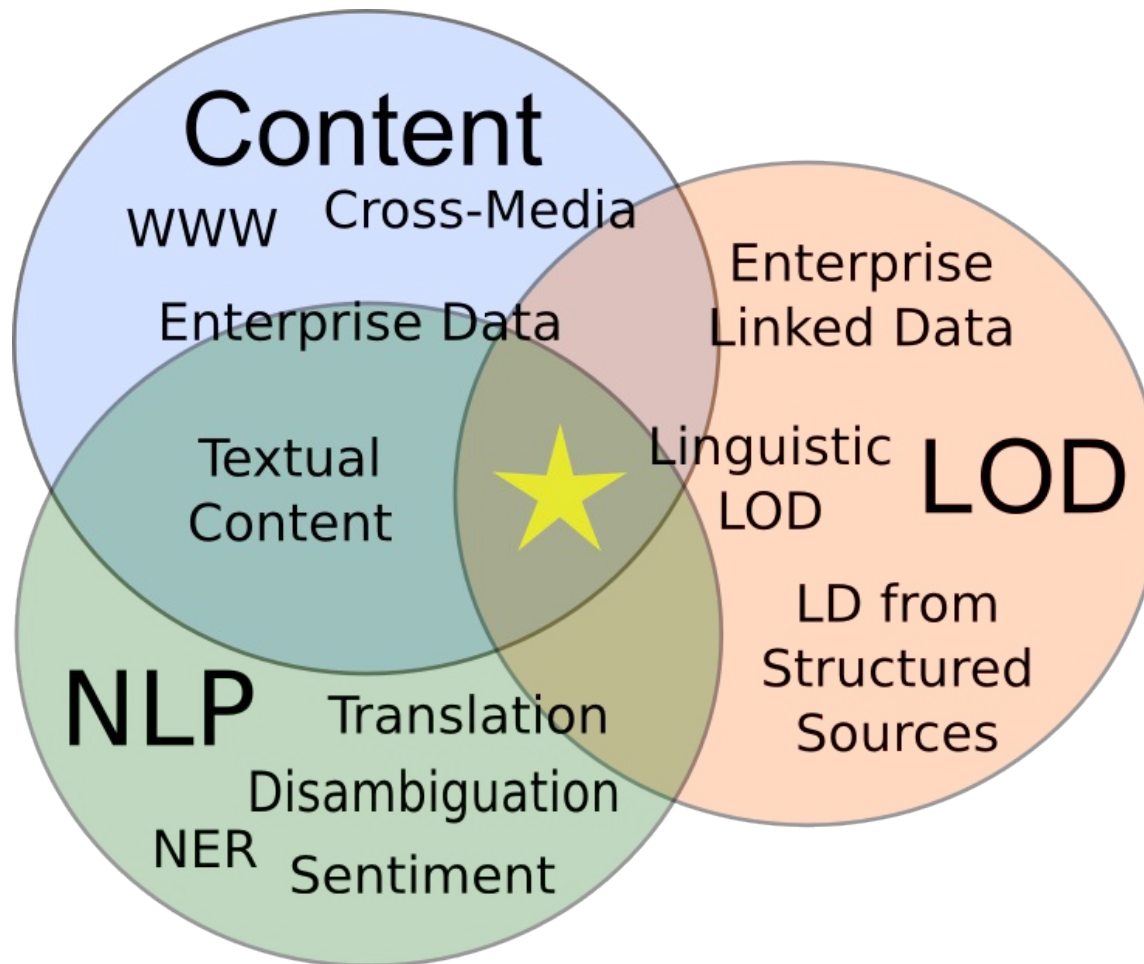




Building the Multilingual Web of Data

Integrating NLP with Linked Data and
RDF using the NLP Interchange
Format

1. Introduction
2. NIF Basics
3. NIF corpora
4. NIF tools & services



Not only data, but also LOD-aware services using:

Lexica and dictionaries in RDF (lemon model)

Training data for NLP in RDF (NIF model)

Service metadata descriptions in RDF

Combination with real world facts (i.e. DBpedia or Geoname)

Long Term Goal

→ index of tools and data

→ users can easily produce ready-made, preconfigured NLP services and pipelines

→ freemium / pay-per use business models



<http://nlp2rdf.org>

Realize this long-term goal

Sustainably maintain and consolidate results from short-term projects

Bootstrap the Eco-System

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

In a nutshell:

- Way to mint URIs for arbitrary strings and content of documents on the web

In a nutshell:

- Way to mint URIs for arbitrary strings and content of documents on the web
- Logical formalisation of strings and annotations via an ontology

In a nutshell:

- Way to mint URIs for arbitrary strings and content of documents on the web
- Logical formalisation of strings and annotations via an ontology
- Quick and easy format

In a nutshell:

- Way to mint URIs for arbitrary strings and content of documents on the web
- Logical formalisation of strings and annotations via an ontology
- Quick and easy format
- Builds on existing standards, e.g. RDF, LAF/GrAF, RFC 5147

In a nutshell:

- Way to mint URIs for arbitrary strings and content of documents on the web
- Logical formalisation of strings and annotations via an ontology
- Quick and easy format
- Builds on existing standards, e.g. RDF, LAF/GrAF, RFC 5147
- Reusability of RDF tools and implementation

In a nutshell:

- Way to mint URIs for arbitrary strings and content of documents on the web
- Logical formalisation of strings and annotations via an ontology
- Quick and easy format
- Builds on existing standards, e.g. RDF, LAF/GrAF, RFC 5147
- Reusability of RDF tools and implementation
- Decreases development cost for integration

Developers nightmare:

- All tools belong to similar class of NLP tools
 - Fulfill similar functions but often not interoperable ecosystems

Developers nightmare:

- All tools belong to similar class of NLP tools
 - Fulfill similar functions but often not interoperable ecosystems

All have:

- Heterogeneous output formats (JSON, XML, ...)

Developers nightmare:

- All tools belong to similar class of NLP tools
 - Fulfill similar functions but often not interoperable ecosystems

All have:

- Heterogeneous output formats (JSON, XML, ...)
- Heterogeneous API parameters

Developers nightmare:

- All tools belong to similar class of NLP tools
 - Fulfill similar functions but often not interoperable ecosystems

All have:

- Heterogeneous output formats (JSON, XML, ...)
- Heterogeneous API parameters
- Heterogeneous way of annotating text:

Developers nightmare:

- All tools belong to similar class of NLP tools
 - Fulfill similar functions but often not interoperable ecosystems

All have:

- Heterogeneous output formats (JSON, XML, ...)
- Heterogeneous API parameters
- Heterogeneous way of annotating text:
 - Some remove HTML internally, offsets not usable

Developers nightmare:

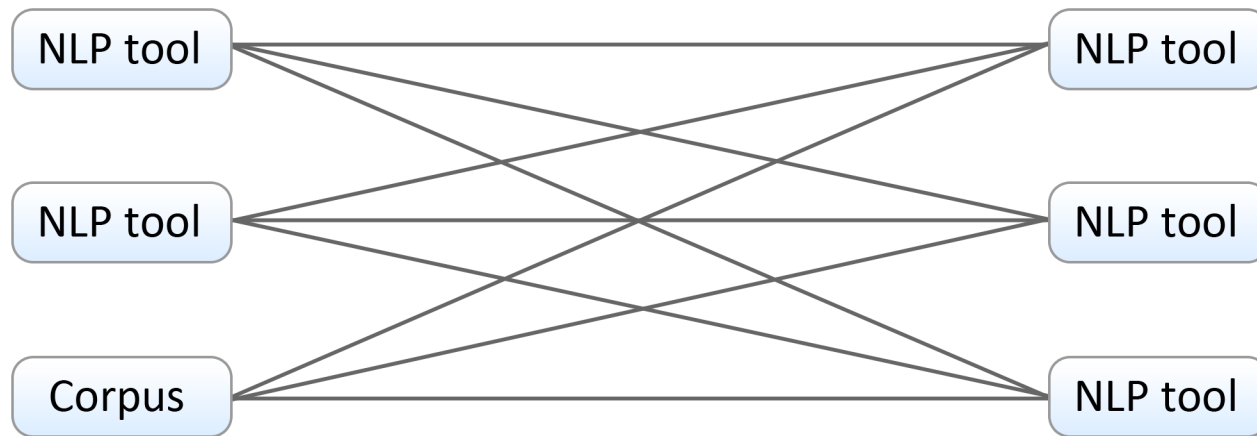
- All tools belong to similar class of NLP tools
 - Fulfill similar functions but often not interoperable ecosystems

All have:

