NLP Interchange Format (NIF)

http://nlp2rdf.org

Sebastian Hellmann

AKSW, Universität Leipzig

NLP2RDF + NIF

- NLP Interchange Format (NIF) is an RDF/OWL-based format that allows to combine and chain several Natural Language Processing (NLP) tools in a flexible, light-weight way.
- NLP2RDF is a LOD2 project providing:
  - documentation
  - reference implementations of NIF
  - collaboration platform
  - tutorials / example source code
  - mailing list for questions and support
  - possible to join on http://nlp2rdf.org

NLP2RDF + NIF

NLP2RDF + NIF

- Motivation and comparison of other NLP frameworks
- URI design
- NLP domain vocabularies
- Applications

NLP2RDF - NIF Use Cases

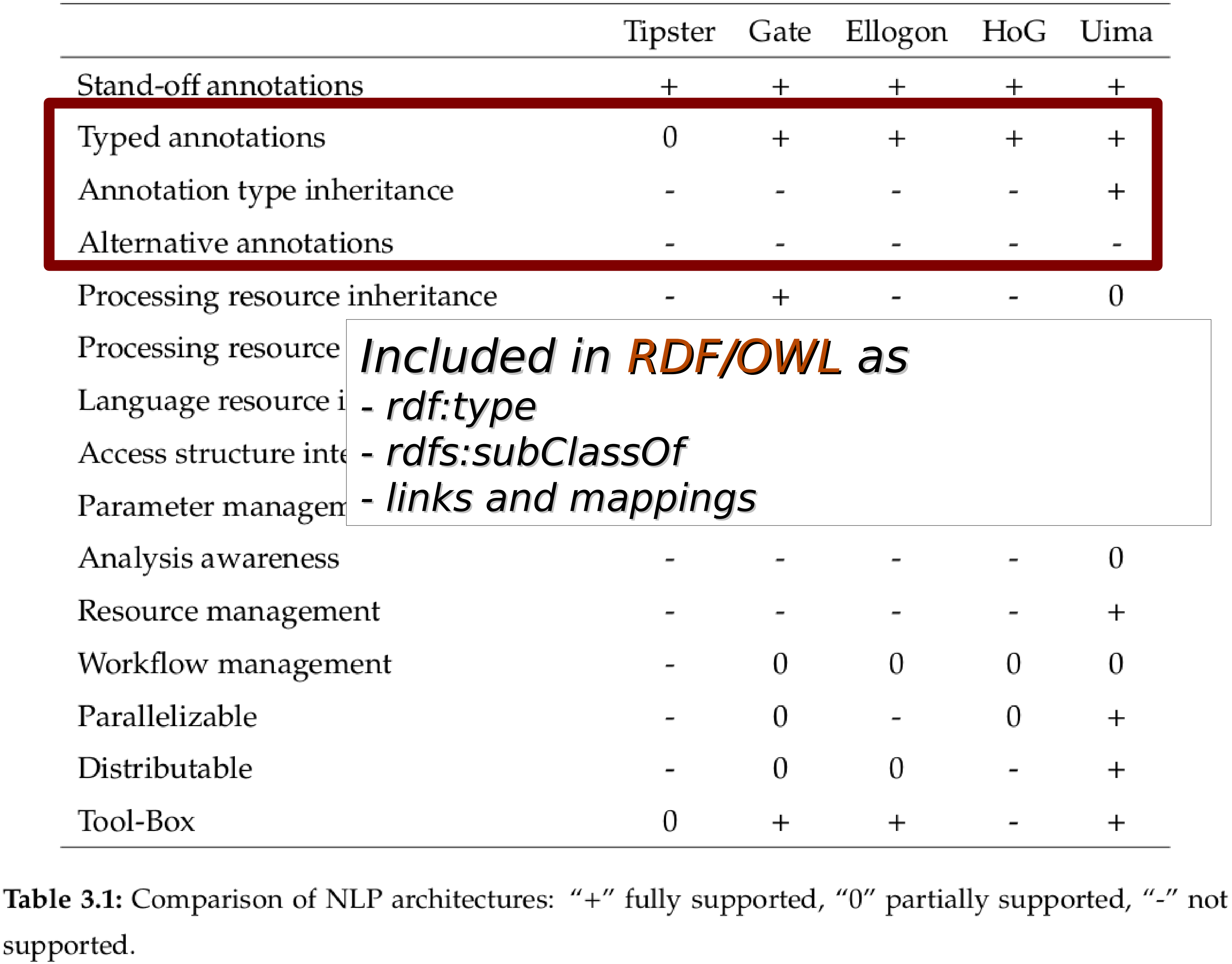
Problem: NLP software is organized in pipelines (UIMA, Gate)

- Integration is done „hard-wired“ (Software has to be developed)
- For each tool and each framework an adapter has to be created (n*m)
- No ad-hoc integration
- Difficult to aggregate output
- Difficult to exchange single components
- Not robust: if step 6 of 20 steps fails no output is produced

	Tipster	Gate	Ellogon	HoG	Uima
Stand-off annotations	+	+	+	+	+
Typed annotations	0	+	+	+	+
Annotation type inheritance	-	-	-	-	+
Alternative annotations	-	-	-	-	-
Processing resource inheritance	-	+	-	-	0
Processing resource interchangeability	0	+	+	+	+
Language resource interchangeability	-	0	-	-	-
Access structure interchangeability	-	0	-	-	-
Parameter management	-	+	+	+	+
Analysis awareness	-	-	-	-	0
Resource management	-	-	-	-	+
Workflow management	-	0	0	0	0
Parallelizable	-	0	-	0	+
Distributable	-	0	0	-	+
Tool-Box	0	+	+	-	+

Table 3.1: Comparison of NLP architectures: “+” fully supported, “0” partially supported, “-” not supported.

