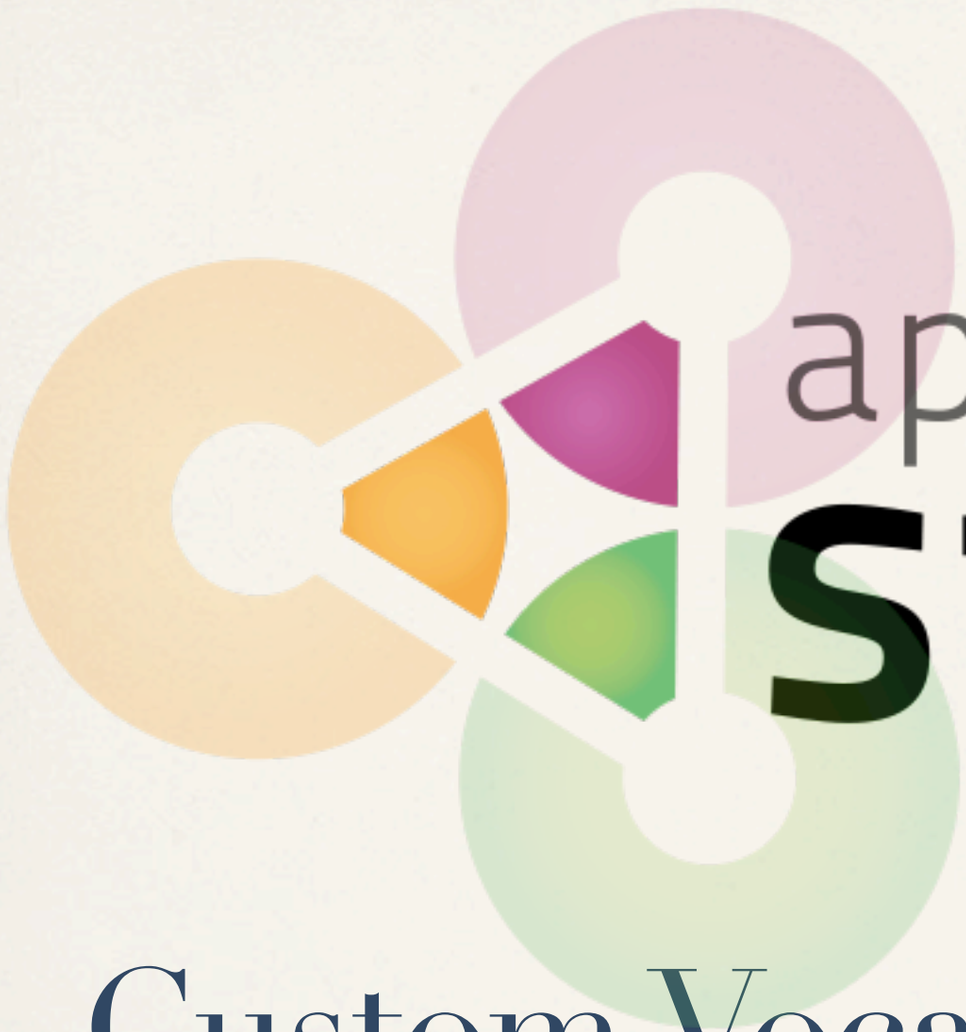


apache

stanbol™

http://stanbol.apache.org

Use Custom Vocabularies with the Stanbol Enhancer

Rupert Westenthaler, Salzburg Research, Austria

07. November, 2012

About Me

- Rupert Westenthaler
  - Apache Stanbol and Clerezza Committer
  - rwesten
  - Affiliation: Salzburg Research, Austria
- Contact Details
  - rupert.westenthaler@gmail.com, rwesten@apache.org
  - twitter: @westei, #stanbol
  - IRC: #stanbol
  - Skype: westei