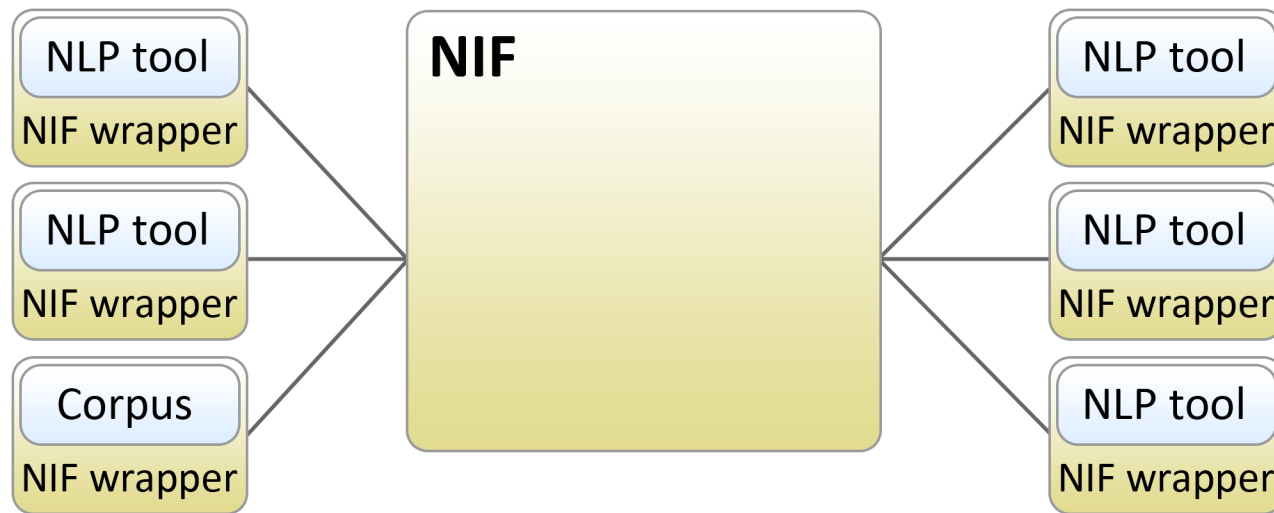
- Heterogeneous output formats (JSON, XML, ...)
- Heterogeneous API parameters
- Heterogeneous way of annotating text:
 - Some remove HTML internally, offsets not usable
 - Some use byte offset instead of char offset

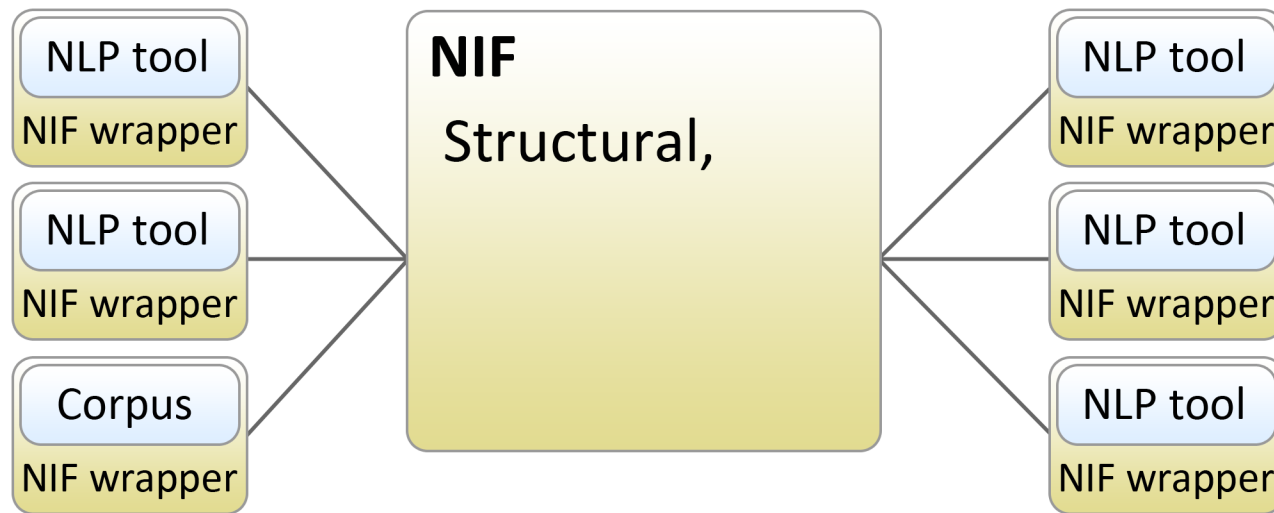
1. Introduction
- 2. NIF Basics**
3. NIF corpora
4. NIF tools & services

