apache stanbol™

http://stanbol.apache.org

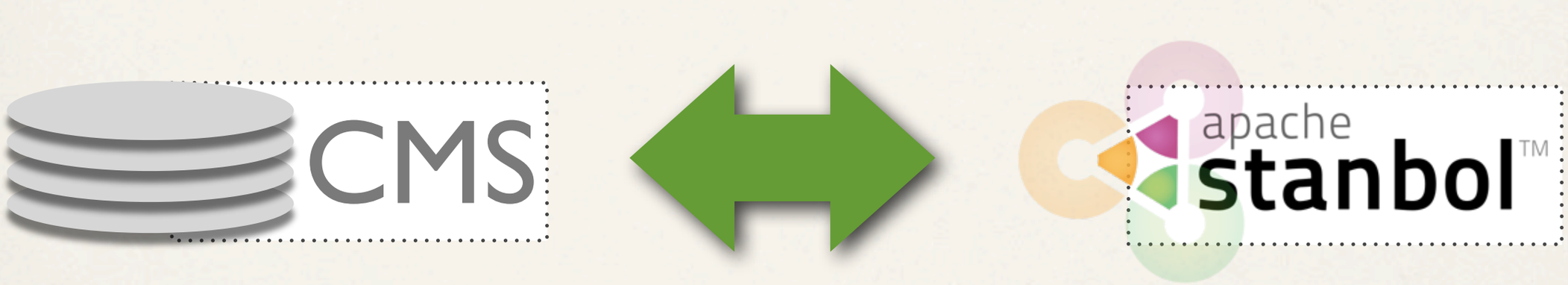
a modular software stack and reusable set of components for semantic content management

19. April, 2012

Semantic Content Management with Apache Stanbol

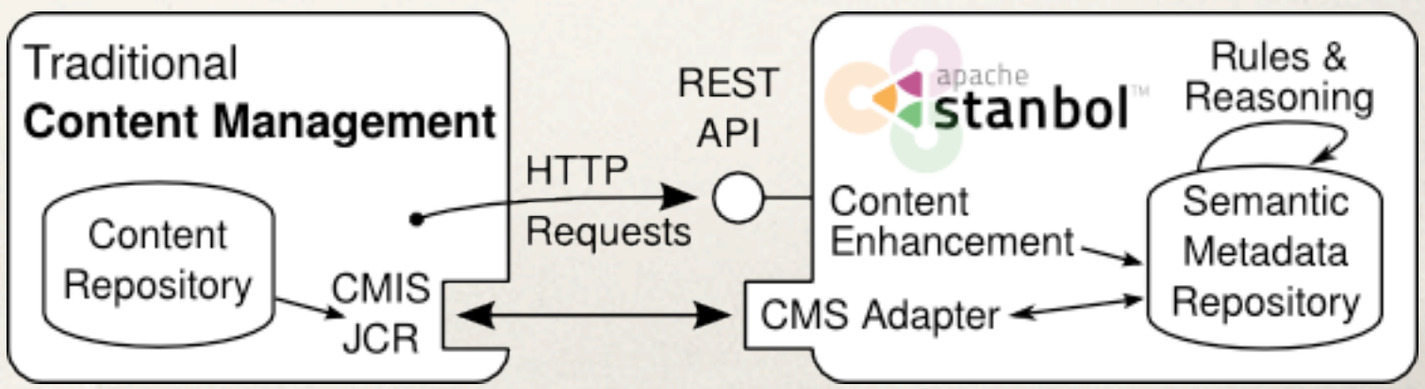
Traditional

Semantic Engine



Semantic Content Management with Apache Stanbol

- **Enhancer:** Extracts Knowledge from parsed Content
- **Entityhub:** Manage Entities and Topics of Interest to your Domain
- **Contenthub:** Semantic Indexing / Search over your - semantic enhanced - Content
- **CMS Adapter:** Sync. your CMS with Apache Stanbol (JCR/CMIS)
- **Ontology Manager:** Manage you formal Domain Knowledge
- **Reasoners & Rules:** Apply Domain Knowledge to improve / validate extracted Information. Refactor / refine knowledge to align it to public schemas such as schema.org

