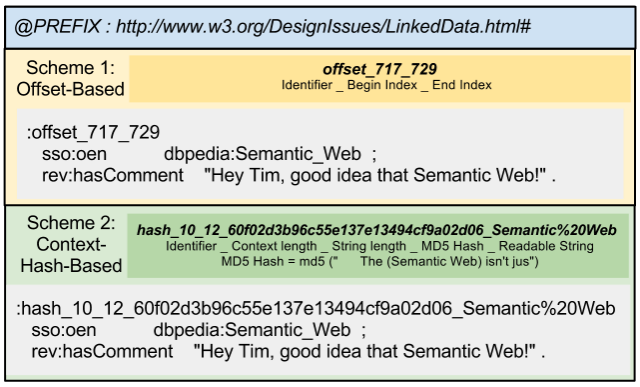

Figure 1: NIF URI schemes: Offset (top) and context-hashes (bottom) are used to create identifiers for strings

1. NIF: AN NLP INTERCHANGE FORMAT

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. The NIF specification has been released in an initial version 1.0 in November 2011 and describes how interoperability between NLP tools, which are exposed as NIF web services can be achieved. Extensive feedback was given on several mailing lists and a community of interest[1] was created to improve the specification. Implementations for 8 different NLP tools (e.g. UIMA, Gate ANNIE and DBpedia Spotlight) exist and a public web demo[2] is available.

In the following, we will first introduce the core concepts of NIF, which are defined in a String Ontology[3] (STR). We will then explain how NIF is used in NERD. The resulting properties and axioms are included into a Structured Sentence Ontology[4] (SSO). While the String Ontology is used to describe the relations between strings (i.e. Unicode characters), the SSO collects properties and classes to connect strings to NLP annotations and NER entities as produced by NERD.

1.1 Core Concepts of NIF

The motivation behind NIF is to allow NLP tools to exchange annotations about documents in RDF. Hence, the main prerequisite is that parts of the documents (i.e. strings) are referenceable by URIs, so that they can be used as subjects in RDF statements. We call an algorithm to create such identifiers *URI Scheme*: For a given text t (a sequence of characters) of length $|t|$ (number of characters), we are looking for a *URI Scheme* to create a URI, that can serve as a *unique* identifier for a substring s of t (i.e. $|s| \leq |t|$). Such a substring can (1) consist of adjacent characters only and it is therefore a unique character sequence within the text, if we account for parameters such as context and position or (2) derived by a function which points to several substrings as defined in (1).

NIF provides two URI schemes, which can be used to represent strings as RDF resources. We focus here on the first scheme using offsets. In the top part of Figure 1, two triples are given that use the following URI as subject:

`http://www.w3.org/DesignIssues/LinkedData.html#offset_717_729`

According to the above definition, the URI points to a substring of a given text t, which starts at character index 717 until the index 729 (counting all characters). NIF currently mandates that the whole string of the document has to be included in the RDF output as an `rdf:Literal` to serve as the reference point, which we will call *inside context* formalized using an OWL class called `str:Context`. The term *document* would be inappropriate to capture the real intention of this concept as we would like to refer to an arbitrary grouping of characters forming a unit, which could also be applied to a *paragraph* or a *sentence* and is highly dependent upon the *wider context* in which the string is actually used such as a Web document reachable via HTTP.

To appropriately capture the intention of such a class, we will distinguish between the notion of outside and inside context of a piece of text. The inside context is easy to explain and formalise, as it is the text itself and therefore it provides a reference context for each substring contained in the text (i.e. the characters before or after the substring). The outside context is more vague and is given by an outside observer, who might arbitrarily interpret the text as a "book chapter" or a "book section".