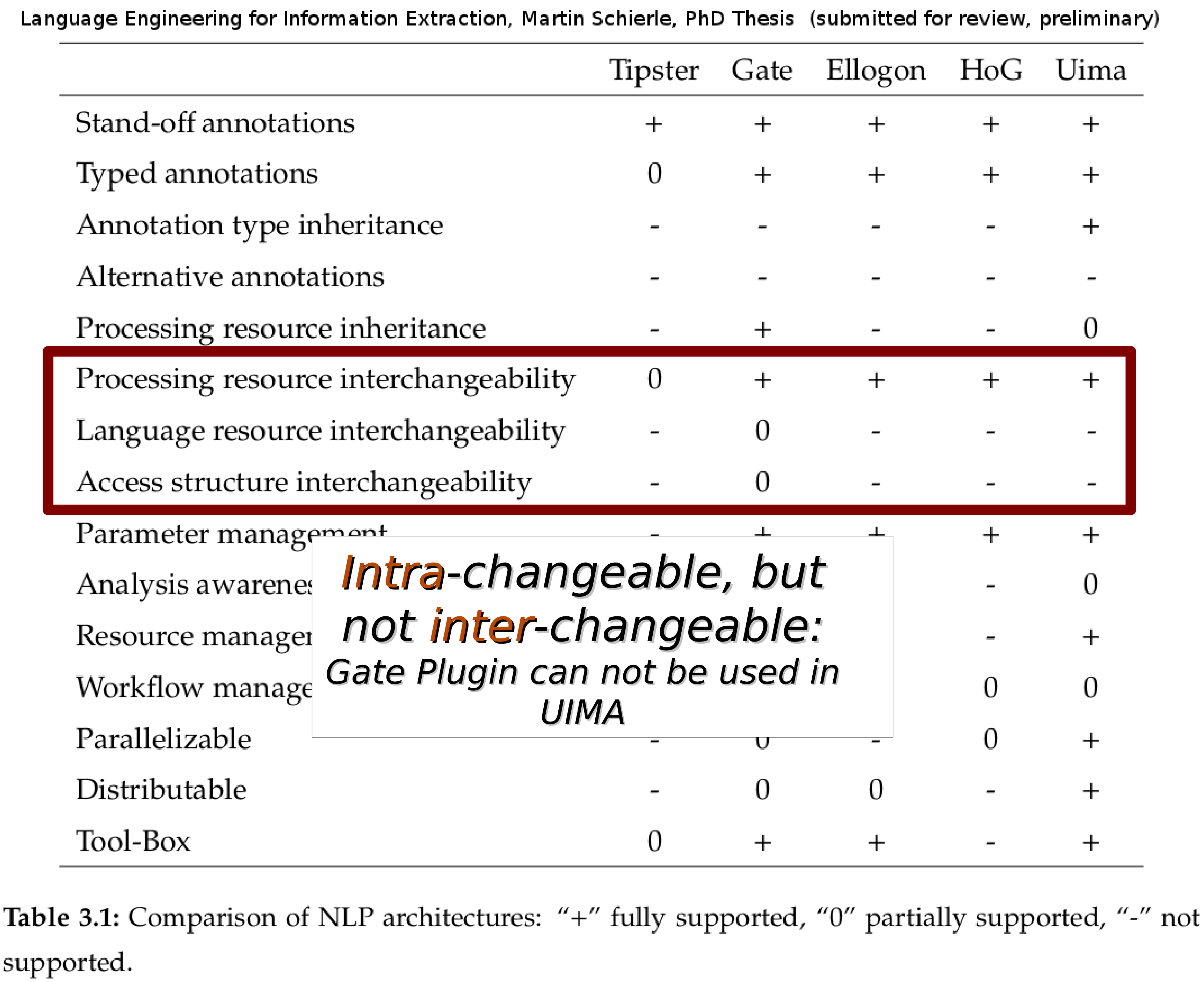
Language Engineering for Information Extraction, Martin Schierle, PhD Thesis (submitted for review, preliminary)



	Tipster	Gate	Ellogon	HoG	Uima
Stand-off annotations	+	+	+	+	+
Typed annotations	0	+	+	+	+
Annotation type inheritance	-	-	-	-	+
Alternative annotations	-	-	-	-	-
Processing resource inheritance	-	+	-	-	0
Processing resource					
Language resource i					
Access structure inte					
Parameter managem					
Analysis awareness	-	-	-	-	0
Resource management	-	-	-	-	+
Workflow management	-	0	0	0	0
Parallelizable	-	0	-	0	+
Distributable	-	0	0	-	+
Tool-Box	0	+	+	-	+

Table 3.1: Comparison of NLP architectures: "+" fully supported, "0" partially supported, "-" not supported.