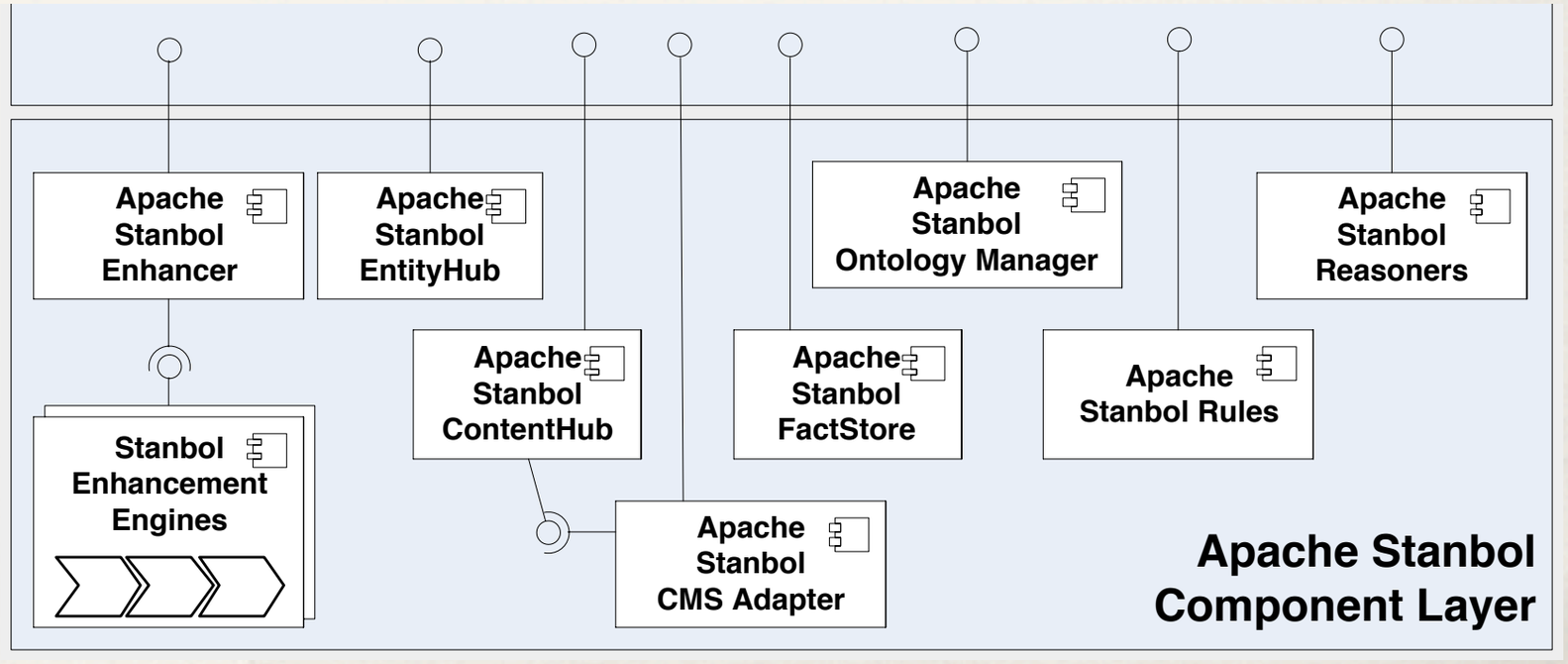


Stanbol Design and Integration Patterns

Don't buy everything.
Take the Components you Need!

- Stanbol Components provide
  - RESTful API
  - Java API and OSGI services

- Stanbol Components do NOT depend on each other
  - however they can be easily combined to



