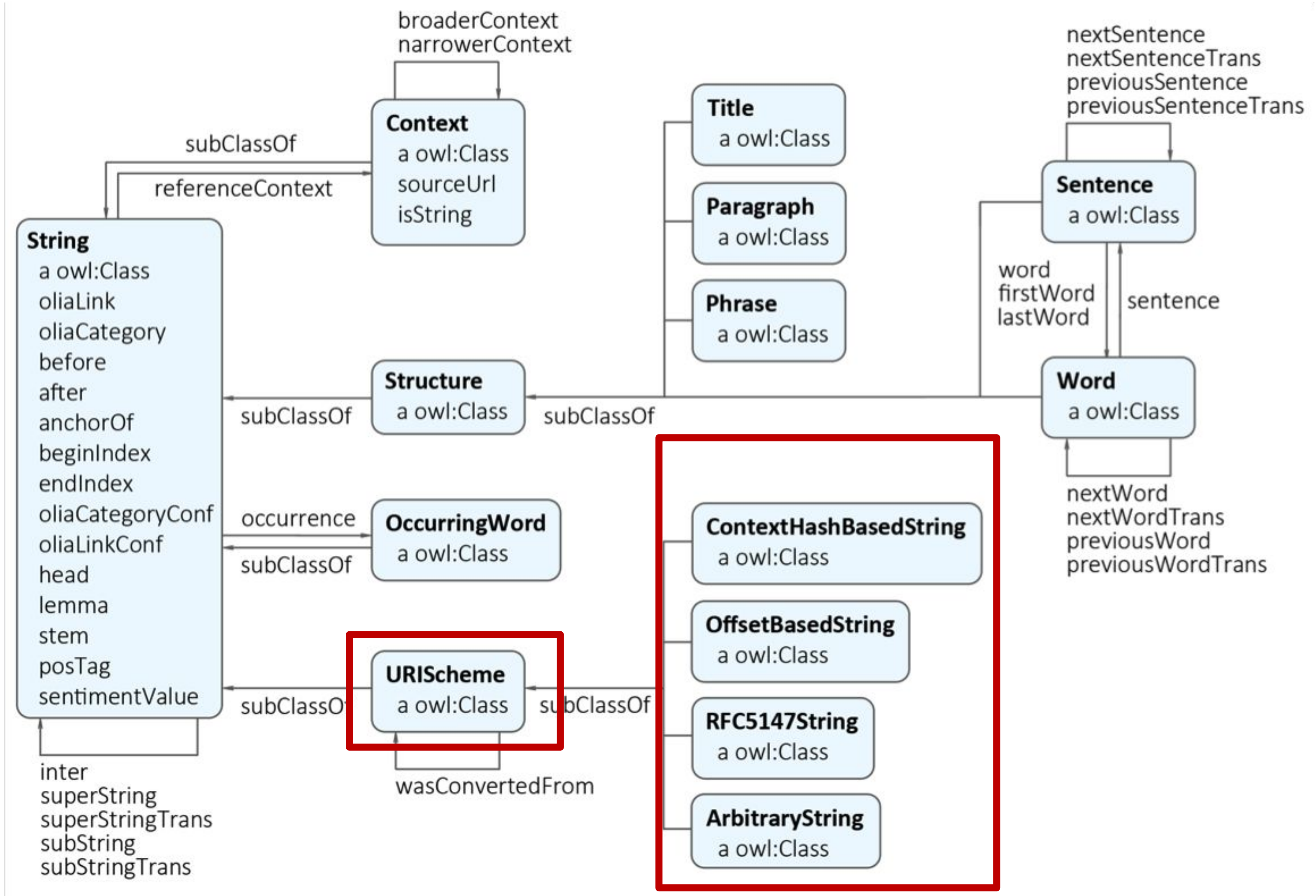
```
@prefix nif:   <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .

<http://cli.nlp2rdf.org/snowball#char=23,30>
      a                    nif:RFC5147String , nif:Word ;
      nif:anchorOf         "Natalie" ;
      nif:beginIndex       "23" ;
      nif:endIndex         "30" ;
      nif:referenceContext  <http://cli.nlp2rdf.org/snowball#char=0,39> ;
      nif:stem             "natali" .

<http://cli.nlp2rdf.org/snowball#char=3,11>
      a                    nif:Word , nif:RFC5147String ;
      nif:anchorOf         "favourite" ;
      nif:beginIndex       "3" ;
      nif:endIndex         "11" ;
      nif:referenceContext  <http://cli.nlp2rdf.org/snowball#char=0,39> ;
      nif:lemma            "favorite" ;
      nif:oliaLink         <http://purl.org/olia/penn.owl#JJ> ;
      nif:oliaCategory     <http://purl.org/olia/penn.owl#Adjective> .
```