Included in RDF/OWL as
- *rdf:type*
- *rdfs:subClassOf*
- *links and mappings*



Language Engineering for Information Extraction, Martin Schierle, PhD Thesis (submitted for review, preliminary)

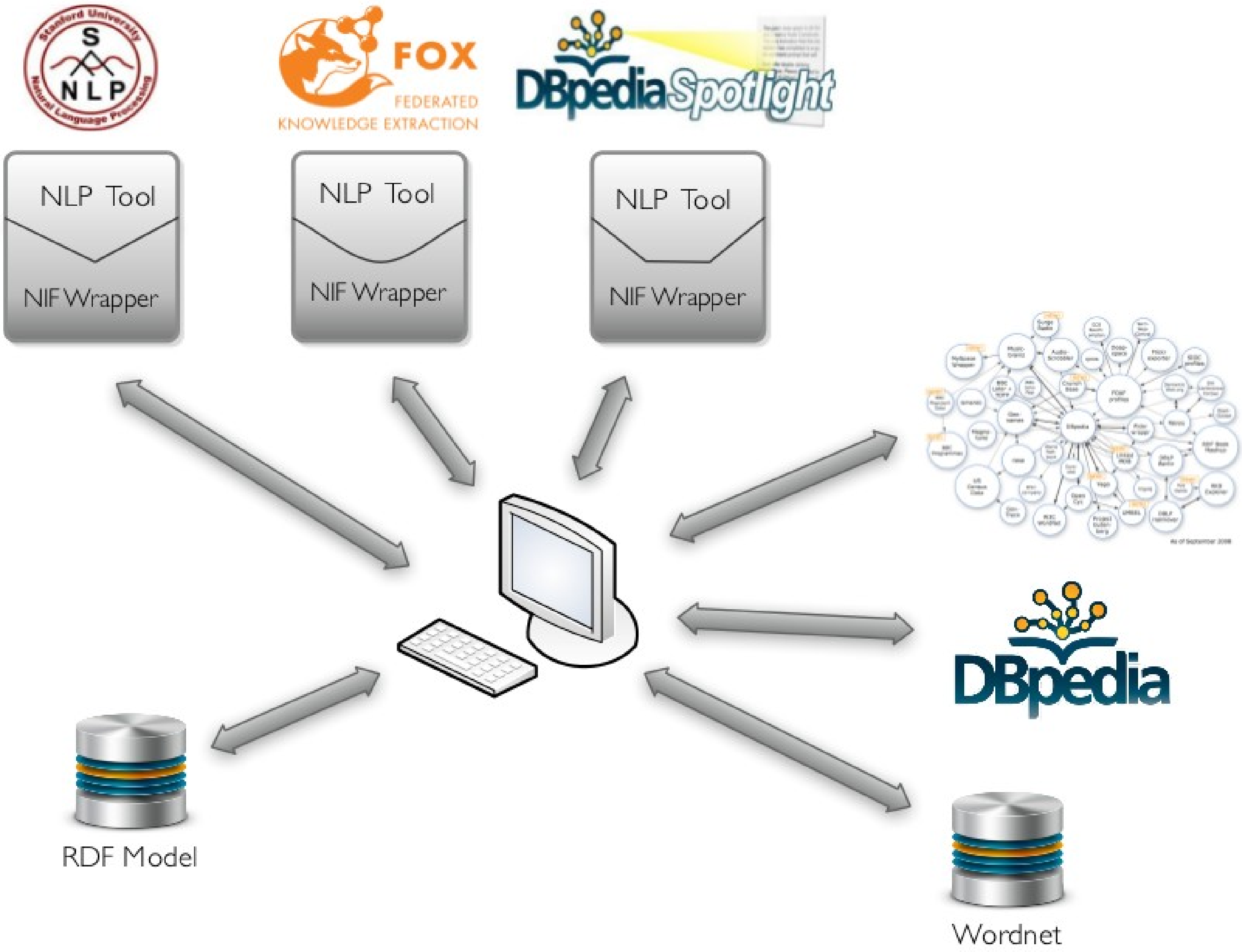
	Tipster	Gate	Ellogon	HoG	Uima
Stand-off annotations	+	+	+	+	+
Typed annotations	0	+	+	+	+
Annotation type inheritance	-	-	-	-	+
Alternative annotations	-	-	-	-	-
Processing resource inheritance	-	+	-	-	0
Processing resource interchangeability	0	+	+	+	+
Language resource interchangeability	-	0	-	-	-
Access structure interchangeability	-	0	-	-	-
Parameter management	-	+	+	+	+
Analysis awarenes	[illegible]	[illegible]	[illegible]	-	0
Resource manage	[illegible]	[illegible]	[illegible]	-	+
Workflow manage	[illegible]	[illegible]	[illegible]	0	0
Parallelizable	-	0	-	0	+
Distributable	-	0	0	-	+
Tool-Box	0	+	+	-	+

Table 3.1: Comparison of NLP architectures: "+" fully supported, "0" partially supported, "-" not supported.