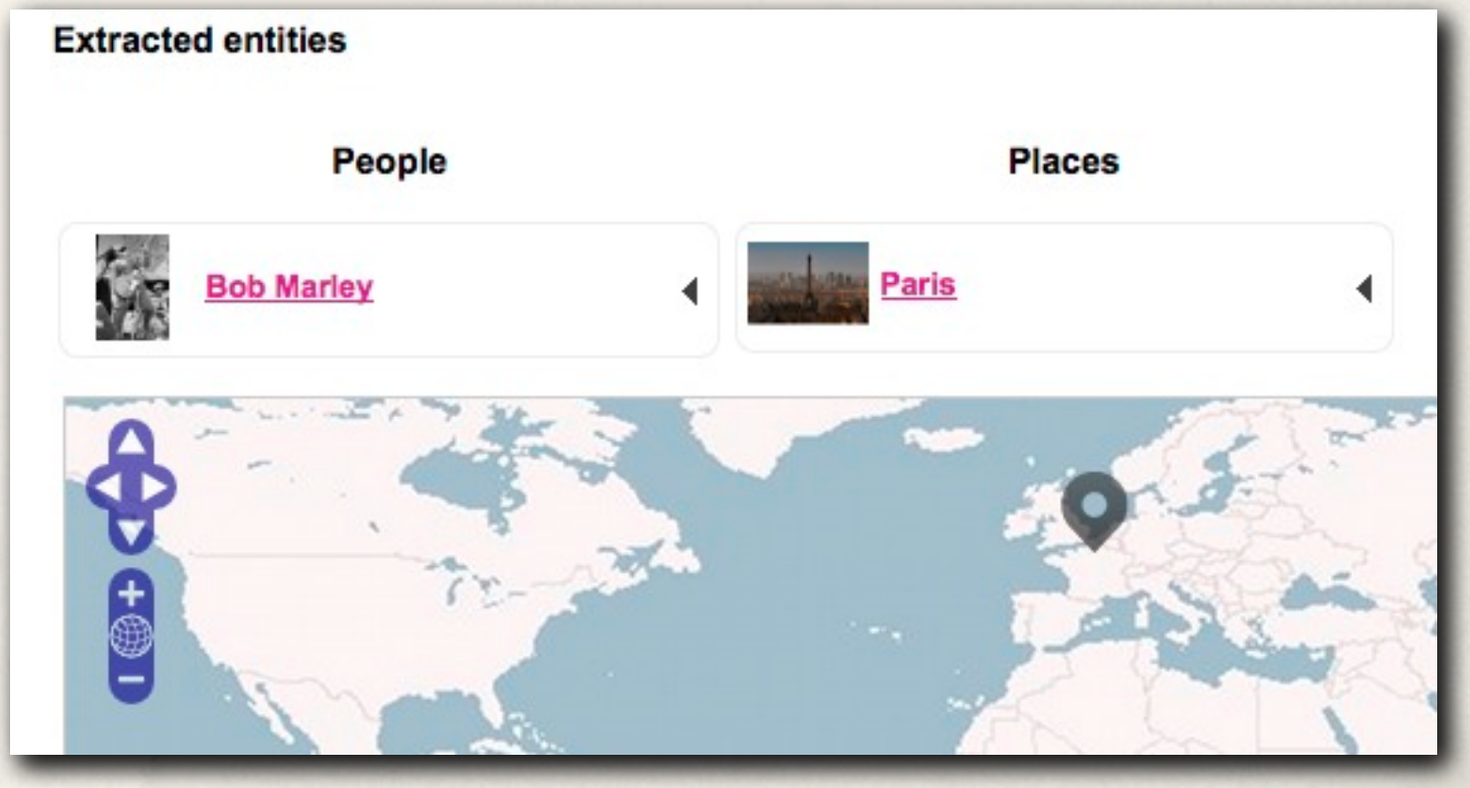
Stanbol Enhancer

Get to **know** your **Content**

```
curl -X POST -H "Accept: text/turtle" -H "Content-type: text/plain" \
    --data "The Stanbol enhancer can detect famous cities such as \
            Paris and people such as Bob Marley." \
    http://localhost:8080/enhancer
```

Enhancement Chain: **default** all **5** engines available

- **tika** (optional , TikaEngine)
- **langid** (required , LangIdEnhancementEngine)
- **ner** (required , NamedEntityExtractionEnhancementEngine)
- **dbpediaLinking** (required , NamedEntityTaggingEngine)



apache stanbol™

Stanbol Enhancer

Get to **know** your **Content**

```
curl -X POST -H "Accept: text/turtle" -H "Content-type: text/plain" \
    --data "The Stanbol enhancer can detect famous cities such as \
            Paris and people such as Bob Marley." \
    http://localhost:8080/enhancer
```

Enhancement Chain: **default** all **5** engines available

- **tika** (optional , TikaEngine)
- **langid** (required , LangIdEnhancementEngine)
- **ner** (required , NamedEntityExtractionEnhancementEngine)
- **dbpediaLinking** (required , NamedEntityTaggingEngine)