apache stanbol™

salzburgresearch

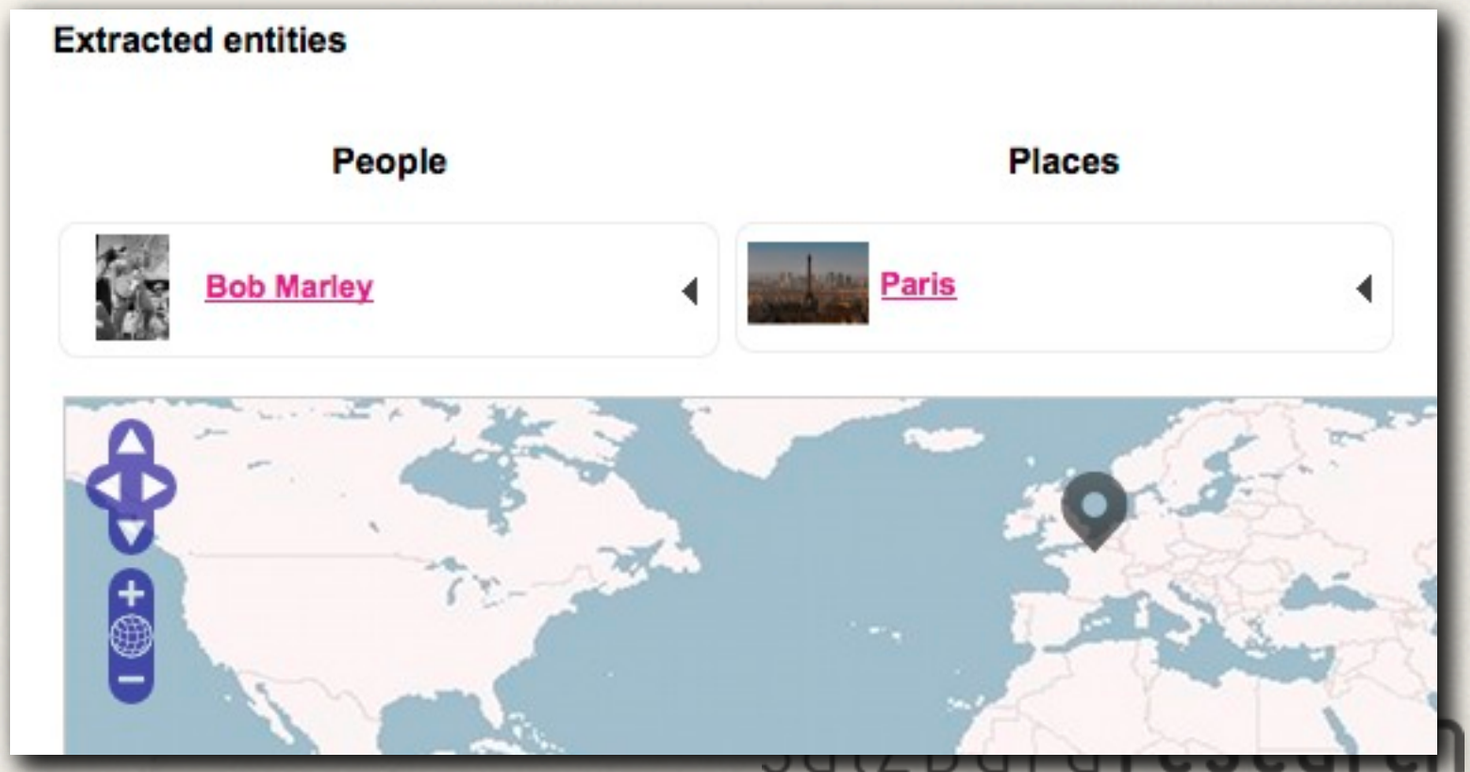
Stanbol Enhancer

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™

Domain Specific Enhancement

Bring your own **Entities**

```
If you have any of these other conditions, you may need a dose
adjustment or special tests to safely take aspirin:

  * asthma or seasonal allergies;
  * stomach ulcers;
  * liver disease;
  * kidney disease;
```

Enhancement Chain: **ehealth** all **4** engines available

- **tika** (optional , TikaEngine)
- **langid** (required , LangIdEnhancementEngine)
- **ehealthExtraction** (required , KeywordLinkingEngine)
- **drugIdExtraction** (required , KeywordLinkingEngine)

Life Sciences

SIDER 2 Side Effect Resource

DrugBank Open Data Drug & Drug Target Database

Diseasome

Extracted entities

Diseases	Drugs
? Asthma	? Aspirin
? Polycystic kidney disease	
? Polycystic liver disease	