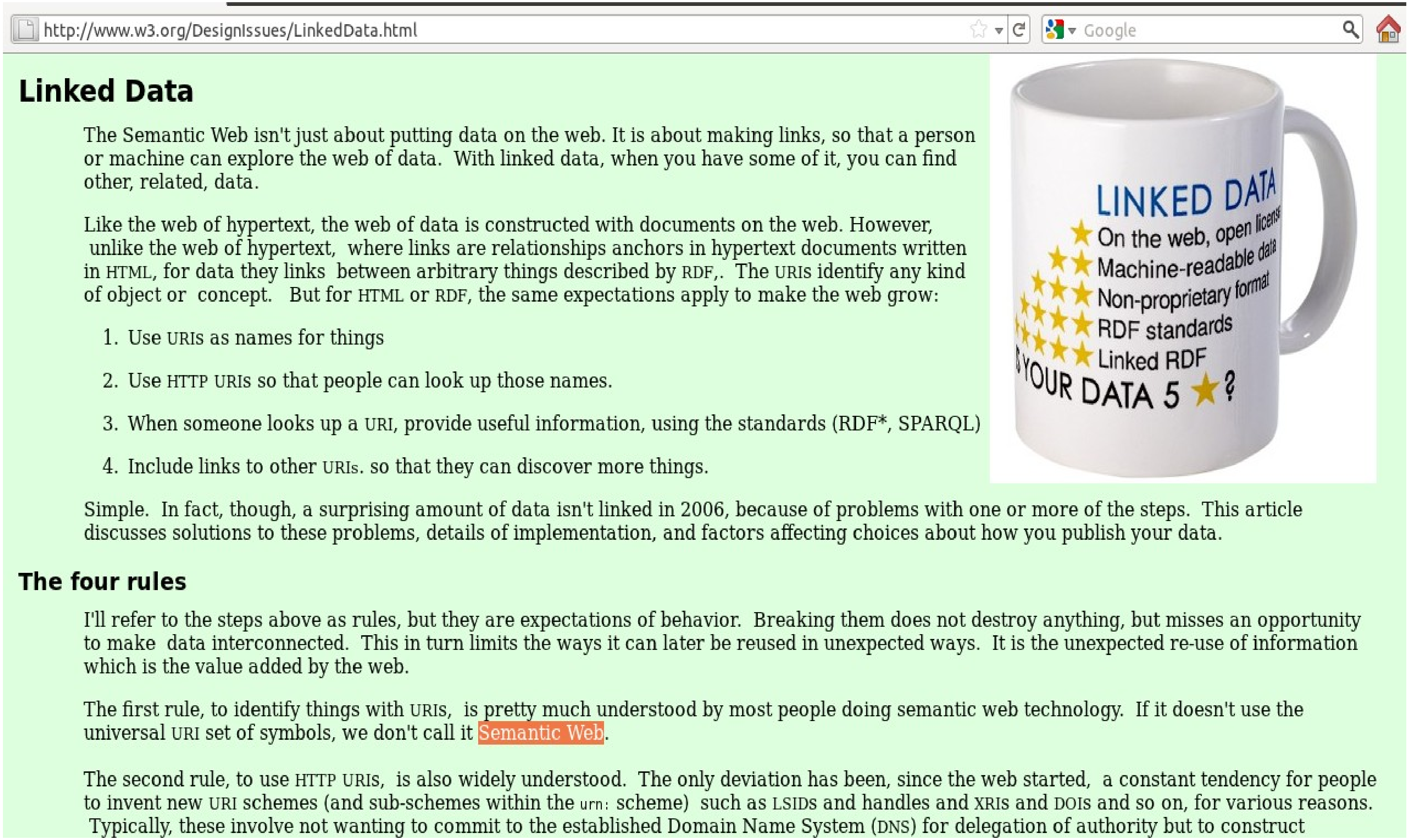
Intra-changeable, but not inter-changeable:

Gate Plugin can not be used in UIMA

NIF – Integration Architecture



NIF – How to address Strings with URIs?



http://www.w3.org/DesignIssues/LinkedData.html
Google
LINKED DATA
★ On the web, open license
★★ Machine-readable data
★★★ Non-proprietary format
★★★★ RDF standards
★★★★★ Linked RDF
IS YOUR DATA 5 ★ ?

Linked Data

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.

Like the web of hypertext, the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links between arbitrary things described by RDF,. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the web grow:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things.

Simple. In fact, though, a surprising amount of data isn't linked in 2006, because of problems with one or more of the steps. This article discusses solutions to these problems, details of implementation, and factors affecting choices about how you publish your data.