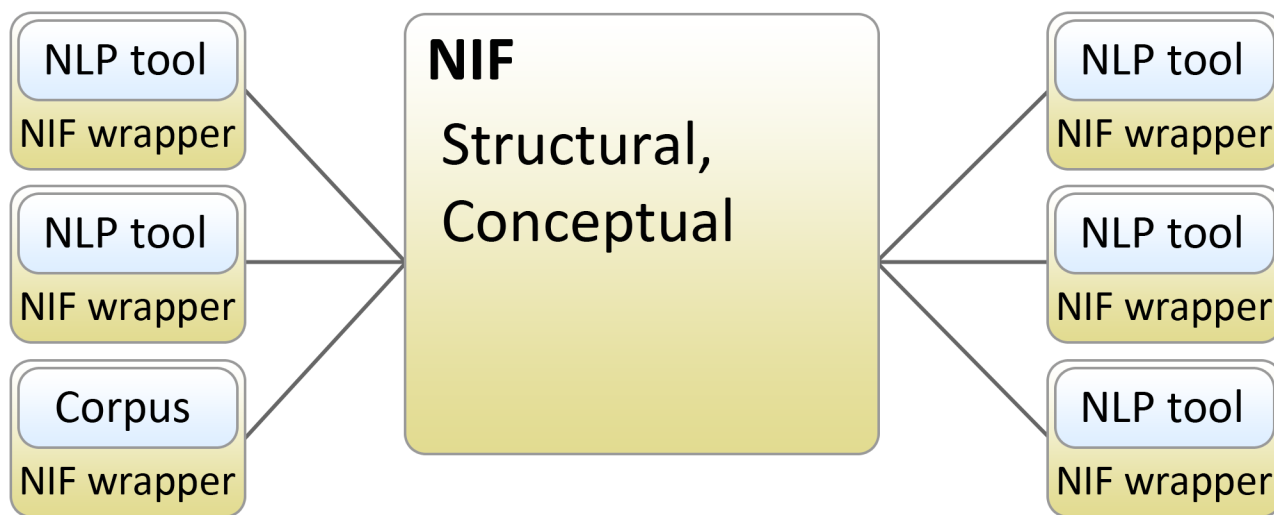


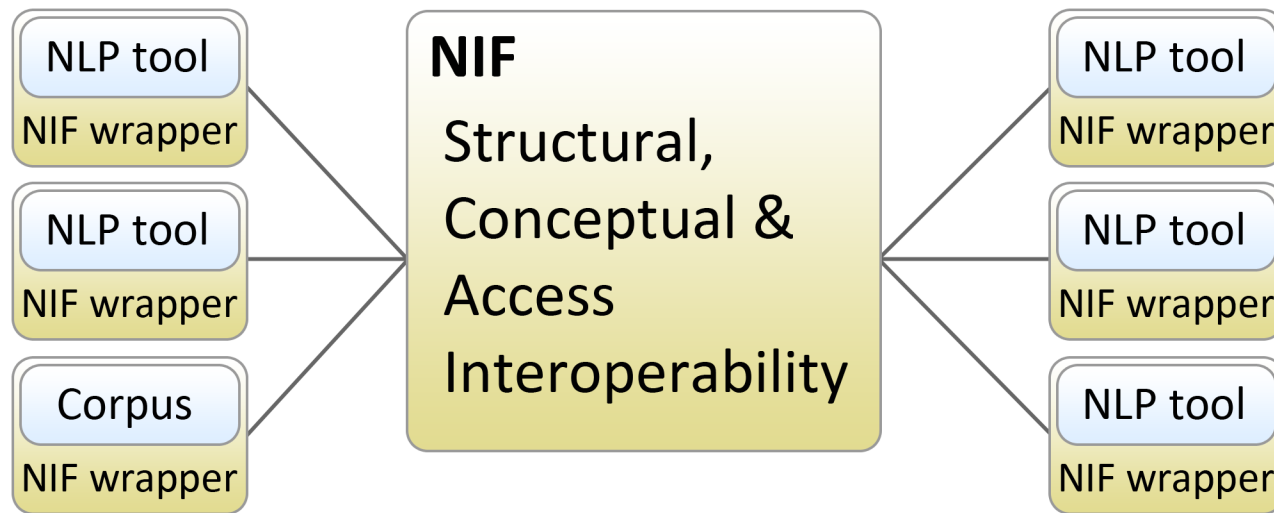


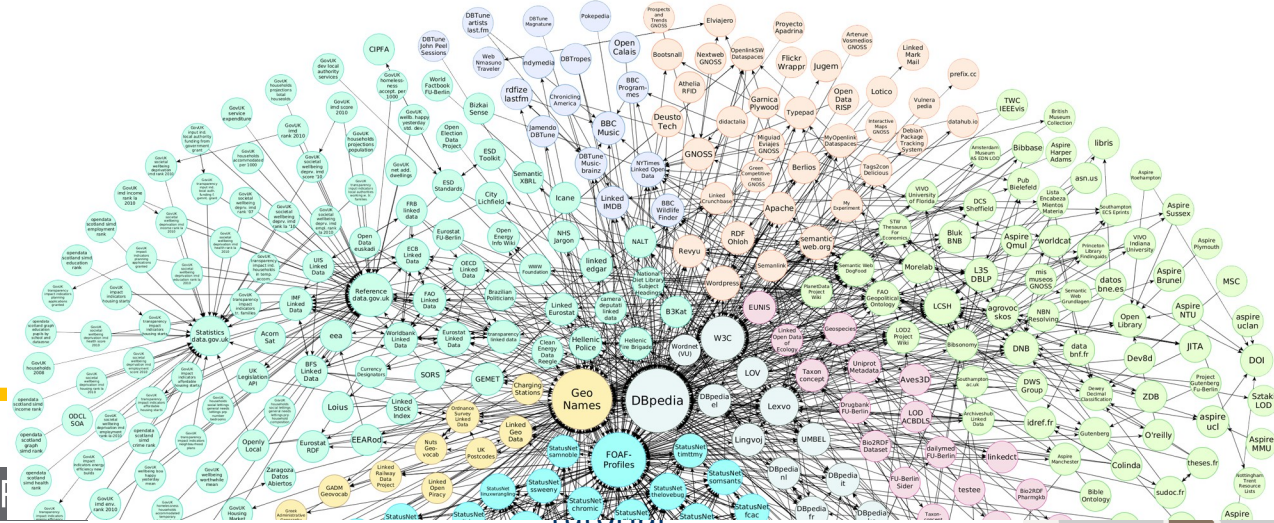
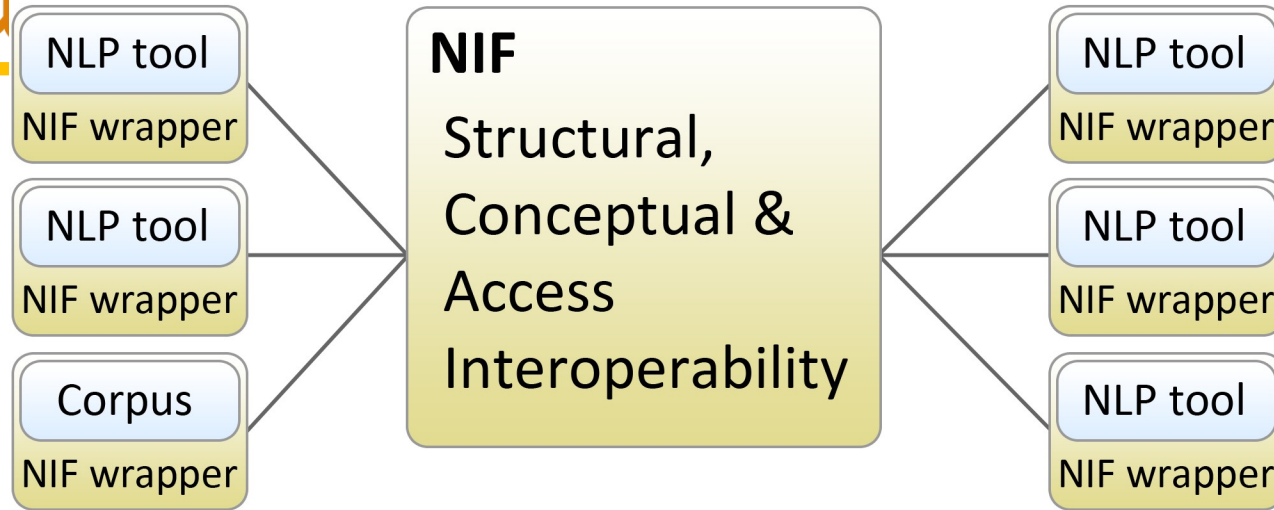