Linking Annotations with LOD using ITS 2.0

```
@prefix nif:   <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .

<http://cli.nlp2rdf.org/snowball#char=23,38>
      a                    nif:RFC5147String , nif:Phrase ;
      nif:anchorOf         "Natalie Portman" ;
      nif:beginIndex        "23" ;
      nif:endIndex         "38" ;
      itsrdf:taIdentRef   <http://dbpedia.org/resource/Natalie_Portman> ;
      itsrdf:taConfidence   “0.10”^^xsd:decimal ;
      nif:referenceContext  <http://cli.nlp2rdf.org/snowball#char=0,39> .
```

- Widely exploited by NER systems
- Reuse of the ITS 2.0 tagset: https://www.w3.org/TR/its20/

URI Scheme for the URIs



Namespace nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>

Minting URIs for strings

- **RFC 5147: “URI fragment identifiers” (spec)**

<http://cli.nlp2rdf.org/snowball#char=23,30>

- The comma char “,” not allowed in the local part of prefixed IRIs

- **Offset based strings**

<http://cli.nlp2rdf.org/snowball#offset_23_30>

- instability with regard to changes in the document
  - In case of a document change (i.e. insertion or deletion of characters), all URIs after the position become invalid.

Minting URIs for strings (cont.)

- **Context-Hash-based URIs**

Following URI for the string “ the “: http://cli.nlp2rdf.org/snowball#

hash_1_5_8dc0d6c8afa469c52ac4981011b3f582_%20the%20

- The URI consists of:
  - string “hash” -> **hash_**
  - context length -> number of chars before and after the string for the hash **_1_**
  - the length of the string -> “ the “ -> **_5_**
  - message digest (MD5) of leftContext(String)rightContext
  - the string itself (URL encoded) -> **_%20the%20**

Outline

1. Motivation
2. NIF in the Nutshell
3. NIF aware Web Services
4. Exercises
5. Q&A

NIF aware Web Services

Web (or local) Services which:

- consume NIF (optional, plain text is acceptable)
- cenerate NIF

… the NLP task is specific to the consumed Web service.

The ultimate goal is to align **various NLP tools and services** to communicate in a common language, i.e. NIF.

- The NIF API defines are common communication protocol.

NIF Web Service API

input: depends on informat/intype
informat: turtle, text
intype: direct, url, file
outformat: turtle, text
urischeme: RFC5147String (default), OffsetBased, ContextHashBased
prefix: namespace for the URIs

… but also consider

- Accept and Content-Type HTTP headers
- HTTP Status Codes: 200, 400, 401, 406, etc.

See https://persistence.uni-leipzig.org/nlp2rdf/specification/api.html

NIF aware Web Services

```
curl --data-urlencode input="My favourite actress is Natalie Portman."
-d informat=text "http://nlp2rdf.lod2.eu/nif-ws.php"
```

HTTP Request:

```
> POST /nif-ws.php HTTP/1.1
> Host: nlp2rdf.lod2.eu
> Content-Length: 70
> Content-Type: application/x-www-form-urlencoded
```

HTTP Response:

```
< HTTP/1.1 200 OK
< Content-Type: text/turtle; charset=UTF-8
```

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix nif:
<http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
<http://nlp2rdf.lod2.eu/nif-ws.php#char=0,40>
  rdf:type nif:RFC5147String , nif:Context ;
  nif:beginIndex "0" ;
  nif:endIndex "40" ;
  nif:isString "My favourite actress is Natalie Portman." .
```

Outline

1. Motivation
2. NIF in the Nutshell
3. NIF aware Web Services
4. Exercises
    - Exercise 1: Stemming using Snowball Stemmer
    - Exercise 2: POS tagging using OpenNLP
5. Q&A

Exercise 1: Stemming

Lets perform some stemming using the Snowball stemmer.

Instructions:

1. Download: NIF_tutorial_hands_on-02-2021.zip (/exercises folder)

https://drive.google.com/file/d/1vY2ekwyrDJBOkvDQChbyLb6nFBd85DFX/view?usp=sharing

1. Open the “instructions.txt” file in a text editor
2. Open a terminal
3. Go to the “jar” folder
4. Copy the first command of the instructions instructions.txt