See Demo:
{stanbol-trunk}/demo/ehealth

apache stanbol™

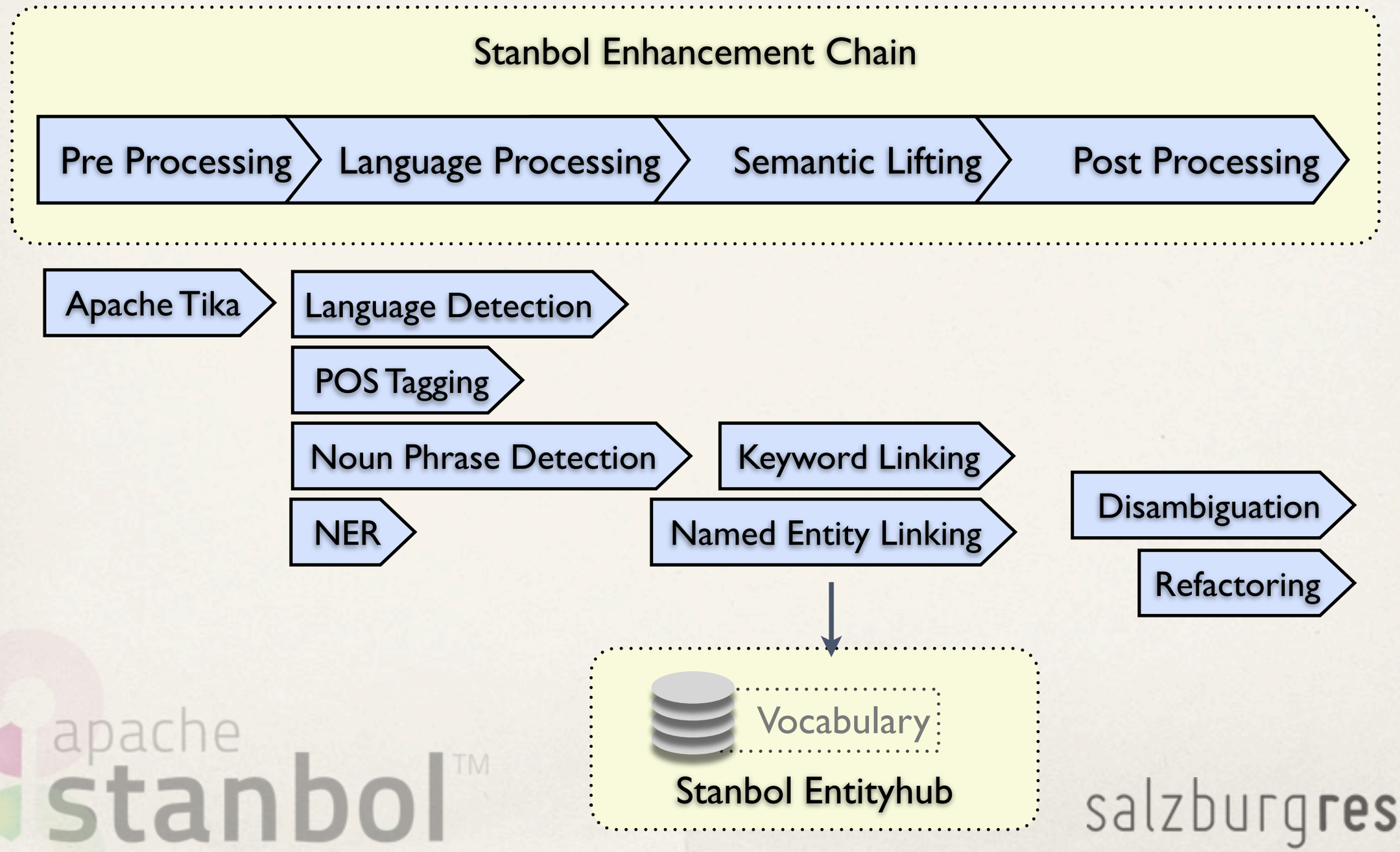
Agenda

- Stanbol Enhancer
  - General Enhancement Workflow
  - Available Enhancement Engines
- Manage Domain Vocabularies with the Stanbol Entityhub
- How to Extract your Entities
  - Named Entity based Linking
  - Word (Phrase) based Linking
- Configure the Stanbol Enhancer for Your Domain Vocabulary