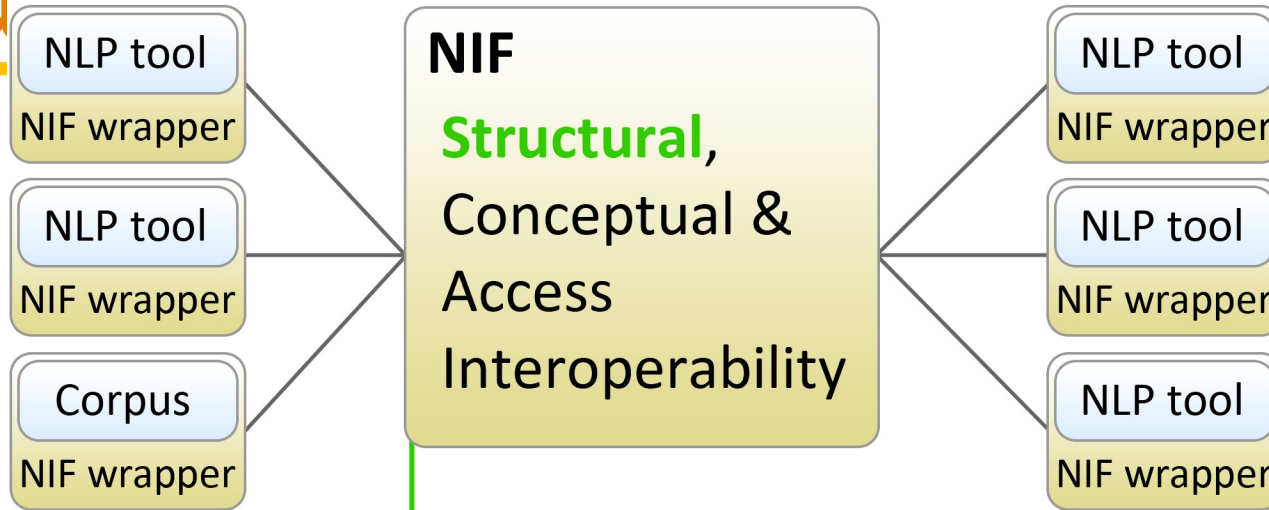




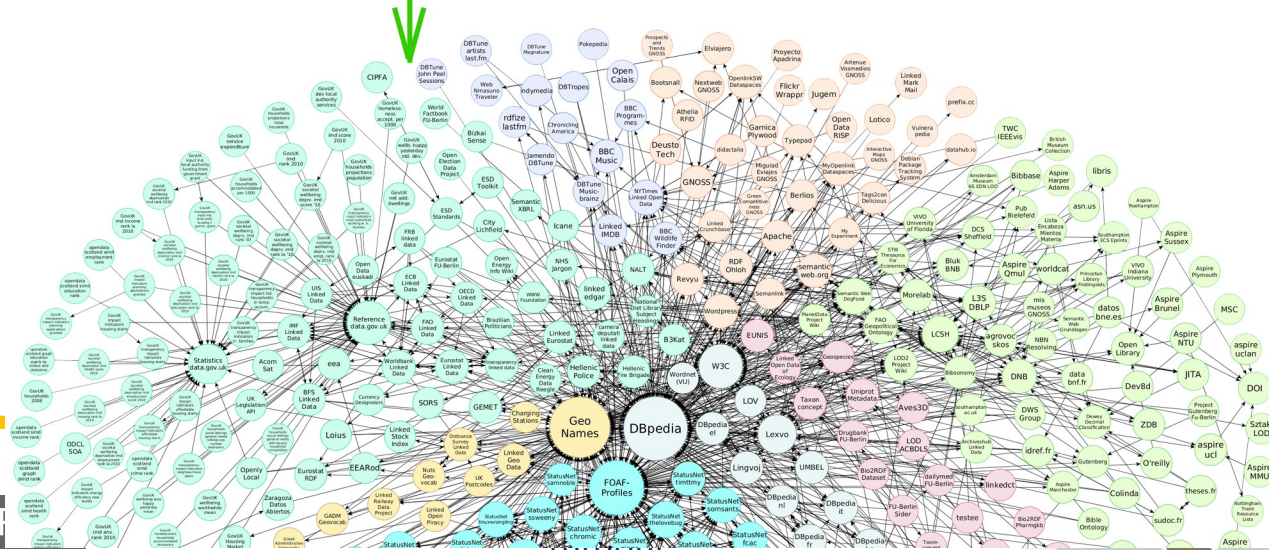








Cross
Linking



“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".
```

Tokenizer

```
<#char=3,12>
```

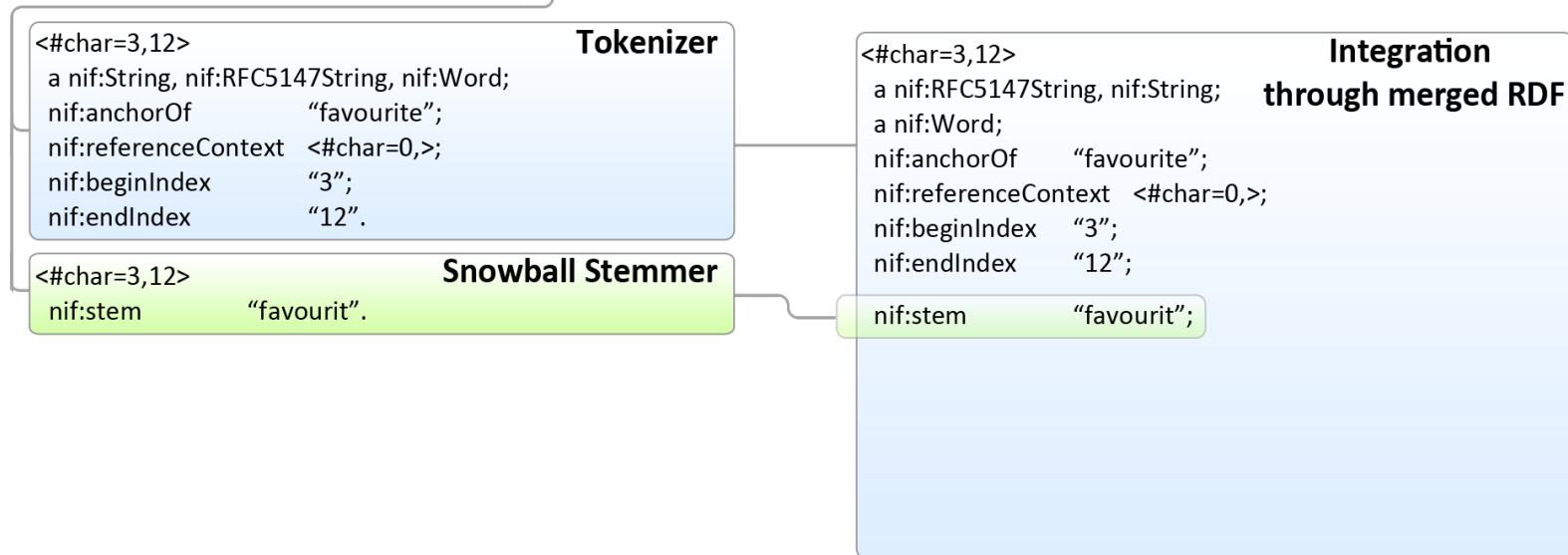
```
a nif:RFC5147String, nif:String;
a nif:Word;
nif:anchorOf      "favourite";
nif:referenceContext <#char=0,>;
nif:beginIndex    "3";
nif:endIndex       "12".
```

Integration

through merged RDF

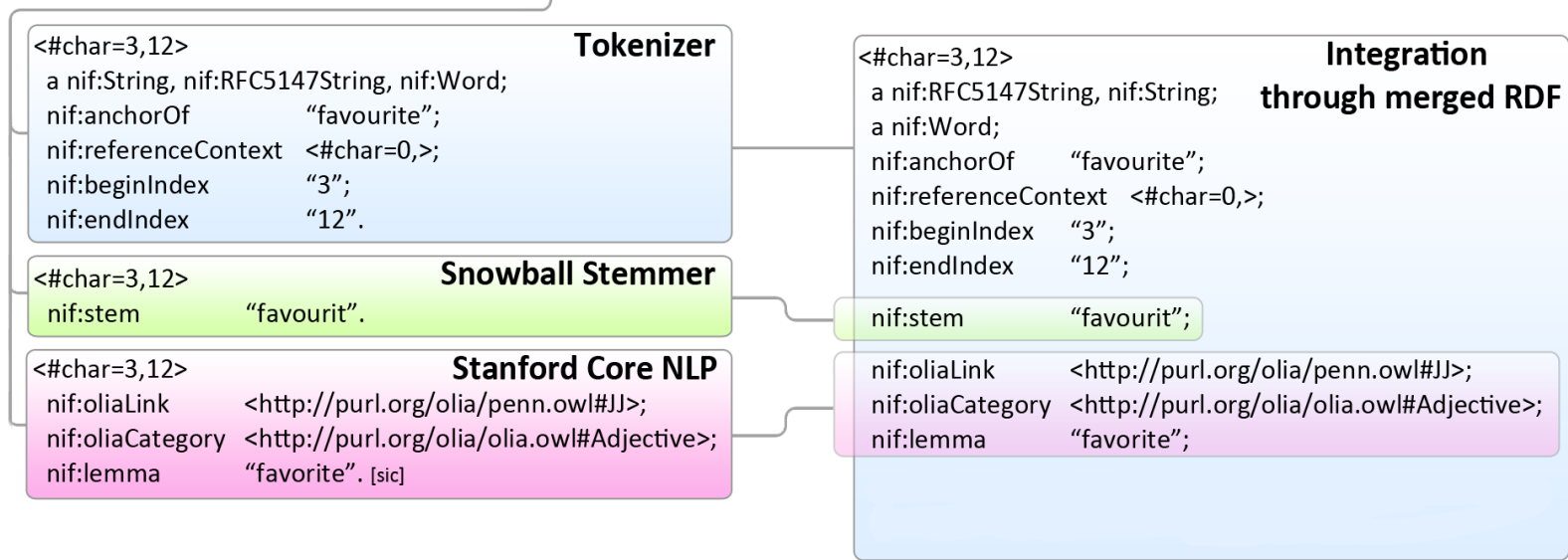
@base <http://example.org/prefix>

"My favourite actress is Natalie Portman."



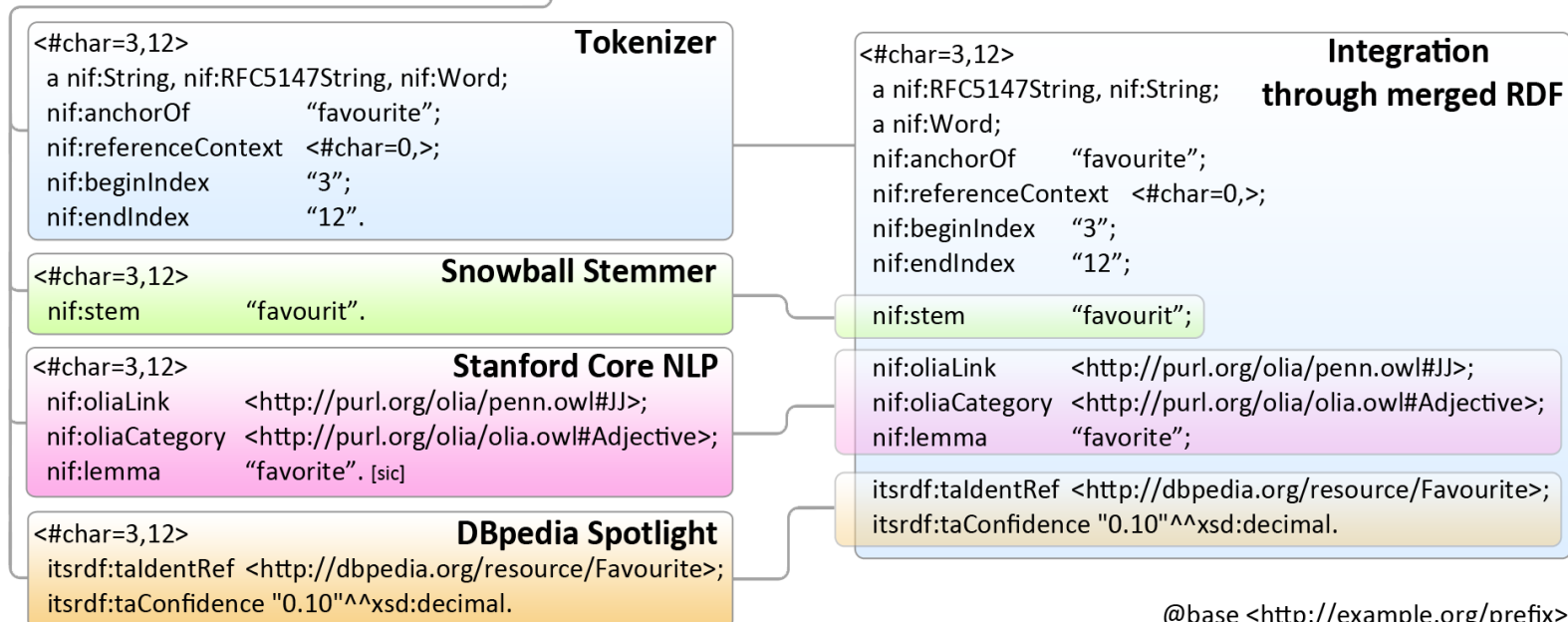
@base <http://example.org/prefix>

“My favourite actress is Natalie Portman.”



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

"My favourite actress is Natalie Portman."



- Corpus contains hotel reviews and review metadata
- 1760 semi-structured files
- In NIF: every file's content becomes one nif:Context resource
- Strings in the file can be addressed via URIs

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

<http://example.org/tripadvisor/hotel_100547.dat#char=0,125958>
  a nif:String, nif:RFC5147String, nif:Context;
  nif:isString
  "" "<Overall Rating>2.5 <Avg. Price>$171 [...]" ""^^xsd:string;
  nif:beginIndex "0"^^xsd:nonNegativeInteger ;
  nif:endIndex "125958"^^xsd:nonNegativeInteger ;
  nif:sourceUrl <http://sifaka.cs.uiuc.edu/~wang296/Data/LARA/TripAdvisor/Review_Texts.zip> .
```

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

<http://example.org/tripadvisor/hotel_100547.dat#char=0,125958>
  a nif:String, nif:RFC5147String, nif:Context;
  nif:isString
  "" "<Overall Rating>2.5 <Avg. Price>$171 [...]" "" ^^xsd:string;
  nif:beginIndex "0" ^^xsd:nonNegativeInteger ;
  nif:endIndex "125958" ^^xsd:nonNegativeInteger ;
  nif:sourceUrl <http://sifaka.cs.uiuc.edu/~wang296/Data/LARA/TripAdvisor/Review_Texts.zip> .
```

- Address the content of a document
- nif:isString contains document content

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

<http://example.org/tripadvisor/hotel_100547.dat#char=0,125958>
  a nif:String, nif:RFC5147String, nif:Context;
  nif:isString
  """<Overall Rating>2.5 <Avg. Price>$171 [...]"""^^xsd:string;
  nif:beginIndex "0"^^xsd:nonNegativeInteger ;
  nif:endIndex "125958"^^xsd:nonNegativeInteger ;
  nif:sourceUrl <http://sifaka.cs.uiuc.edu/~wang296/Data/LARA/TripAdvisor/Review_Texts.zip> .
```