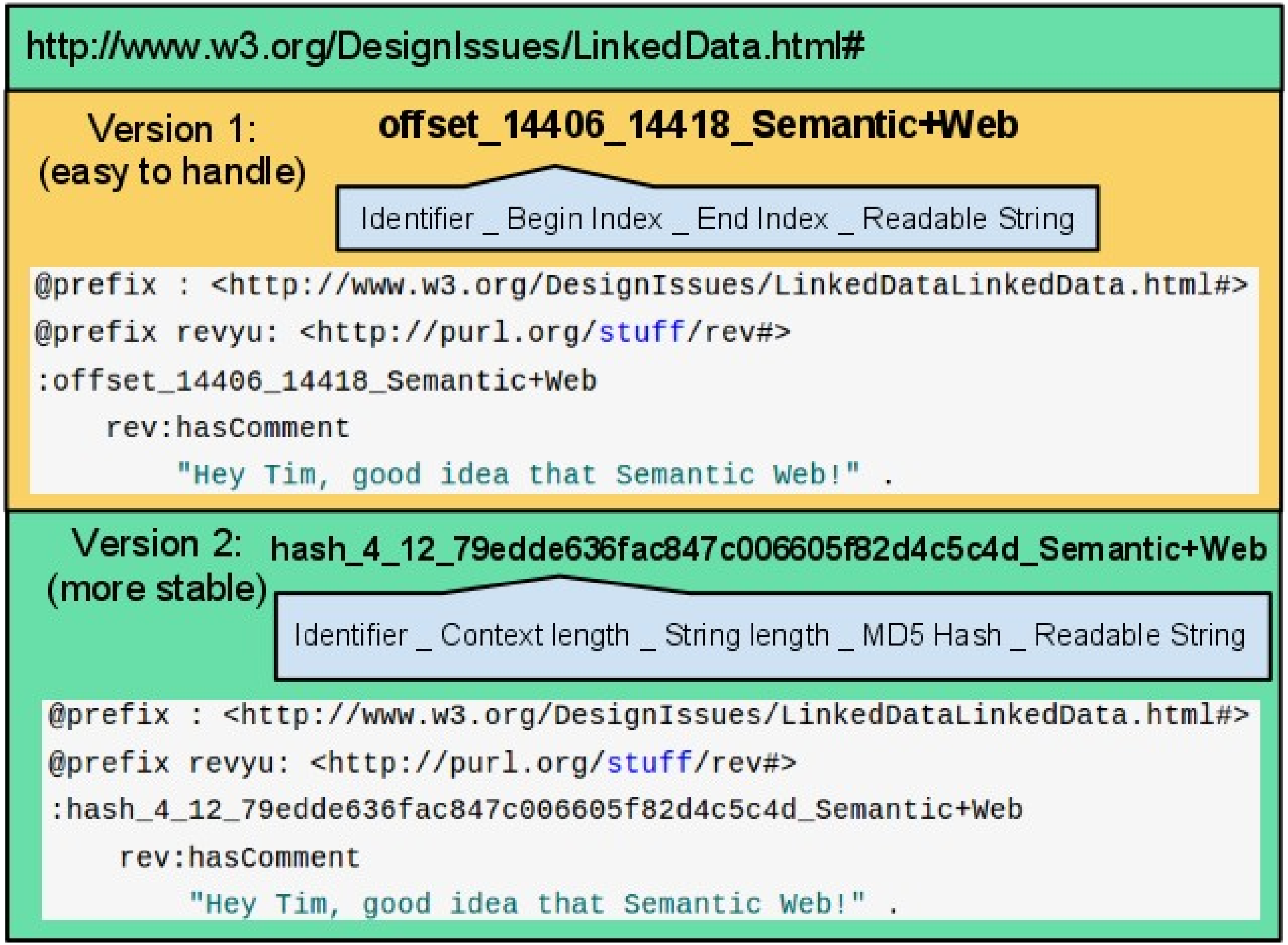
The four rules

I'll refer to the steps above as rules, but they are expectations of behavior. Breaking them does not destroy anything, but misses an opportunity to make data interconnected. This in turn limits the ways it can later be reused in unexpected ways. It is the unexpected re-use of information which is the value added by the web.

The first rule, to identify things with URIs, is pretty much understood by most people doing semantic web technology. If it doesn't use the universal URI set of symbols, we don't call it Semantic Web.

The second rule, to use HTTP URIs, is also widely understood. The only deviation has been, since the web started, a constant tendency for people to invent new URI schemes (and sub-schemes within the urn: scheme) such as LSIDs and handles and XRIs and DOIs and so on, for various reasons. Typically, these involve not wanting to commit to the established Domain Name System (DNS) for delegation of authority but to construct

NIF – How to address Strings with URIs?



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

Version 1:
(easy to handle)

offset_14406_14418_Semantic+Web

Identifier _ Begin Index _ End Index _ Readable String

```
@prefix : <http://www.w3.org/DesignIssues/LinkedDataLinkedData.html#>
@prefix revyu: <http://purl.org/stuff/rev#>
:offset_14406_14418_Semantic+Web
    rev:hasComment
        "Hey Tim, good idea that Semantic Web!" .
```

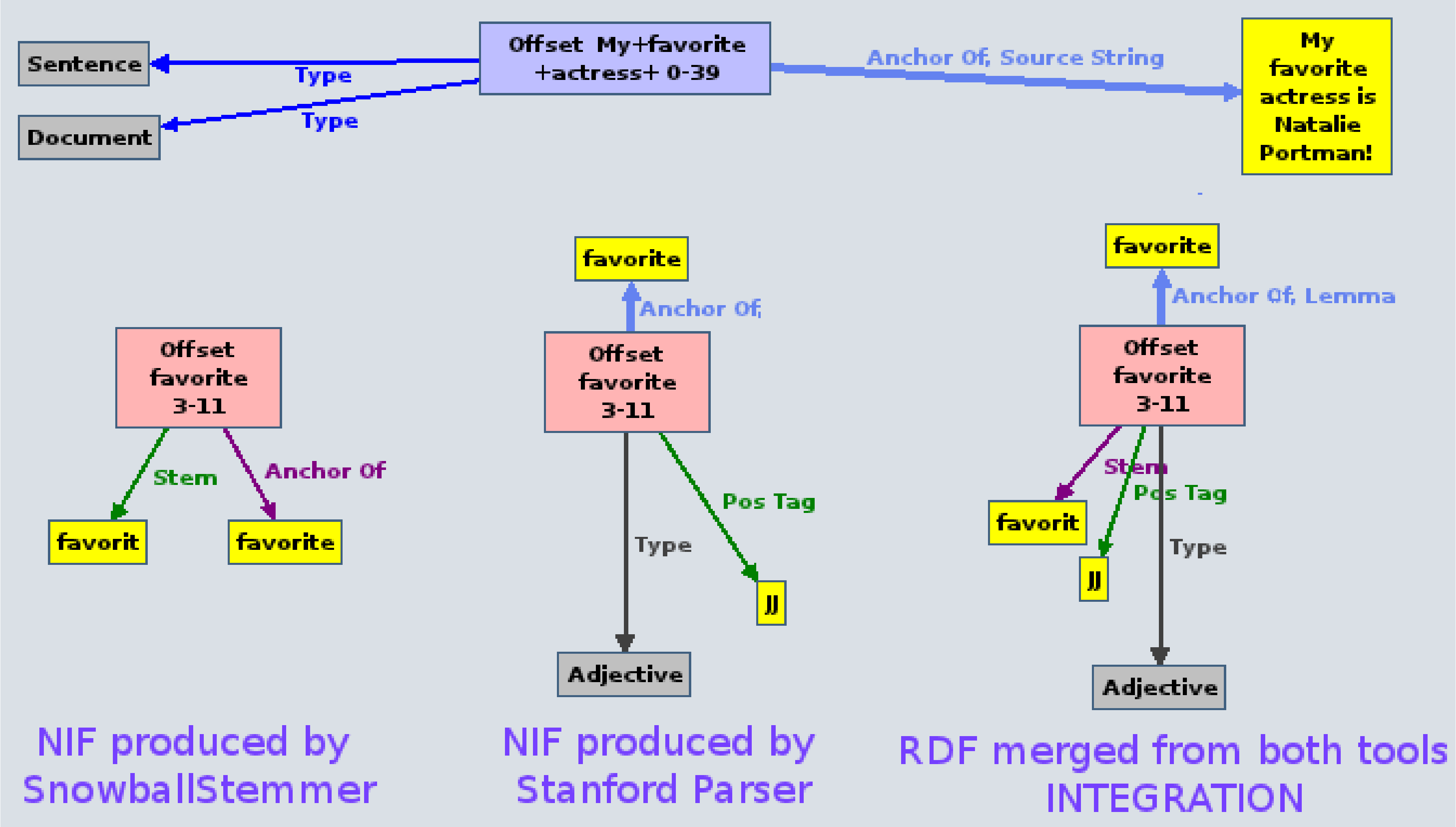
Version 2:
(more stable)