```
java -jar snowball.jar -f text -i 'My favorite actress is Natalie Portman.’
```

1. Paste the command in the terminal

Results from stemming

```
java -jar snowball.jar -f text -i "I am connected."
```

Standard NIF annotations

```
<http://cli.nlp2rdf.org/snowball#char=5,14>
      a                    nif:Word , nif:RFC5147String ;
      nif:anchorOf         "connected" ;
      nif:beginIndex       "5" ;
      nif:endIndex         "14" ;
      nif:nextWord         <http://cli.nlp2rdf.org/snowball#char=14,15> ;
      nif:previousWord     <http://cli.nlp2rdf.org/snowball#char=2,4> ;
      nif:referenceContext <http://cli.nlp2rdf.org/snowball#char=0,15> ;
      nif:sentence         <http://cli.nlp2rdf.org/snowball#char=0,15> ;
      nif:stem             "connect" .
```

String offsets

Snowball stem annotation

Exercise 2: POS tagging

Lets do some POS tagging using OpenNLP. In the terminal enter:

```
java -jar opennlp.jar -f text -i "My favorite actress is Natalie Portman."
-modelFolder ../model/
```

- The `-modelFolder` parameter set the folder that contains the POS tagging trained models and tokenization
- You might add the parameter `--outfile output.ttl` to store the NIF triples in a file

Results from the POS tagging

```
<http://cli.nlp2rdf.org/opennlp#char=31,38>
      a                nif:Word , nif:RFC5147String ;
      nif:anchorOf       "Portman" ;
      nif:beginIndex      "31" ;
      nif:endIndex        "38" ;
      nif:oliaCategory     olia:Noun , olia:ProperNoun ;
      nif:oliaLink        <http://purl.org/olia/penn.owl#NNP> ;
      nif:referenceContext  <http://cli.nlp2rdf.org/opennlp#char=0,39> .