- Address the content of a document
- nif:isString contains document content
- Note that in NIF the document is != content of the document
- two different documents can have the same content
=> must not have the same URI

```
<http://example.org/tripadvisor/hotel_100547.dat#char=9010,9529>
  a nif:String, nif:RFC5147String ;
  nif:anchorOf
  """<Author>kekeScotland
<Content>Good Exeprience We got a good deal compared to other Seattle prices and as we we
hotel as a base so we didn't expect too much. Location of the hotel was good and the room
clean. We had a safe and free wi-fi. Only downside was the room was on the small side and
tiny.[...] """^^xsd:string;
  nif:beginIndex "9010"^^xsd:nonNetgativeInteger ;
  nif:endIndex "9529"^^xsd:nonNetgativeInteger ;
  nif:referenceContext <http://example.org/tripadvisor/hotel_100547.dat#char=0,125958> .
```

```
<http://example.org/tripadvisor/hotel_100547.dat#char=9010,9529>
  a nif:String, nif:RFC5147String ;
  nif:anchorOf
  """<Author>kekeScotland
<Content>Good Exeprience We got a good deal compared to other Seattle prices and as we we
hotel as a base so we didn't expect too much. Location of the hotel was good and the room
clean. We had a safe and free wi-fi. Only downside was the room was on the small side and
tiny.[...] """^^xsd:string;
  nif:beginIndex "9010"^^xsd:nonNetgativeInteger ;
  nif:endIndex "9529"^^xsd:nonNetgativeInteger ;
  nif:referenceContext <http://example.org/tripadvisor/hotel_100547.dat#char=0,125958> .
```

- Address arbitrary strings in the document
- Use string offsets in relation to context to adress strings
- nif:anchorOf contains the string
- Additional properties can now be added to the string


```
<http://example.org/tripadvisor/hotel_100547.dat#char=9010,9529>
  a nif:String, nif:RFC5147String ; a tripadvisor:Review ;
  nif:anchorOf
  """<Author>kekeScotland
<Content>Good Exeprience We got a good deal compared to other Seattle prices and as we we
hotel as a base so we didn't expect too much. Location of the hotel was good and the room
clean. We had a safe and free wi-fi. Only downside was the room was on the small side and
tiny.[...] """^^xsd:string;
  nif:beginIndex "9010"^^xsd:nonNetgativeInteger ;
  nif:endIndex "9529"^^xsd:nonNetgativeInteger ;
  nif:referenceContext <http://example.org/tripadvisor/hotel_100547.dat#char=0,125958> .
```

- Address arbitrary strings in the document
- Use string offsets in relation to context to adress strings
- nif:anchorOf contains the string
- Additional properties can now be added to the string

```
<http://example.org/tripadvisor/hotel_100547.dat#char=9041,9045>
  a nif:String, nif:RFC5147String, nif:Word;
  nif:anchorOf
  "" "Good" "" ^^xsd:string;
  nif:beginIndex "9041" ^^xsd:nonNegativeInteger ;
  nif:endIndex "9045" ^^xsd:nonNegativeInteger ;
  nif:referenceContext <http://example.org/tripadvisor/hotel_100547.dat#char=0,125958> .

<http://example.org/tripadvisor/hotel_100547.dat#char=9041,9056>
  a nif:String, nif:RFC5147String, nif:Phrase;
  nif:anchorOf
  "" "Good Exeprience" "" ^^xsd:string;
  nif:beginIndex "9041" ^^xsd:nonNegativeInteger ;
  nif:endIndex "9056" ^^xsd:nonNegativeInteger ;
  nif:referenceContext <http://example.org/tripadvisor/hotel_100547.dat#char=0,125958> .
```

Sentiment values, POS tags and other annotations etc can now be added to words and phrases

```

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

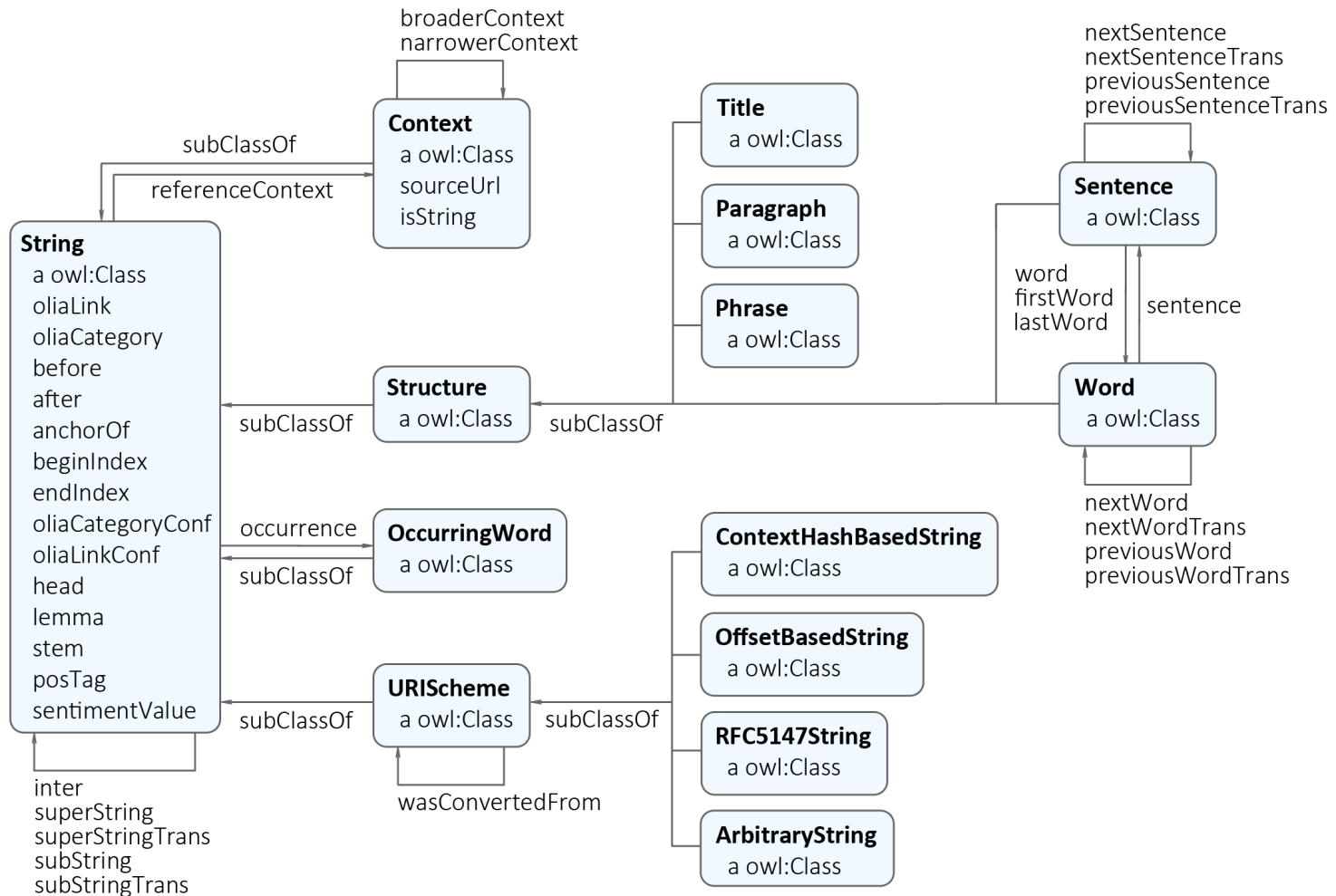
```

The anchors for each word are the following:

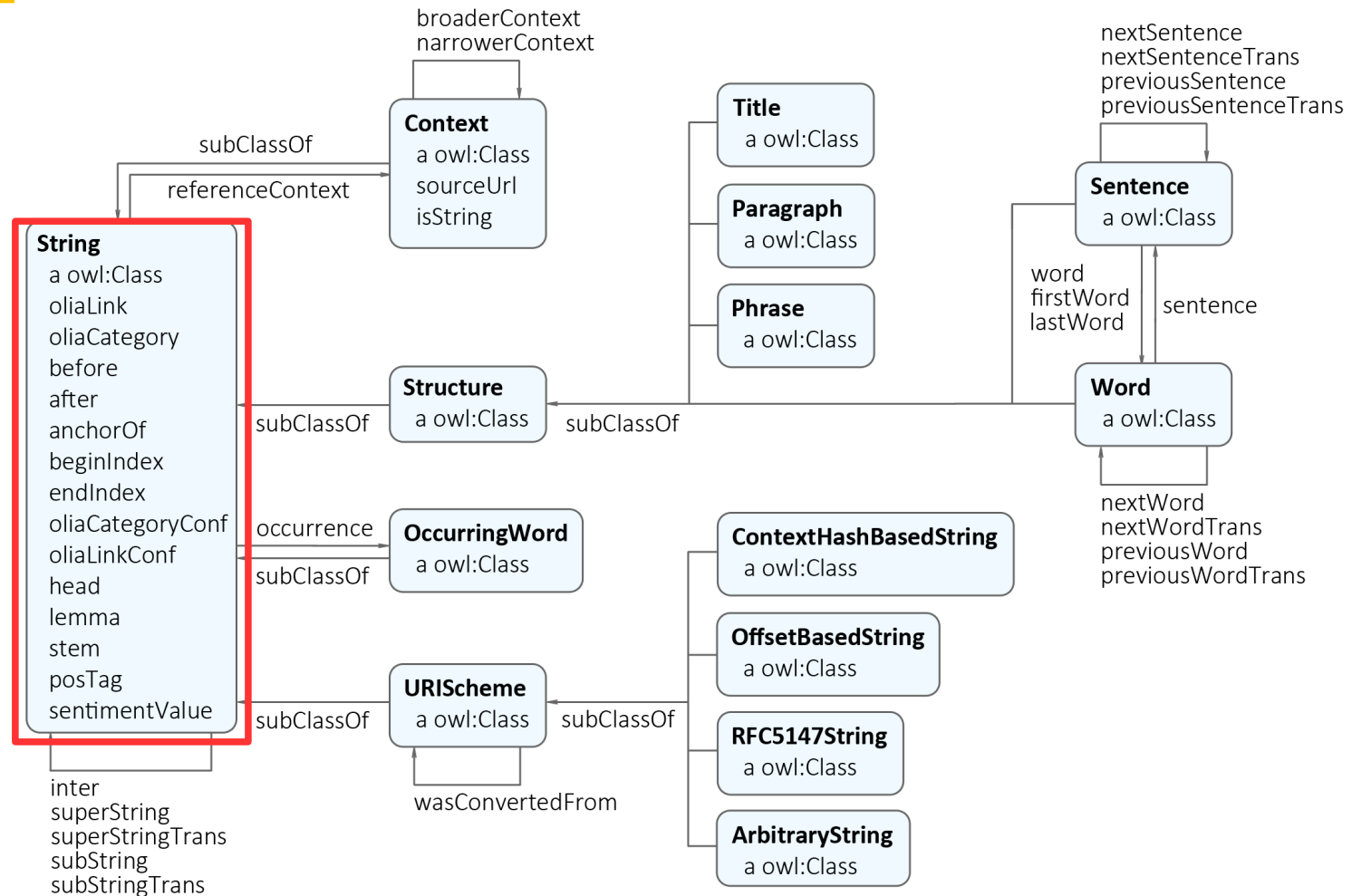
```

My: start=0, end=2
dog: start=3, end=6
has: start=7, end=10
fleas: start=11, end=16

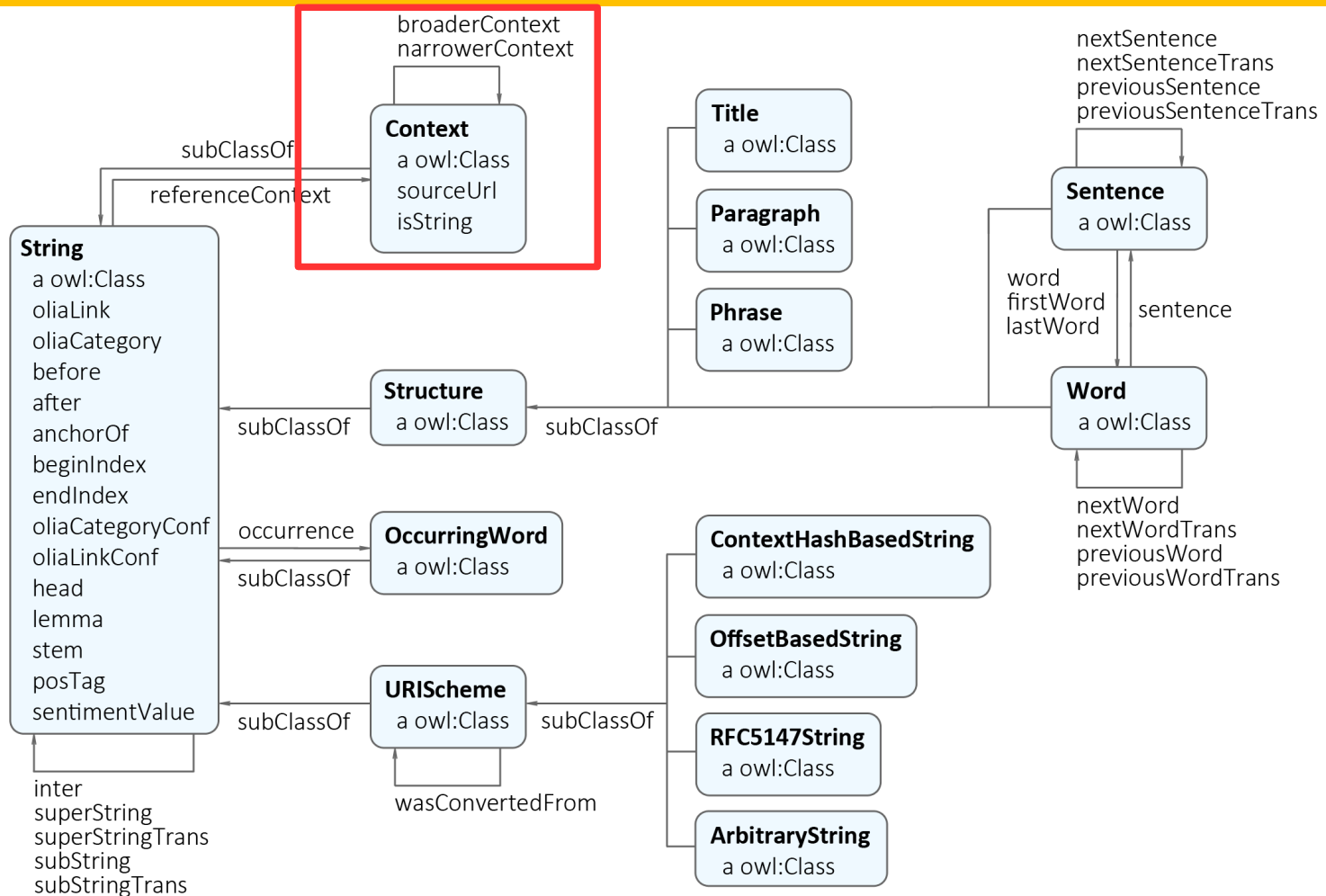
```



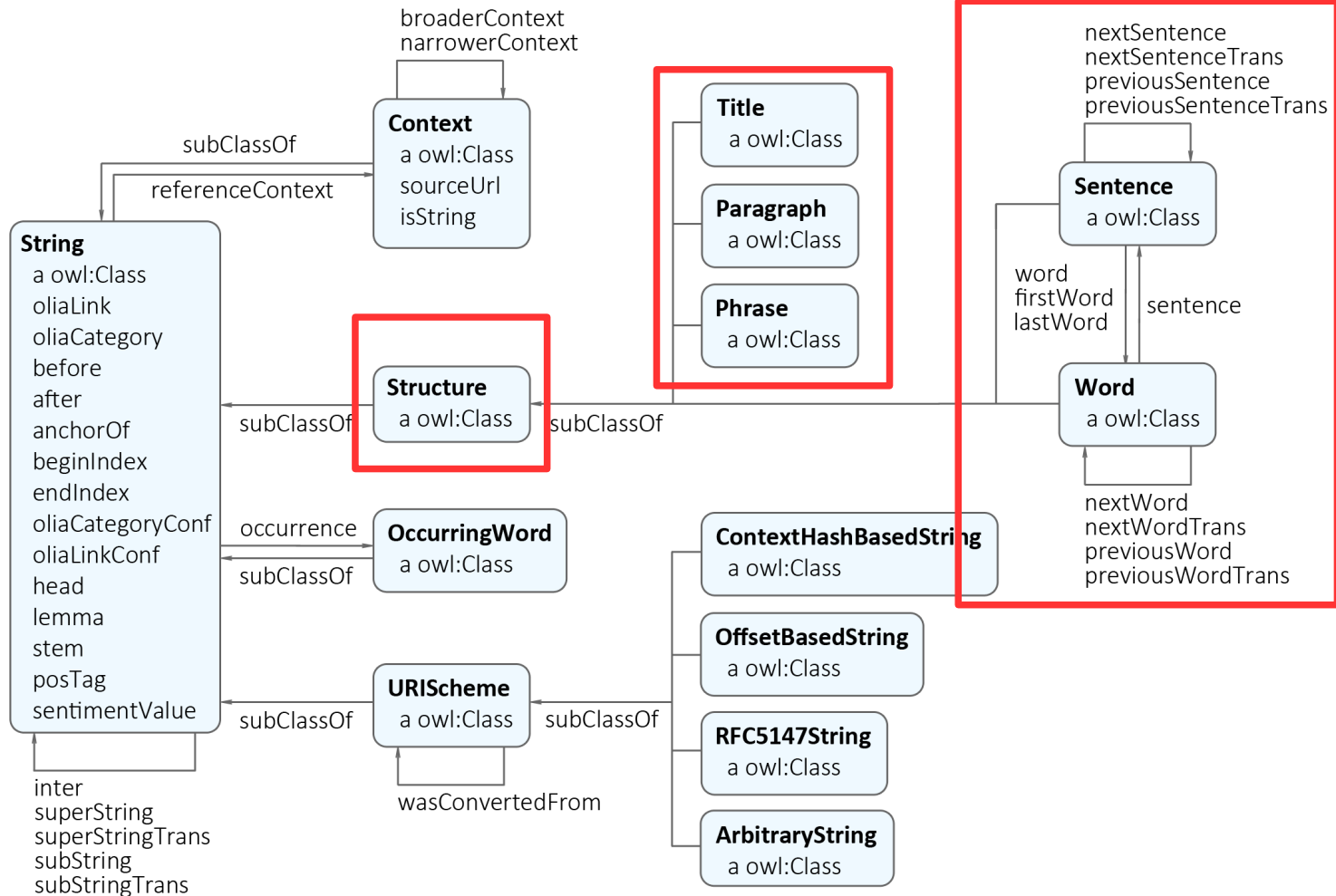
Namespace nif: <<http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>>