hash_4_12_79edde636fac847c006605f82d4c5c4d_Semantic+Web

Identifier _ Context length _ String length _ MD5 Hash _ Readable String

```
@prefix : <http://www.w3.org/DesignIssues/LinkedDataLinkedData.html#>
@prefix revyu: <http://purl.org/stuff/rev#>
:hash_4_12_79edde636fac847c006605f82d4c5c4d_Semantic+Web
    rev:hasComment
        "Hey Tim, good idea that Semantic Web!" .
```

NIF – Combined RDF

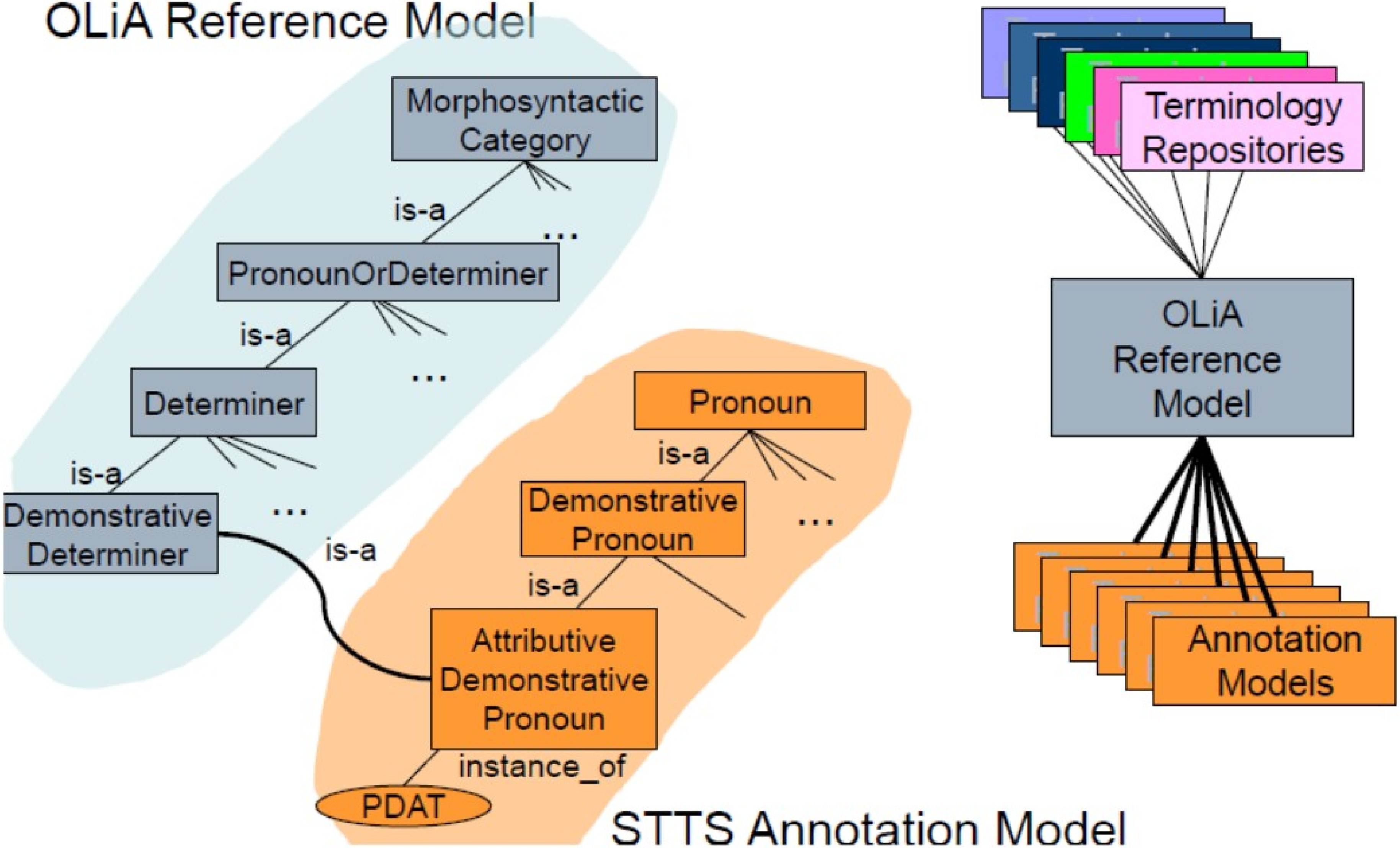


NLP2RDF – NIF – 1.0

- NIF-1.0 provides
  - URI recipes to anchor annotation in documents
  - Ontologies to describe the relations between these URIs:
    - e.g. subString, String, Word, Sentence, Document
    - http://nlp2rdf.lod2.eu/schema/string/
    - http://nlp2rdf.lod2.eu/schema/sso/
  - Vocabularies for certain NLP tasks and domains
    - e.g. OLiA [Chiarcos 2008, 2010] http://nachhalt.sfb632.uni-potsdam.de/owl/

OLiA

Ontologies of Linguistic Annotation



OLiA Reference Model

Morphosyntactic Category

is-a

...

