# Linked Data Corpus Creation with NIF

This document describes best practices to follow for the generation of Linked Data text corpora, using the NLP Interchange Format (NIF).

## Target audience

Corpus creators and users seeking to make corpora interoperable and to publish them as linked data. Basic knowledge of RDF is mandatory for conversion. Basic knowledge of linked data and web server access is needed for publication.

## Scope

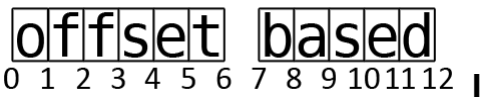
Conversion of existing corpora into RDF using NIF, as well as creation of linked data corpora from textual data.

## Core concepts

**Corpus**

We understand a corpus as a collection of documents. Documents contain text, represented as strings of characters and annotations that provide more information about these strings. NIF provides a way to identify strings using URIs and annotate them using an ontology.

**String identification**




Strings are identified using a URI scheme consisting of: the URI of the document itself; the **character indices** of beginning and end of the string; and a separator between document URI and string position identifier. Character indices in NIF are counted *offset based*, starting at zero before the first character and counting the gaps between the characters until after the last character of the referenced string:

http://example.org/corpus/document#char=**4,10**

**This URI scheme is valid for text/plain.** Other mime types may require different URI schemes.

**String annotation**

After assigning URIs to meaningful strings of the corpus, these URIs can be annotated using the NIF core ontology (see page 2).

# Example

**Document** The Semantic Web is a good idea.

## Context

- Contains document text in nif:isString
- nif:beginIndex is always 0

```
<http://example.org/sem#char=0,32>
 a nif:String , nif:Context , nif:RFC5147String ;
 nif:isString "The Semantic Web is a good idea."@en ;
 nif:beginIndex "0"^^xsd:nonNegativeInteger ;
 nif:endIndex "32"^^xsd:nonNegativeInteger .
```

## Sentence

- Contains the string in nif:anchorOf
- refers to Context with nif:referenceContext

```
<http://example.org/sem#char=0,32>
 a nif:String , nif:Sentence , nif:RFC5147String ;
 nif:anchorOf "The Semantic Web is a good idea."@en ;
 nif:beginIndex "0"^^xsd:nonNegativeInteger ;
 nif:endIndex "32"^^xsd:nonNegativeInteger ;
 nif:referenceContext
   <http://example.org/sem#char=0,32> .
```

## Words, Phrases

- Contain the string in nif:anchorOf
- refers to Context with nif:referenceContext
- POS tags mapped via OLiA
- Entity references via itsrdf:taIdentRef

```
<http://example.org/sem#char=4,16>
 a nif:String , nif:Phrase , nif:RFC5147String ;
 nif:anchorOf "Semantic Web"@en ;
 nif:beginIndex "4"^^xsd:nonNegativeInteger ;
 nif:endIndex "16"^^xsd:nonNegativeInteger ;
 nif:oliaLink <http://purl.org/olia/penn.owl#NNP> ;
 itsrdf:taIdentRef
   <http://dbpedia.org/resource/Semantic_Web> ;
 nif:referenceContext
   <http://example.org/sem#char=0,32> .
```

Find a real world example at http://brown.nlp2rdf.org

## Namespaces and Ontologies

nif: http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#
olia: http://purl.org/olia
itsrdf: http://www.w3.org/2005/11/its/rdf#

Website: http://site.nlp2rdf.org
Github: http://github.com/nlp2rdf
Example corpus: http://brown.nlp2rdf.org

mylider