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

19. April, 2012
Semantic Content Management with Apache Stanbol
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 your 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

- 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

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

www.iks-project.eu
Stanbol Enhancer

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)

Extracted entities

People

- Bob Marley

Places

- Paris
Stanbol Enhancer

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

```json
{
   "@subject": "urn:enhancement-784296de-60ee-95c8-f984-839a1e24d1b9",
   "@type": [
      "enhancer:Enhancement",
      "enhancer:EntityAnnotation"
   ],
   "dc:relation": "urn:enhancement-9298dc8-6e5e-e44c-4c1d-c669f96d80d7",
   "enhancer:confidence": 17396.67,
   "enhancer:entity-label": {
      "@literal": "Bob Marley",
      "@language": "en"
   },
   "enhancer:entity-type": [
      "dbp:ont:MusicalArtist",
      "foaf:Person",
      "dbp:ont:Artist",
      "dbp:ont:Person",
      "owl:Thing"
   ],
   "enhancer:extracted-from": "urn:content-item-sha1-4186ce8dd9b27663a8ea60fca7acebceef820174"
}
```
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:

  ![Diagram](attachment:image.png)
Stanbol NLP Processing (STANBOL-733)

NLP Processing Chain

- LandId
- Tokenize
- SentDetect
- POS
- NounPhrase
- SentiTagging
- Entity-Linking

Vocabulary

plain Content

enhanced Content

Dienstag, 25. September 12
Stanbol NLP Processing (STANBOL-733)

NLP Processing Chain

Plain Content

Enhanced Content

Tokens  Sentence  POS Tags  Chunks  Link Entities  Phrase Cloud

Word Sentiments  Phrase/Sentence Sentiment

Dienstag, 25. September 12
Analyzed Text (1/2)

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

- **Spans support Annotations**

```
Leibzig ist eine schöne Stadt
```

Span Types
- Token
- Chunk
- Sentence
- Text Section
- Analyzed Text

```
"stanbol.enhancer.nlp.{type}"[1..*]
```

{annotation}

probability
Analyzed Text (2/2)

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

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

- Sentiment Annotation
  - SentimentTag :: Double

  - Lemma, Tense, Number (plural/singular), ...

Dienstag, 25. September 12
We are looking for

- Connect NLP Frameworks
  - POS tagging, Chunking, Relation detection …
  - in {your} language
  - including commercial frameworks

- Make Datasets available
  - Dictionary based Sentiment Tagging, Lemmatizer
Stanbol Facts

- **Web:**  [http://stanbol.apache.org/](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]

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