PronounOrDeterminer

is-a

...

Determiner

is-a

...

Demonstrative Determiner

is-a

Pronoun

is-a

...

Demonstrative Pronoun

is-a

Attributive Demonstrative Pronoun

instance_of

PDAT

STTS Annotation Model

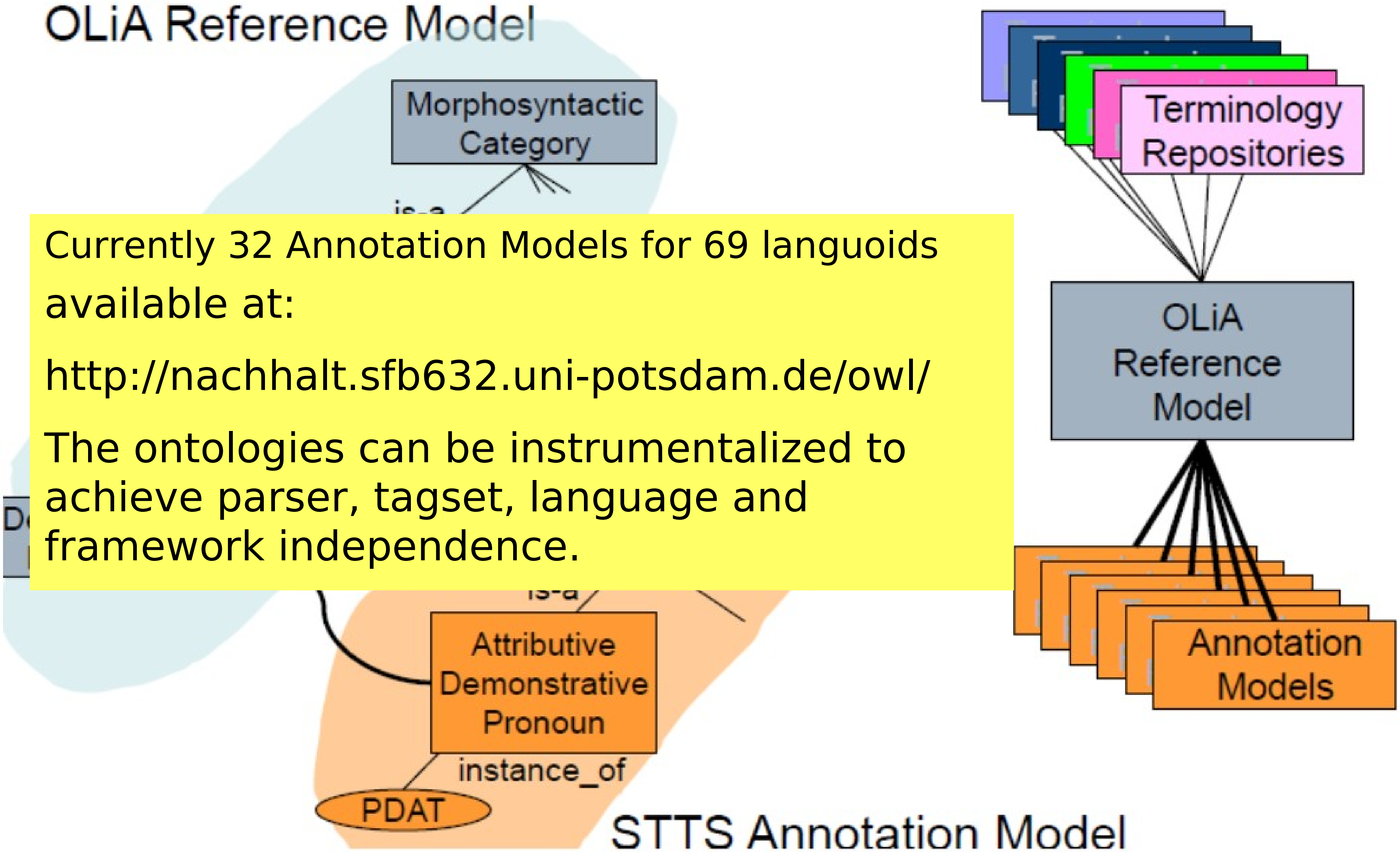
Terminology Repositories

OLiA Reference Model

Annotation Models

OLiA

Ontologies of Linguistic Annotation



Currently 32 Annotation Models for 69 languoids

available at:

http://nachhalt.sfb632.uni-potsdam.de/owl/

The ontologies can be instrumentalized to achieve parser, tagset, language and framework independence.