apache stanbol™

salzburgresearch

Enhancement Workflow



apache stanbol™

salzburgresearch

Enhancement Engines 1/3

- Apache Tika Engine / Metaxa Engine
  - Plain Text extraction; Metadata Extraction; Content Type detection
- Language Detection
  - Tika LangId; Language-Detection Engine
- Topic Classification
  - Trainingset / Classifier for your Topics
  - supports hierarchical Classification Schemes
- Apache OpenNLP NER
  - extracts Persons / Organizations / Places
  - custom NER models trained for custom Entity Types

Enhancement Engines 2/3

- Named Entity Linking
  - Links recognized Entities with Controlled Vocabularies
- Keyword Extraction
  - Label based extraction of Entities
- Refactor Engine
  - Rule based post-processing of Enhancements results
- Integrated “external” Services:

GeoNames

Unstructured Information Management Architecture
An Apache Project.

Zemanta

DBpedia Spotlight

Language Grid

apache stanbol™

CALAIS
Powered by Thomson Reuters

salzburgresearch

Enhancement Engines 3/3
STANBOL-733: "Stanbol NLP Processing Module"

- OpenNLP based
  - Tokenizing, Sentence Detection, POS tagging, Chunking
- Sentiment
  - Dictionary / POS based Word Sentiment tagging
  - Sentiment Summarization (Phrase, Sentence, Document)
- Lemmatization
  - CELI / linguagrid.org Lemmatize Engine
- Bring your own NLP framework

salzburgresearch

Stanbol Entityhub

manage the **Entities** of your **Domain**

- Manage multiple Entity Source - "Sites"
  - Entities are stored using Apache Solr or apache clerezza
- Entity Retrieval
- Query for Entities
- LDpath [1] support for:
  - graph path retrieval, schema translation and simple reasoning

jena

openRDF.org Sesame

```
curl http://localhost:8080/entityhub/site/dbpedia/entity\
    id={entity-id}
```

```
curl -X POST -d "name=lyon&limit=10" \
    http://localhost:8080/entityhub/site/dbpedia/find
```

```
friend-names = foaf:knows/foaf:name
```

```
schema:name = rdfs:label[@en];
schema:description = rdfs:comment[@en];
schema:image = foaf:depiction;
schema:url = foaf:homepage;
```

```
skos:broaderTransitive = (skos:broader)+;
skos:related = (skos:related | ^skos:related);
```

[1] http://code.google.com/p/ldpath/

apache stanbol™

salzburgresearch

Vocabulary Management
-- 2 possibilities --

- Entityhub Managed Site
  - Use RESTful API to manage the Vocabulary

- Entityhub Referenced Site
  - using a full local index
  - created by using the Entityhub Indexing Tool
  - installable to Apache Stanbol

salzburgresearch

Managed Site

- Configure a ManagedSite
  - Solr Yard (stores the Entity Data)
  - Managed Site (provides Access to the Entities)
- RESTful API
  - curl -i -X POST -H "Content-Type: application/rdf+xml" \
    -T {file.rdf} "http://localhost:8080/entityhub/site/{name}/entity"
  - curl -i -X DELETE "http://localhost:8080/entityhub/site/{name}/entity?id={entity-uri}"

apache stanbol™

salzburgresearch

Entityhub Indexing Tool 1/2

- Standalone Application to Index Vocabularies
  - Expert Level Tool!
  - Good default Configuration

```
java -jar org.apache.stanbol.entityhub.indexing.genericrdf-*.jar init
```

- Configuration: “indexing/config/indexing.properties”
  - Indexed Fields: “indexing/config/mappings.txt”
- RDF data: “indexing/resources/rdfdata”

Entityhub Indexing Tool 2/2

- Indexing:

  java -Xmx1024m -jar org.apache.stanbol.entityhub.indexing.genericrdf-*.jar index

- Results: “indexing/dist”
  - “{name}.solrindex.zip”
    - copy to the Stanbol “datafile” directory
  - “org.apache.stanbol.data.site.{name}-1.0.0.jar”
    - install to the OSGI environment