<http://cli.nlp2rdf.org/opennlp#char=12,19>
      a                nif:RFC5147String , nif:Word ;
      nif:anchorOf       "actress" ;
      nif:beginIndex      "12" ;
      nif:endIndex        "19" ;
      nif:oliaCategory     olia:Noun , olia:CommonNoun ;
      nif:oliaLink        <http://purl.org/olia/penn.owl#NN> ;
      nif:referenceContext  <http://cli.nlp2rdf.org/opennlp#char=0,39> .
```

Want more? Lets try the Stanford library

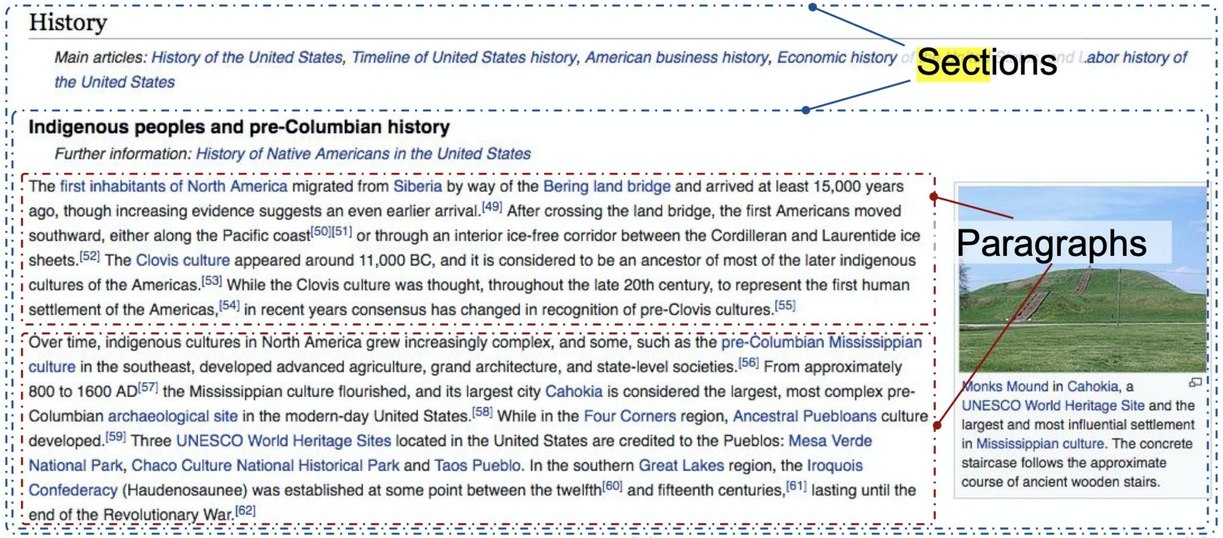
```
java -jar opennlp.jar -f text -i "My favorite actress is Natalie Portman."
-modelFolder ../model/
```

… and perform at once:

- tokenization
- sentence splitting
- POS tagging
- lemmatization

Spotlight: DBpedia NIF

- Open, Large-Scale and Multilingual Knowledge Extraction Corpus
  - The content of all articles for 128 Wikipedia languages.
  - The structure and content described using NIF.
  - Sections, paragraphs, titles and links.



History

Main articles: History of the United States, Timeline of United States history, American business history, Economic history o nd Labor history of the United States

Indigenous peoples and pre-Columbian history

Further information: History of Native Americans in the United States

The first inhabitants of North America migrated from Siberia by way of the Bering land bridge and arrived at least 15,000 years ago, though increasing evidence suggests an even earlier arrival.[49] After crossing the land bridge, the first Americans moved southward, either along the Pacific coast[50][51] or through an interior ice-free corridor between the Cordilleran and Laurentide ice sheets.[52] The Clovis culture appeared around 11,000 BC, and it is considered to be an ancestor of most of the later indigenous cultures of the Americas.[53] While the Clovis culture was thought, throughout the late 20th century, to represent the first human settlement of the Americas,[54] in recent years consensus has changed in recognition of pre-Clovis cultures.[55]

Over time, indigenous cultures in North America grew increasingly complex, and some, such as the pre-Columbian Mississippian culture in the southeast, developed advanced agriculture, grand architecture, and state-level societies.[56] From approximately 800 to 1600 AD[57] the Mississippian culture flourished, and its largest city Cahokia is considered the largest, most complex pre-Columbian archaeological site in the modern-day United States.[58] While in the Four Corners region, Ancestral Puebloans culture developed.[59] Three UNESCO World Heritage Sites located in the United States are credited to the Pueblos: Mesa Verde National Park, Chaco Culture National Historical Park and Taos Pueblo. In the southern Great Lakes region, the Iroquois Confederacy (Haudenosaunee) was established at some point between the twelfth[60] and fifteenth centuries,[61] lasting until the end of the Revolutionary War.[62]

Monks Mound in Cahokia, a UNESCO World Heritage Site and the largest and most influential settlement in Mississippian culture. The concrete staircase follows the approximate course of ancient wooden stairs.

Sections

Paragraphs

Get it from: https://databus.dbpedia.org/dbpedia/text/

Q&A

Thank you for your attention!

… and looking to your further exploration and exploitation of the NIF format!

Feel free to contact me at:

dojcinovski.milan@gmail.com

Acknowledgements

This presentation was given at the Eurolan 2021 Training School. The school has been primarily supported by the NexusLinguarum COST Action

Nexus Linguarum

COST Action CA18209 - NexusLinguarum “European network for Web-centred linguistic data science (https://nexuslinguarum.eu/)

References used

Web Resources

NIF 2.0 Core Ontology:
https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html

NIF API Spec:
https://persistence.uni-leipzig.org/nlp2rdf/specification/api.html

NIF Core Spec:
https://persistence.uni-leipzig.org/nlp2rdf/specification/core.html

RFC 5147:
https://tools.ietf.org/html/rfc5147

Turtle Spec:
https://www.w3.org/TR/turtle/

Literature

Milan Dojchinovski, Julio Hernandez, Markus Ackermann, Amit Kirschenbaum, & Sebastian Hellmann. (2018). DBpedia NIF: Open, Large-Scale and Multilingual Knowledge Extraction Corpus. https://arxiv.org/abs/1812.10315

Hellmann, Sebastian, Jens Lehmann, and Sören Auer. "Linked-data aware uri schemes for referencing text fragments." International Conference on Knowledge Engineering and Knowledge Management. Springer, Berlin, Heidelberg, 2012. http://jens-lehmann.org/files/2012/ekaw_nif.pdf

Ideas for some home works

- Process your own content
  - own corpus
  - local newspaper or Wikipedia or BBC or … your favourite website.

- Analyze the content
  - number of sentences, words, phrases, POS tags, entities, etc.

- Query the results
  - load the data in your favourite triple store
  - and run some cool SPARQL queries