NIF RoadMap

- RoadMap:
  - NIF 1.0 is published and implementation has started
  - http://nlp2rdf.org allows to browse the implementations
  - Benchmarking of String URI properties (stability)
  - Interactive *Tutorial challenges* online
  - NIF 2.0-draft will be refined based on the experience gained during the implementation of NIF 1.0
  - Several organisations already use NIF (especially LOD2)

Contact

Address

University of Leipzig
Faculty of Mathematics and Computer Science
Institute of Computer Science
Department of Business Information Systems

Postfach 100920
04009 Leipzig
Germany

Project: http://lod2.eu
Organisation: http://uni-leipzig.de, http://aksw.org
Presenter: http://bis.informatik.uni-leipzig.de/SebastianHellmann
NLP2RDF page: http://nlp2rdf.org

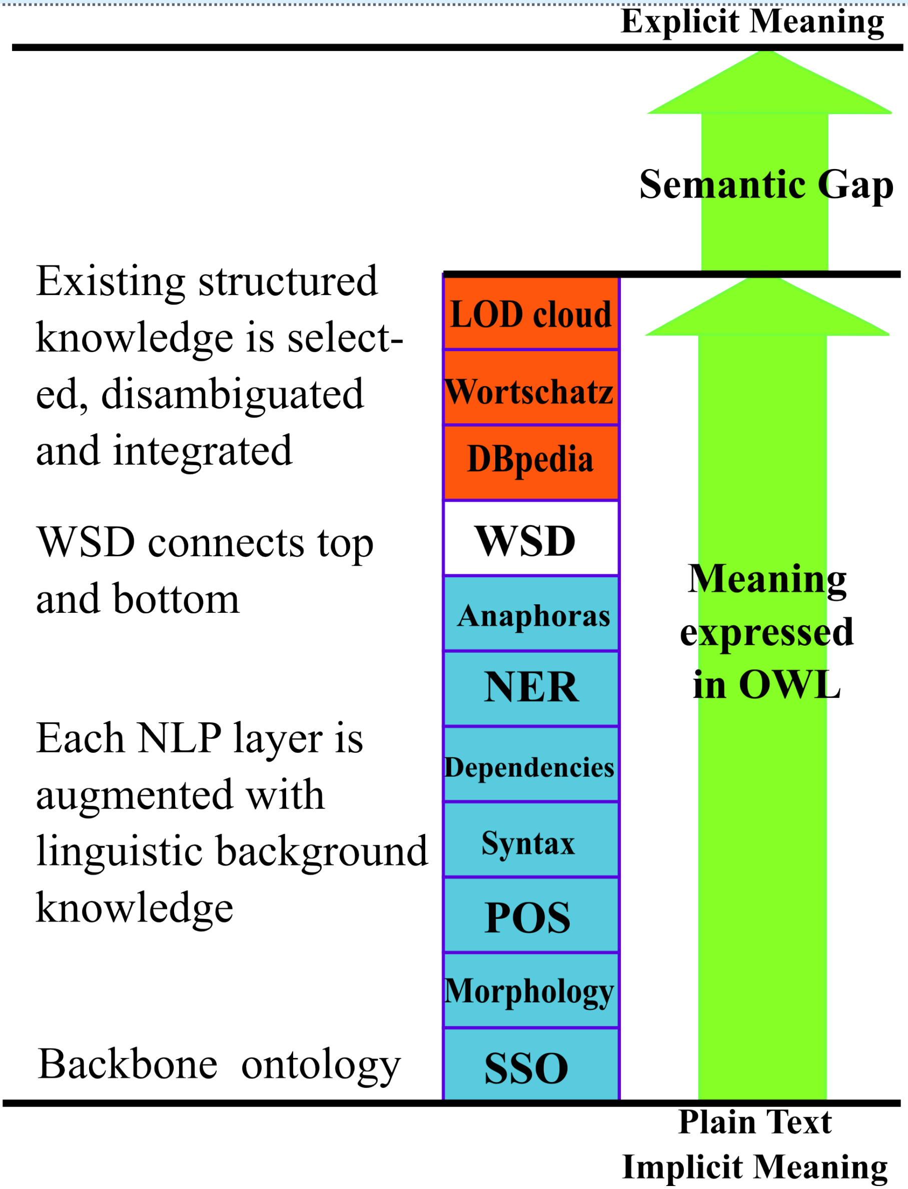
Thanks for your attention!

Meaning Representation Language

Advantages of RDF/OWL

- RDF makes data integration easy: URIref, LinkedData
- OWL is based on Description Logics (Guarded Fragment)
- Availability of open data sets (access and licence)
- Reusability of Vocabularies and Ontologies
- Diverse serializations for annotations: XML, Turtle, RDFa+XHTML
- Scalable tool support (Databases, Reasoning)
- Data is flexible and can produce indexes

Meaning Representation Language



Knowledge Extraction with SPARQL

Classical approach:

- POS tag / Dependency parser (e.g. Stanford)
- create a rule/pattern language to extract knowledge

Lot's of home-made solutions and problems!

Knowledge Extraction with SPARQL

Johanna Völker – Learning Expressive Ontologies (LExO)

```
# Example:

# A fish is any aquatic vertebrate animal that is covered with scales, and equipped with
two sets of paired fins and several unpaired fins.

# [fish] subClassOf [any aquatic vertebrate animal that is covered ...]


Construct {?sub rdfs:subClassOf ?super} {

 ?is a penn:BePresentTense .

 ?is nlp:superToken ?is_any_aquatic_.

 ?is_any_aquatic_ a olia:VerbPhrase .

 ?is_any_aquatic_ nlp:syntacticSubToken [ nlp:normUri ?super] .

 ?animal nlp:cop ?is .

 ?animal nlp:nsubj ?fish .?fish nlp:superToken [ nlp:normUri ?sub] .

}
```