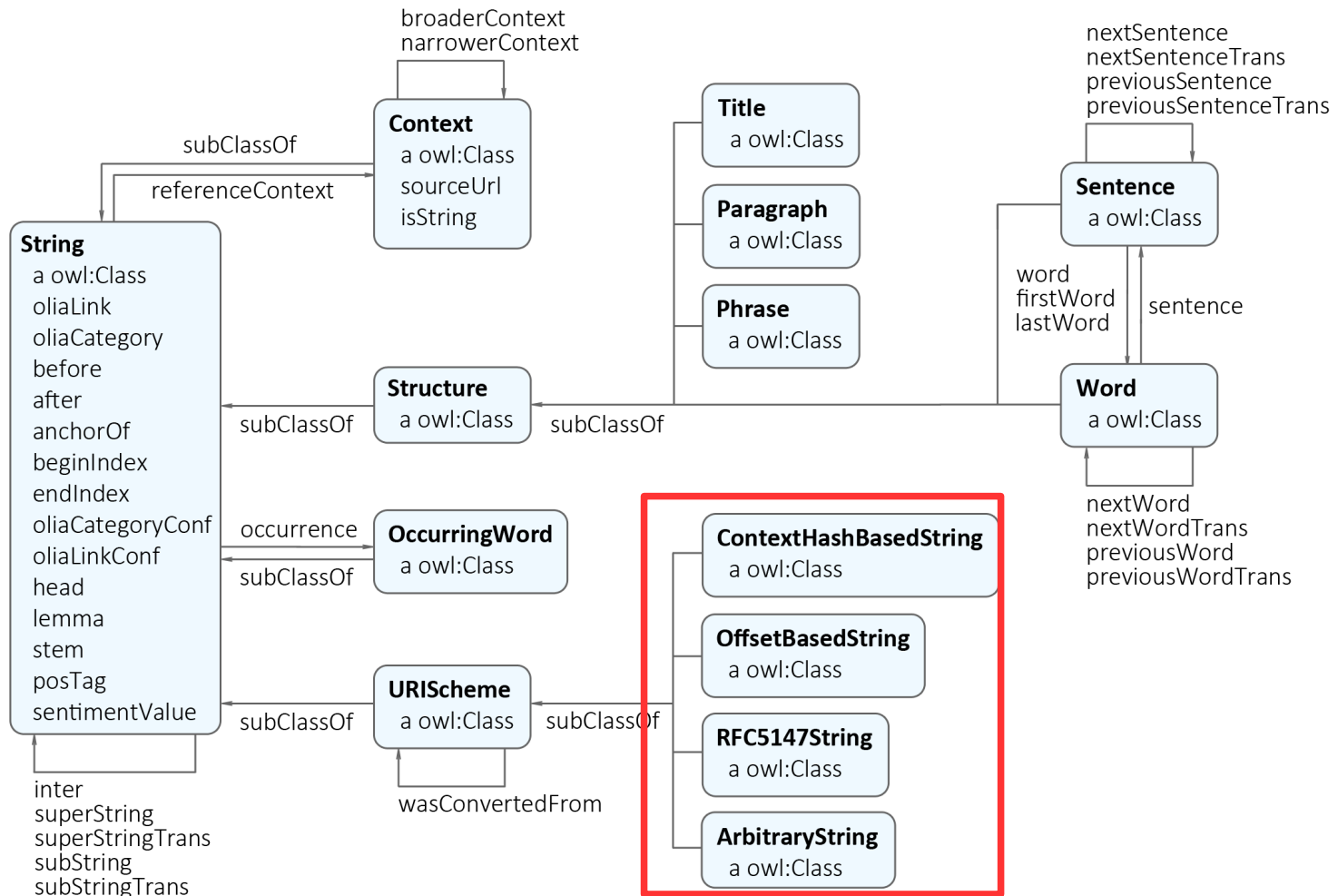
Namespace nif: <<http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>>



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



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



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

NIF Combinator: Combining NLP Tool Output

This web site show cases and combines available demo nif web services. Note that the list displayed here is not complete. More tools will be added soon. Read the [full documentation](#) [Bibtext]. The code for this demo is available [here](#). This service also allows you to retrieve the raw data. Use the *format* parameter which supports *rdxml*, *turtle*, *rdjson*, *ntriples* as values. You can also use the *Accept*: header with the appropriate content-type. Note that this service is currently using NIF-2.0 specification.

Action

- ☒ display and merge
- ☐ validate
- ☐ summarize

URI Recipe

- ☒ offset
- ☐ context-hash

Prefix

<http://prefix.given.by/theClient#>

Tools

- ☒ [Snowball Stemmer](#)
- ☐ [Stanford CoreNLP](#)
- ☐ [OpenNLP](#)
- ☐ [DBpedia Spotlight](#) (Confidence:)

Other NIF service:

Coming soon:

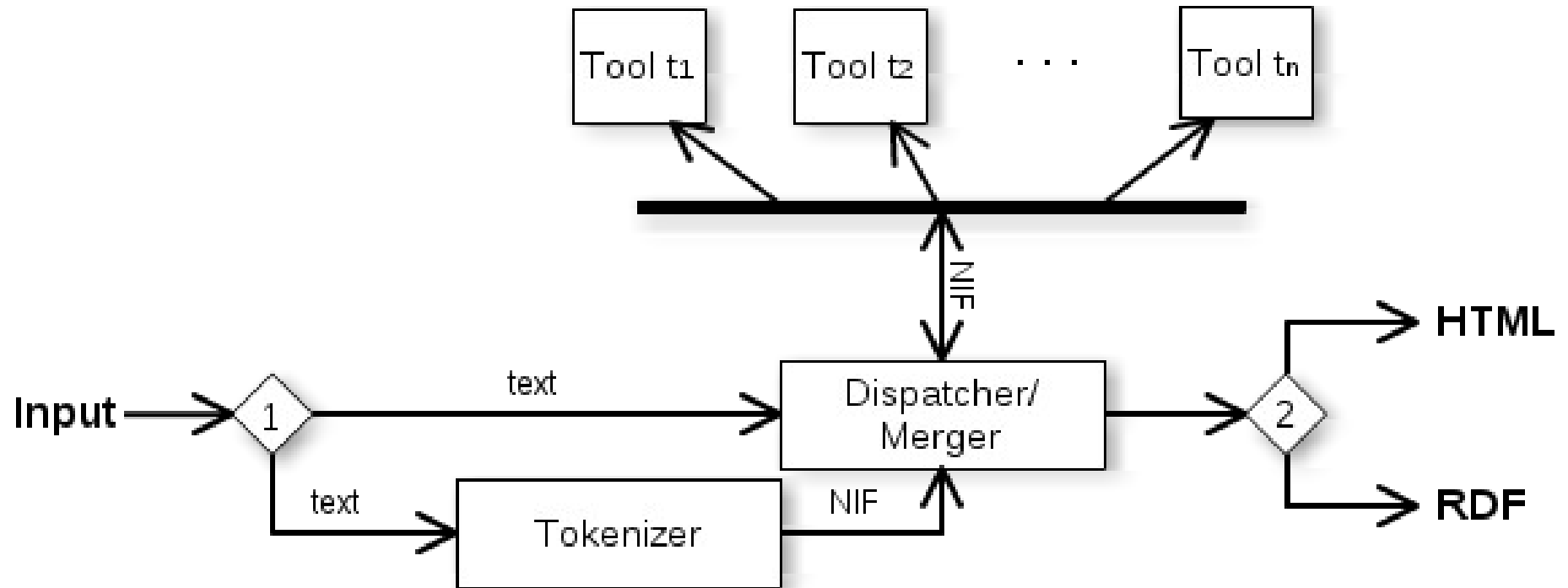
- ☐ [UIMA](#)
- ☐ [Mallet](#)
- ☐ [Gate ANNIE](#)
- ☐ [DBpedia Spotlight](#)



Input text

President Obama on Monday will call for a new minimum tax rate for individuals making more than \$1 million a year.

Merged Output



1. Introduction
2. NIF Basics
- 3. NIF corpora**
4. NIF tools & services

Name	Size (Triple)	URL
Wikilinks	500M	
News-100	13K	
RSS-500	10K	
Reuters-128	7K	
Spotlight	3K	
KORE50	2K	
Brown	500K	

Wikipedia abstract corpus in progress

<http://datahub.io/dataset?tags=nif&q=nif> : Tag „nif“ on datahub

- Large scale coreference resolution corpus by Umass / Google
- Over 10M crawled websites that contain text (Named Entities) that link to Wikipedia
- Converted to the NIF format and published as LOD:
<http://wiki-link.nlp2rdf.org/>
- Additional processing done to extract relevant text snippets, add Dbpedia ontology classes and coarse-grained classes (entity types)
- Over 500 million triples, 79GB LOD, 12GB gzipped dumps
- Over 30 million links to over 3 million entities

- Converted to the NIF format and published as LOD:
<http://brown.nlp2rdf.org>
- Corpus showcases handling of POS tags in NIF
- POS tags: mapped via OliA to predefined categories

```
<#char=643,647>
  a nif:String , nif:Word , nif:RFC5147String ;
  nif:anchorOf "Jury"^^xsd:string ;
  nif:referenceContext <#char=0,> ;
  nif:oliaLink brown:NN ;
  nif:sentence <#char=619,777> ;
  nif:beginIndex "643"^^xsd:nonNegativeInteger ;
  nif:endIndex "647"^^xsd:nonNegativeInteger .
```

- Categories can be used to query all resources of a certain POS regardless of tagset used in the corpus