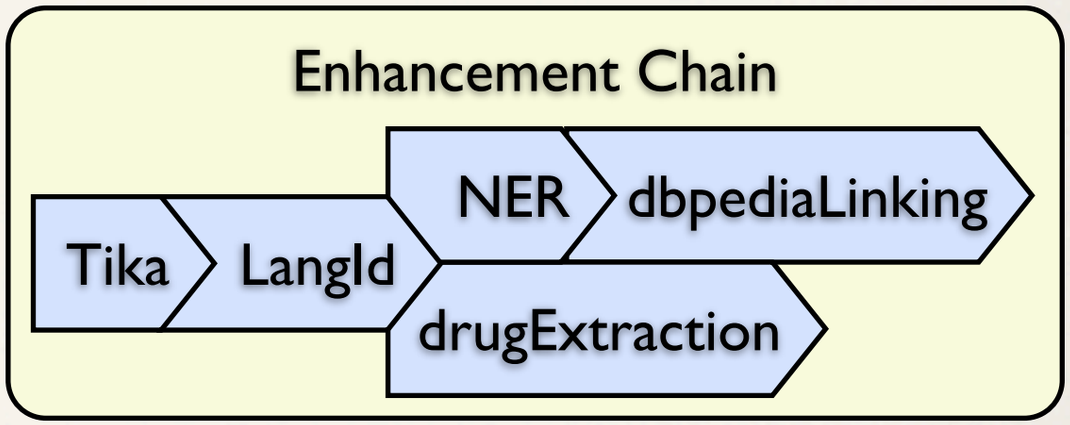
RDF


```
{
  "@subject": "urn:enhancement-784296de-6aee-95a8-8f84-839a1e24d1b9",
  "@type": [
    "enhancer:Enhancement",
    "enhancer:EntityAnnotation"
  ],
  "dc:created": "2012-04-13T13:43:56.016Z",
  "dc:creator": "org.apache.stanbol.enhancer.engines.entitytagging.impl.NamedEntityTaggingEngine",
  "dc:relation": "urn:enhancement-929e0dc8-6c5e-e44c-4c1d-c669f96d00d7",
  "enhancer:confidence": 17396.67,
  "enhancer:entity-label": {
    "@literal": "Bob Marley",
    "@language": "en"
  },
  "enhancer:entity-reference": "http://dbpedia.org/resource/Bob_Marley",
  "enhancer:entity-type": [
    "dbp-ont:MusicalArtist",
    "foaf:Person",
    "dbp-ont:Artist",
    "dbp-ont:Person",
    "owl:Thing"
  ],
  "enhancer:extracted-from": "urn:content-item-sha1-4186ce0dd89b27663a8ea60fc7acebceefa20174"
},
```

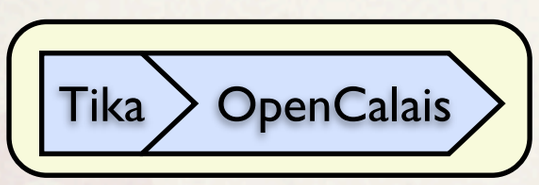

apache stanbol™

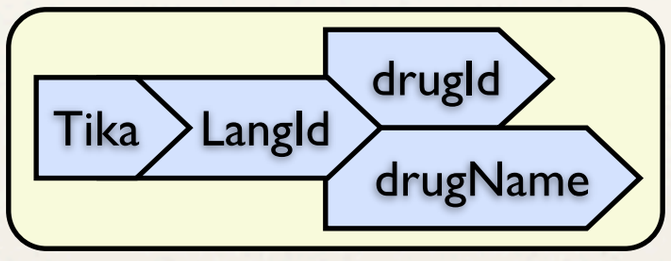
Enhancement Chains

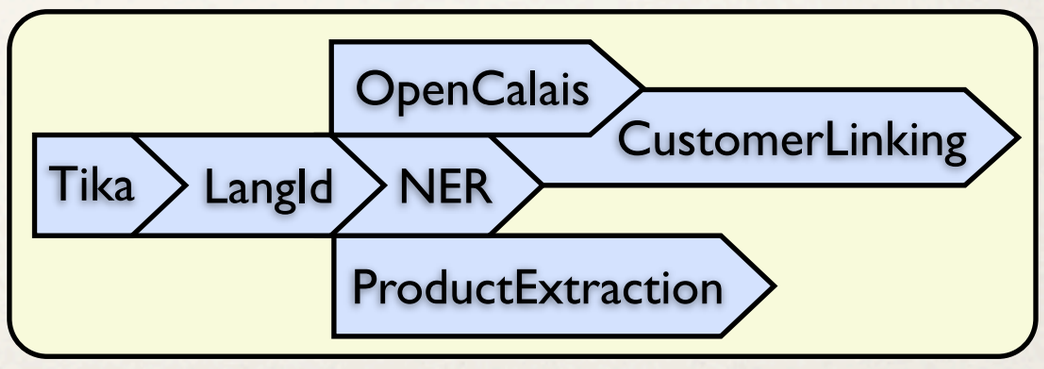
- Define how Content is processed by the Enhancer
  - `/enhancer` calls the default Chain
  - use multiple Chains `/enhancer/chain/{name}`
  - call single EnhancementEngines `/enhancer/engine/{name}`