The class `str:Context` now provides a clear reference point for all other relative URIs used in this context and blocks the addition of information from a larger (outside) context by definition. By definition `str:Context` is disjoint with `foaf:Document` as labeling a context resource as a document is an information, which is not contained within the context (i.e. the text) itself. It is legal, however, to say that the string of the context occurs in (`str:occursIn`) a `foaf:Document`. Additionally, `str:Context` is a subclass of `str:String` and therefore its instances denote Unicode text as well. The main benefit to limit the context is that an OWL reasoner can now infer that two contexts are the same, if they consist of the same string, because an inverse-functional data type property (`str:isString`) is used to attach the actual text to the context resource.

```
:offset_0_26546  a  str:Context ;
#the exact retrieval method is left underspecified
  str:occursIn <http://www.w3.org/DesignIssues/
       LinkedData.html> ;
# [...] are all 26547 characters as rdf:Literal
    str:isString   "[...]" .
:offset_717_729  a  str:String ;
  str:referenceContext :offset_0_26546 .
```

A complete formalisation is still work in progress, but the idea is explained here. The NIF URIs will be grounded on Unicode Characters(especially Unicode Normalization Form C [5]. For all resources of type *str:String*, the universe of discourse will then be the powerset over the concatenation of Unicode characters. Perspectively, we hope that this will allow for an unambiguous interpretation of NIF by machines.

Within the framework of RDF and the current usage of NIF for the interchange of output between NLP tools, the definition of the semantics is sufficient to produce a working system. However, problems arise if additional interoperabil-

[1] `http://nlp2rdf.org/get-involved`
[2] `http://nlp2rdf.lod2.eu/demo.php`
[3] `http://nlp2rdf.lod2.eu/schema/string`
[4] `http://nlp2rdf.lod2.eu/schema/sso`
[5] `http://www.unicode.org/reports/tr15/#Norm_Forms` counted in Code Units `http://unicode.org/faq/char_combmark.html#7`

ity with Linked Data or fragment identifiers and ad-hoc retrieval of content from the Web is demanded. The actual retrieval method (such as content negotiation) to retrieve and validate the content for `#offset_717_729_Semantic%20Web` or its reference context is left underspecified as is the relation of NIF URIs to fragment identifiers for MIME types such as text/plain (see RFC 5147[6]). As long as such issues remain open, the complete text has to be included as RDF Literal.

1.2 Connecting String to Entities

For NERD, three relevant concepts have to be expressed in RDF and were included into the Structured Sentence Ontology (SSO): OEN, OED and NERD ontology types.

One Entity per Name (OEN) can be modeled in a straightforward way, by introducing a property `sso:oen`, which connects the string with an arbitrary entity.

```
:offset_717_729  sso:oen dbpedia:Semantic_Web .
```

One Entity per Document (OED). As *document* is an outside interpretation of a string, the notion of context in NIF has to be used. The property `sso:oec` is used to attach entities to a given context. We furthermore add the following DL-Axiom:

$$sso{:}oec \supseteq str{:}referenceContext^{-1} \circ sso{:}oen$$

As the property `oen` contains more specific information, `oec` can be inferred by the above role chain inclusion. In case the context is enlarged, any materialized information attached via the `oec` property needs to be migrated to the larger context resource.

The connection between NERD types and strings is done via a linked data URI, which disambiguates the entity. Overall three cases can be distinguished: In case, the NER extractor has provided a linked data URI to disambiguate the entity, we simply re-use it as in the following example:

```
# this URI points to the string "W3C"
:offset_23107_23110
  rdf:type              str:String ;
  str:referenceContext  :offset_0_26546 ;
  sso:oen               dbpedia:W3C ;
  str:beginIndex        "23107"  ;
  str:endIndex          "23110"  .
dbpedia:W3C rdf:type    nerd:Organization .
```

If, however, the NER extractor provides no disambiguation link at all or just a non-linked data URI for the entity (typically, the `foaf:homepage` of an organization such as *http://www.w3.org/*), we plan to mint a new linked data URI for the respective entity that could then be further sameAs with other identifiers in a data reconciliation process.

[6] `http://tools.ietf.org/html/rfc5147`