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

<http://brown.nlp2rdf.org>

Query Text

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

select ?word, ?oliaPos, ?uri where {
  ?uri nif:oliaLink ?oliaPos.
  ?uri nif:anchorOf ?word.
  ?oliaPos rdf:type/rdfs:subClassOf* olia:Noun} Limit 100
```

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

<http://brown.nlp2rdf.org>

Query Text

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

select ?word, ?oliaPos, ?uri where {
  ?uri nif:oliaLink ?oliaPos.
  ?uri nif:anchorOf ?word.
  ?oliaPos rdf:type/rdfs:subClassOf* olia:Noun} Limit 100
```

- Querying all nouns using the OIa mapping

word	oliaPos	uri
"Henry"^^<http://www.w3.org/2001/XMLSchema#string>	http://purl.org/olia/brown.owl#NP	http://brown.nlp2rdf.org/linkeddata.php?t=url&f=xml&i=http://brown.nlp2rdf.org/corpus/a01.xml#char=5755,5760
"Fulton"^^<http://www.w3.org/2001/XMLSchema#string>	http://purl.org/olia/brown.owl#NP	http://brown.nlp2rdf.org/linkeddata.php?t=url&f=xml&i=http://brown.nlp2rdf.org/corpus/a01.xml#char=2767,2773
"W."^^<http://www.w3.org/2001/XMLSchema#string>	http://purl.org/olia/brown.owl#NP	http://brown.nlp2rdf.org/linkeddata.php?t=url&f=xml&i=http://brown.nlp2rdf.org/corpus/a01.xml#char=6510,6512
"Atlanta"^^<http://www.w3.org/2001/XMLSchema#string>	http://purl.org/olia/brown.owl#NP	http://brown.nlp2rdf.org/linkeddata.php?t=url&f=xml&i=http://brown.nlp2rdf.org/corpus/a01.xml#char=955,962
"Bellwood"^^<http://www.w3.org/2001/XMLSchema#string>	http://purl.org/olia/brown.owl#NP	http://brown.nlp2rdf.org/linkeddata.php?t=url&f=xml&i=http://brown.nlp2rdf.org/corpus/a01.xml#char=4977,4985
"Georgia"^^<http://www.w3.org/2001/XMLSchema#string>	http://purl.org/olia/brown.owl#NP	http://brown.nlp2rdf.org/linkeddata.php?t=url&f=xml&i=http://brown.nlp2rdf.org/corpus/a01.xml#char=6206,6213
"Tabb"^^<http://www.w3.org/2001/XMLSchema#string>	http://purl.org/olia/brown.owl#NP	http://brown.nlp2rdf.org/linkeddata.php?t=url&f=xml&i=http://brown.nlp2rdf.org/corpus/a01.xml#char=13059,13063

- Querying all nouns using the OIa mapping

1. Introduction
2. NIF Basics
3. NIF corpora
- 4. NIF tools & services**

Available NIF tools :

- Stanford Core NLP
- OpenNLP
- RDFace
- Validator
- ConLL converter ...



Confidence:
0.0

Language:
English

Annotation score:
0.0

☐ n-best candidates

SELECT TYPES...

ANNOTATE

First documented in the 13th century, [Berlin](#) was the capital of the Kingdom of [Prussia](#) (1701–1918), the [German Empire](#) (1871–1918), the [Weimar Republic](#) (1919–33) and the [Third Reich](#) (1933–45). [Berlin](#) in the 1920s was the third largest m http://dbpedia.org/resource/Weimar_Republic [War II](#), the city became divided into [East Berlin](#) -- the capital of [East Germany](#) -- and [West Berlin](#), a [West German exclave](#) surrounded by the [Berlin Wall](#) from 1961–89. Following [German reunification](#) in 1990, the city regained its status as the capital of [Germany](#), hosting 147 foreign embassies.

BACK TO TEXT

<https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki>

```
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/
  ontologies/nif-core#> .
# POS tags produced by Stanford Parser online demo
# http://nlp.stanford.edu:8080/parser/index.jsp
<http://nlp2rdf.lod2.eu/nif-ws.php#char=13,20>
  rdf:type nif:RFC5147String ;
  nif:anchorOf "actress" ;
  nif:beginIndex "13" ;
  nif:endIndex "20" ;
  nif:referenceContext <http://nlp2rdf.lod2.eu/nif-ws.php#char
    =0,40> ;
# provenance link
nif:oliaProv <http://nlp.stanford.edu:8080/parser/index.jsp> ;
# the tag
nif:oliaLink <http://purl.org/olia/penn.owl#NN> ;
# mappings can be found here: https://github.com/NLP2RDF/
  software
nif:oliaCategory <http://purl.org/olia/olia.owl#Noun> ,
  <http://purl.org/olia/olia.owl#CommonNoun> .
```

```
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/
  ontologies/nif-core#> .
# POS tags produced by Stanford Parser online demo
# http://nlp.stanford.edu:8080/parser/index.jsp
<http://nlp2rdf.lod2.eu/nif-ws.php#char=13,20>
  rdf:type nif:RFC5147String ;
  nif:anchorOf "actress" ;
  nif:beginIndex "13" ;
  nif:endIndex "20" ;
  nif:referenceContext <http://nlp2rdf.lod2.eu/nif-ws.php#char
    =0,40> ;
# provenance link
nif:oliaProv <http://nlp.stanford.edu:8080/parser/index.jsp> ;
# the tag
nif:oliaLink <http://purl.org/olia/penn.owl#NN> ;

nif:oliaCategory <http://purl.org/olia/olia.owl#Noun> ,
  <http://purl.org/olia/olia.owl#CommonNoun> .
```

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

# POS tags produced by Stanford Parser online demo
# http://nlp.stanford.edu:8080/parser/index.jsp
<http://nlp2rdf.lod2.eu/nif-ws.php#char=24,39>
  rdf:type nif:NamedEntity ;
  nif:anchorOf "Natalie Portman" ;
  nif:beginIndex "24";
  nif:endIndex "33" ;
→ itsrdf:taIdentRef <http://dbpedia.org/resource/Natalie_Portman>
  .
```

```
<http://nlp2rdf.lod2.eu/nif-ws.php#char=24,39>
  itsrdf:taClassRef <http://dbpedia.org/ontology/Actor> ;
  itsrdf:taClassRef <http://dbpedia.org/ontology/Artist> ;
  itsrdf:taClassRef <http://dbpedia.org/ontology/Person> ;
  itsrdf:taClassRef <http://xmlns.com/foaf/0.1/Person> .
```



```
<http://nlp2rdf.lod2.eu/nif-ws.php#char=24,39>
  itsrdf:taIdentRef <http://dbpedia.org/resource/Natalie_Portman>
    .
  itsrdf:taIdentRef <http://rdf.freebase.com/ns/m.09l3p> .
# entails that
<http://dbpedia.org/resource/Natalie_Portman>
  owl:sameAs <http://rdf.freebase.com/ns/m.09l3p> .
```

```
<http://nlp2rdf.lod2.eu/nif-ws.php#char=24,39>
  itsrdf:taIdentRef <http://dbpedia.org/resource/Natalie_Portman
    > ;
  itsrdf:taClassRef <http://dbpedia.org/ontology/Actor> .
# the tool states that the type found is a general property of
  the found entity beyond the context of the current nif:
  Context
<http://dbpedia.org/resource/Natalie_Portman>
  rdf:type <http://dbpedia.org/ontology/Actor> .
```

NIF 2.0 Dashboard.

This dashboard reflects the actual state of resources of the NIF 2.0 project. Here you can also find documentation, corpora, ontologies, demos, wikis and specifications.

Online REST Webservices		
Service Name	URL	Status
OpenNLP	http://demo.nlp2rdf.org:9998/opennlp?f=text&i=This+is+a+test.&t=direct&modelFolder=model	Online!
Stanford	http://demo.nlp2rdf.org:9999/stanfordcorenlp?f=text&i=This+is+a+test.&t=direct	Online!
Snowball	http://demo.nlp2rdf.org:9996/snowball?f=text&i=This+is+my+favorite+test.&t=direct	Online!
NIF Validator	http://demo.nlp2rdf.org:9990	Online!
NIF DBpedia Spotlight	http://demo.nlp2rdf.org:9995/spotlight?f=text&i=This+is+Germany.&t=direct&confidence=0.3	Online!
NIF Combinator	http://demo.nlp2rdf.org	Online!

NIF Corpus		
Corpus Name	URL	DataID File
Add Corpus		

Ontology	
NIF Validator TestCase Suite@en	Details
STC - Simple Test Case Ontology@en	Details
RLOG - an RDF Logging Ontology@en	Details
NIF 2.0 Core Ontology - Validation Model@en	Details
NIF 2.0 Core Ontology - Inference Model@en	Details
NIF 2.0 Core Ontology - Terminological	

NIF Specifications	
W3C Conversion of NIF to ITS 2.0@en	Details
W3C Conversion of ITS 2.0 to NIF@en	Details
NIF 2.0 Core Specification@en	Details
NIF 2.0 Stanbol Profile Specification@en	Details
Public API Specification@en	Details

Wikis	
How to publish a txt corpora with NIF as Linked Data@en	Details

Poster	
Promotional poster for Meta Forum@en	Details

Software	
DBpedia Spotlight implementation of NIF@en	Details
NIF 2.0 Validator Tool@en	Details
NIF 2.0 Stanford Core Implementation@en	Details

Demo	
NIF Conversion Demo Service@en	Details
NIF Combinator Demo@en	Details

Thank you very much!

<http://nlp2rdf.org>
<http://github.com/NLP2RDF>