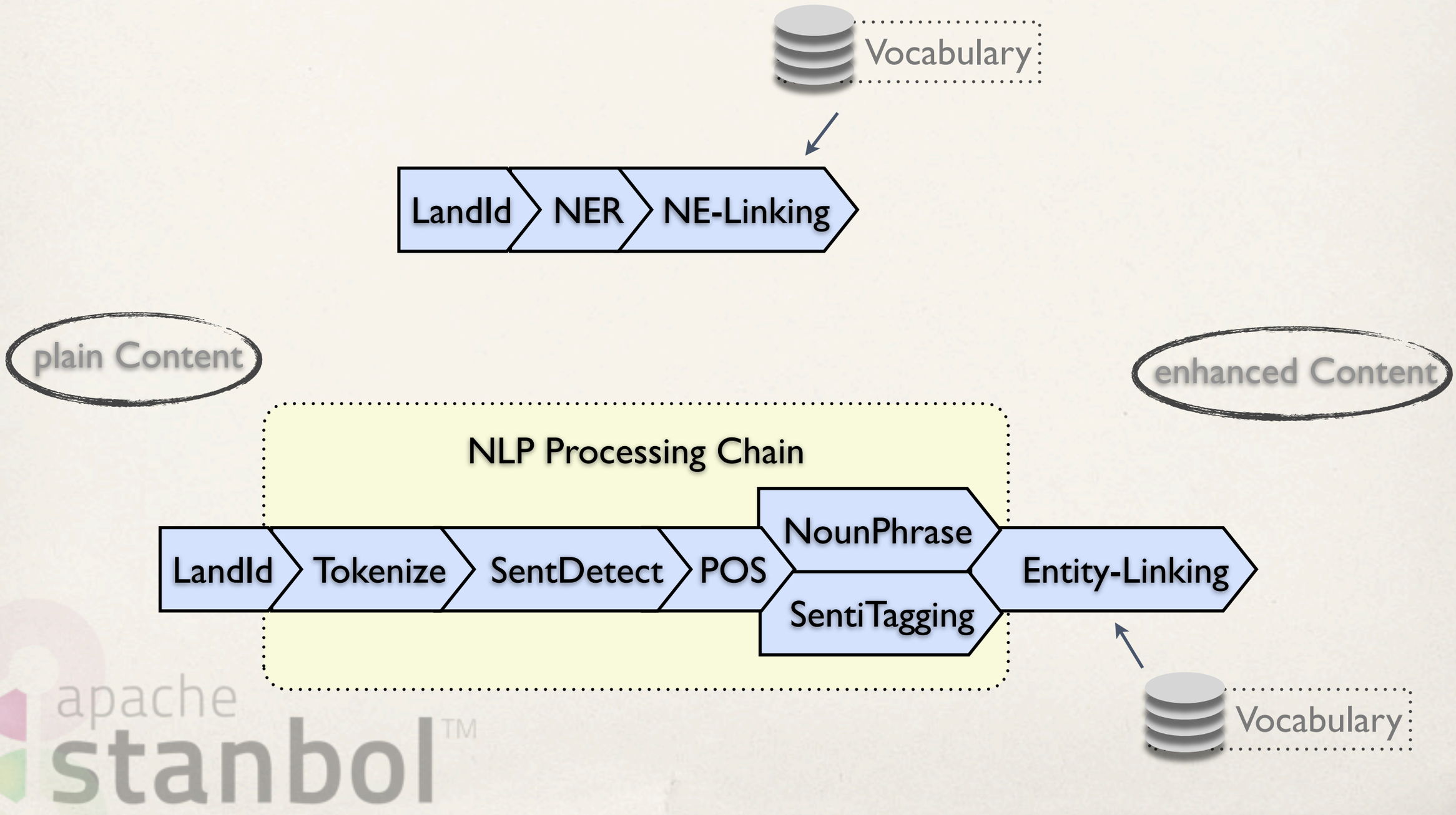
- Some Examples:



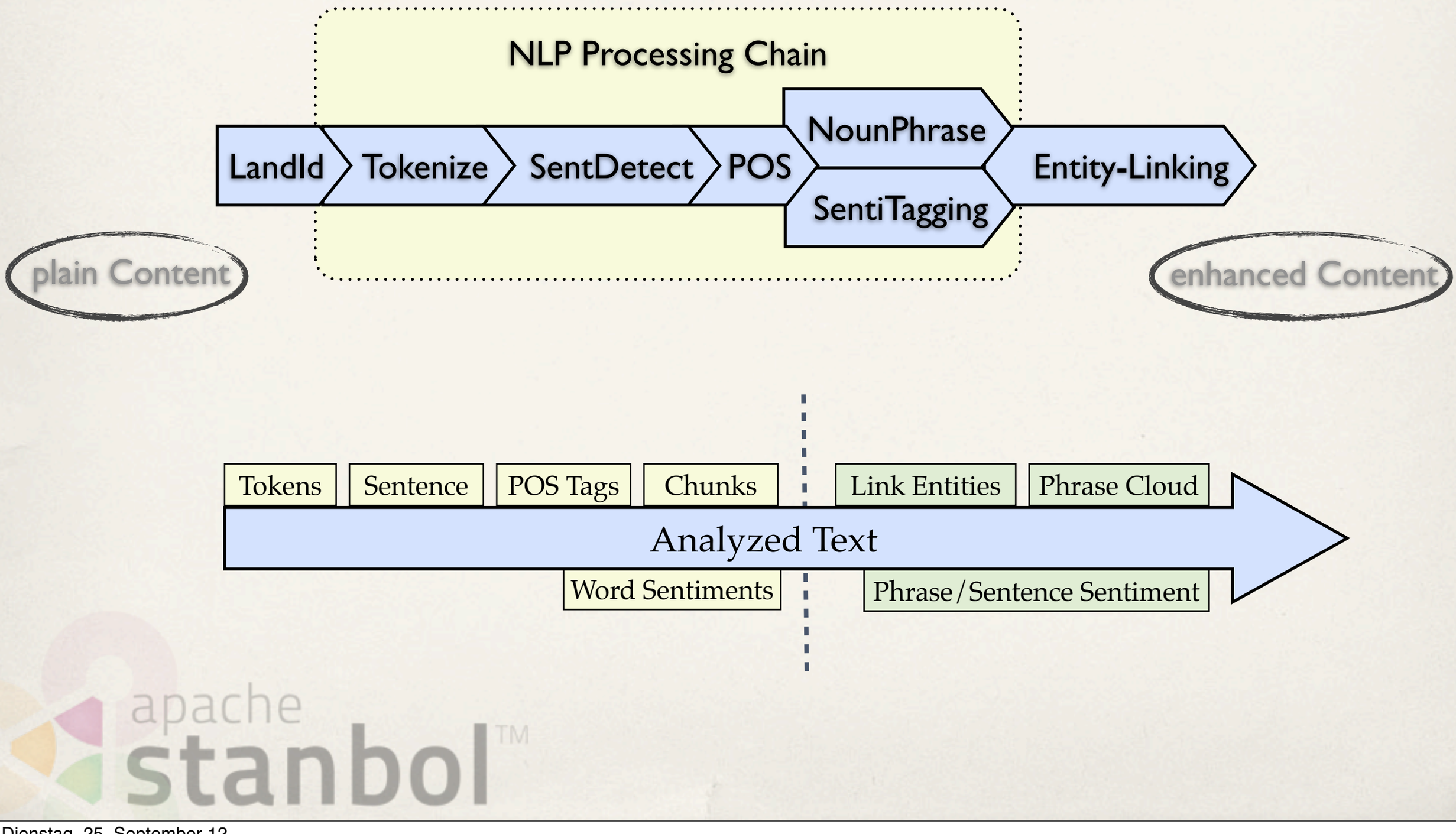




Stanbol NLP Processing (STANBOL-733)