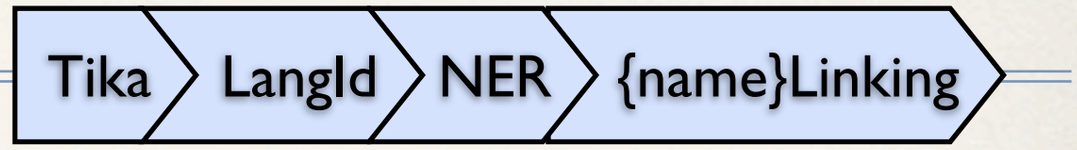
Enhancement Workflows
-- 2 possibilities --

- Named Entity based Linking
  - extract Named Entities
  - lookup Named Entities in your Vocabulary
- Keyword (-phrase) based Lookup
  - Detect “interesting” Words -> Keywords
    - Tokenize, POS tagging, Noun Phrase Detection
  - lookup Keywords in your Vocabulary

apache stanbol™

salzburgresearch

Named Entity Linking 1/2

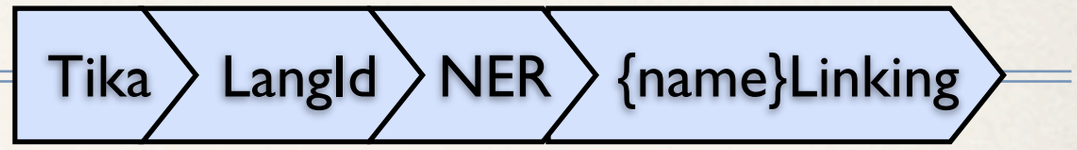


- Named Entity Extraction
  - Entity Label
  - Entity Type (Person, Organization, Place, {other-trained-types}
- Named Entity Linking
  - Filter based on Type
  - Query for Label

apache stanbol™

salzburgresearch

Named Entity Linking 2/2

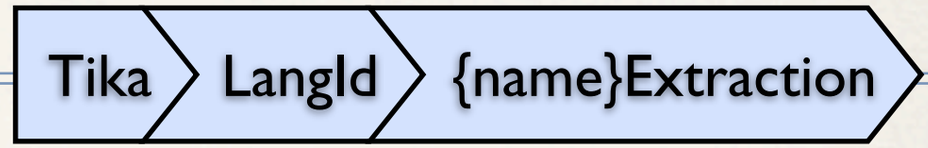


- Requires
  - Language of the Text
  - NER model for the Language AND the Type of Named Entities
- NER support for Persons, Organizations, Places
  - English, Spanish, Dutch (OpenNLP)
  - English, French, Spanish (OpenCalais)
  - French, Italian (via CELI NER engine)
  - Your can use your own OpenNLP NER Model
- Configure a NamedEntityLinking Engine for Your Vocabulary

apache stanbol™

salzburgresearch

Keyword Linking 1/2



- NLP processing
  - Tokenizing
  - POS Tagging
  - Noun Phrase Detection

Noun → Common Noun
Noun → Proper Noun

STANBOL-733: NLP Processing

- Keyword Linking
  - Query for Entities that match selected Tokens in the Text

apache stanbol™

salzburgresearch

Keyword Linking 2/2

- Requires
  - Language of the Text
  - Tokenized Text
- Language Support
  - POS: Danish, Dutch, English, German, Portuguese, Spanish, Swedish
  - Noun Phrase Detection: English, German
  - STANBOL-733: Bring your own NLP models / framework!
- Configure a KeywordLinkingEngine for Your Vocabulary

apache stanbol™

salzburgresearch

Enhancement Chains

- Define how Content is processed by the Enhancer
  - `/enhancer` calls the default Chain
  - call chains by name `/enhancer/chain/{name}`
  - call single EnhancementEngines `/enhancer/engine/{name}`
- Chain Implementations:
  - Weighted Chain - Engines are sorted by their “ORDERING”
  - List Chain - Engines are executed in the configured order
  - Graph Chain - Configure dependencies between Engines

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 Apache TLP on 19.August 2012
  - incubated based on code developed by the IKS project [1]

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

apache stanbol™

salzburgresearch