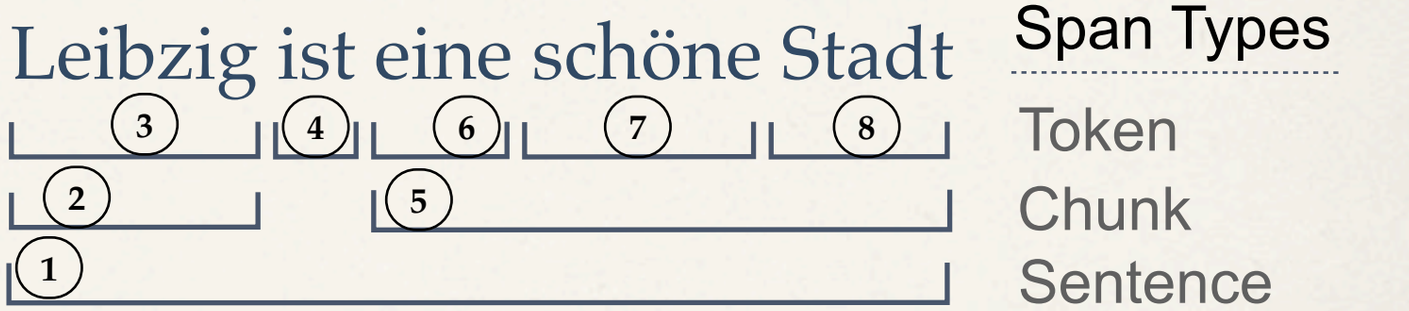


Stanbol NLP Processing (STANBOL-733)



Analyzed Text (1/2)

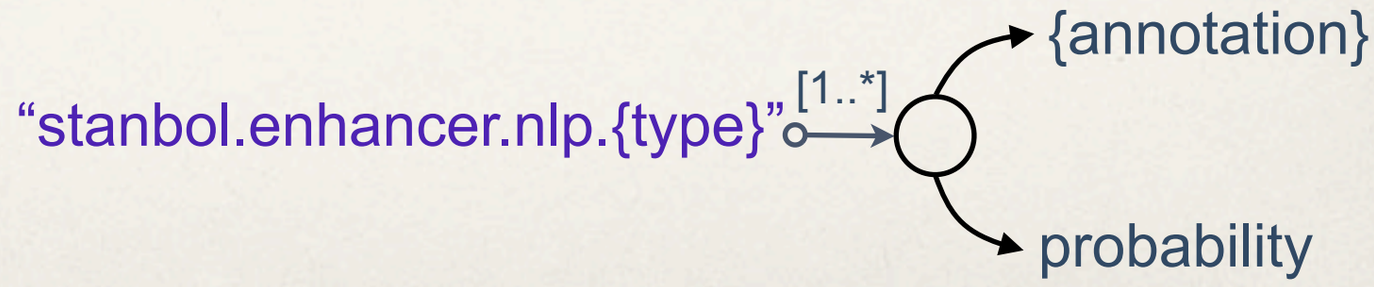
- Navigateable Map with Spans
  - Span sorted by Natural Order
  - Iterator based API that allows concurrent Modifications



Span Types

- Token
- Chunk
- Sentence
- Text Section
- Analyzed Text

- Spans support Annotations



Analyzed Text (2/2)

- Pos Annotation
- Phrase Annotation
- Sentiment Annotation
- TODO:
  - Lemma, Tense, Number (plural/singular), ...

PosTag

tag (e.g. "NE")
lexical-category
(e.g. olia:Noun)

PhraseTag

tag (e.g. "NP")
lexical-category
(e.g. olia:NounPhrase)

SentimentTag :: Double

We are looking for

Work with the Stanbol Community

- Connect NLP Frameworks
  - POS tagging, Chunking, Relation detection …
  - in {your} language
  - including commercial frameworks
- Make Datasets available
  - Dictionary based Sentiment Tagging, Lemmatizer
-

openNLP™

Language Grid

Lucene

Wiktionary
['wɪkʃənrɪ] *n.*,

WORTSCHATZ
UNIVERSITÄT LEIPZIG

WordNet
A lexical database for English

apache stanbol™

Stanbol Facts

- Web: http://stanbol.apache.org/
- Mailing List: dev@stanbol.apache.org
- Releases:
  - 0.9.0-incubation
  - Entityhub: 0.10.0-incubation
- Graduated to full Apache 19.August 2012
  - based on code developed by the IKS project [1]

apache stanbol™

[1] http://www.iks-project.eu