Automated Linking Data with

Olivier Grisel
http://www.nuxeo.com

Rupert Westenthaler
http://www.salzburgresearch.at

19. April, 2012

http://www.iks-project.eu
Semantic Content Management with Apache Stanbol

Traditional CMS ↔ Semantic Engine

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 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
Enhancement Engines 1/2

- Apache Tika Engine / Metaxa Engine
  - Plain Text extraction; Metadata Extraction; Content Type detection
- Language Detection
- Topic Classification
  - Trainingset / Classifier for your Topics
  - supports hierarchical Classification Schemes
- Named Entity Recognition
  - extracts Persons / Organizations / Places

soon:
Enhancement Engines 2/2

- 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:
  - Zemanta®
  - GeoNames
  - apacastanbol™
  - CALAIS
Domain Specific Enhancement

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)

Extracted entities

<table>
<thead>
<tr>
<th>Diseases</th>
<th>Drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>? Asthma</td>
<td>? Aspirin</td>
</tr>
<tr>
<td>? Polycystic kidney disease</td>
<td></td>
</tr>
<tr>
<td>? Polycystic liver disease</td>
<td></td>
</tr>
</tbody>
</table>
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:
Using Stanbol in Web Applications

HALLO — ANNOTATING CONTENT WITH LINKED DATA

Before taking Lotrel

You should not use Lotrel if:

- you are allergic to amlodipine (Norvasc) or benazepril (Lotensin);
- you have ever had angioedema (hives or severe swelling of deep skin tissues sometimes caused by allergies);
- you are allergic to any other ACE inhibitor, such as captopril (Capoten), fosinopril (Monopril), enalapril (Vasotec Zestril), moexipril (Univasc), perindopril (Aceon), quinapril (Accupril), ramipril (Altace), or trandolapril (Mavik).

To make sure you can safely take Lotrel, tell your doctor if you have any of these other conditions:

- kidney disease (or if you are on dialysis);
- liver disease;
- heart disease or congestive heart failure;
- diabetes; or
- if you are on a low sodium diet.

FDA pregnancy category C. Lotrel can cause harm to an unborn baby. Women who are or might become pregnant should not use Lotrel unless your doctor tells you that the benefits of Lotrel for you are greater than the possible harm to your unborn baby. If you become pregnant while using Lotrel, tell your doctor right away. You should use a form of birth control while taking Lotrel. Do not breastfeed a nursing baby.

Demo: http://bergie.github.com/hallo/annotate.html

http://viejs.org
We are looking for

- RDFa / Microdata support
  - Knowledge extraction while keeping positioning within the Content
- Entity Disambiguation
  - Entity-Linking + Disambiguation (e.g. by using Solr MLT)
  - Disambiguation of already linked Entities
- More Domain specific Customizations
  - Share as “/demo” with the Stanbol Community!
- <Your> Service as EnhancementEngine
Stanbol Entityhub

- Manage your Entities
  - e.g. Employees, Projects, User Tags, Company Thesaurus

- Support for Linked Data
  - Used for Entity-Linking by the Enhancer

- Specific Resources
  - Web Resources
    - linkedct.org
  - Clinical Trails

- Remote Sites
  - Reviews for Entities
    - revyu.com

- Local Cache
  - Support specific Sites via Extensions
  - Used to Retrieve additional Information for Entities
  - eHealth
  - dbpedia
  - revyu cache
  - linkedct cache
  - reuters cache

- Referenced Sites
  - apache
  - stanbol™
Stanbol Entityhub

- Manage multiple Entity Source - Referenced Sites
  - Supports fast local Caches using or
- Query for Entities
  - used by the Stanbol Enhancer
- LDpath [1] support for:
  - graph path retrieval
  - schema translation
  - simple reasoning

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

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

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

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

You can help by

- Integrate with Data Reconciliation Tools
  - Google Refine:
    - Silk: Entity Link discovery Framework
  - Support for <your> Dataset
    - direct access via EntityDereferencer implementation
    - provide as Entityhub ReferencedSite (or RDF dump)
Stanbol Contenthub

curl -i -X POST -H "Content-Type:text/plain" \ 
--data "Add your content here" \ 
http://localhost:8080/contenthub/contenthub/store

Enhancer

plain Content

data = "Add your content here"

Semantic Indexing

Semantic Search

Simple Faceted Search

Configure your Semantic Index Layout

RESTful API

Apache Solr

cms_adapter

CMS Adapter

enhanced Content

plain Content
Stanbol Contenthub

- Add Semantic Search to your CMS
  - RESTful Faceted Search Interface
  - Related Keyword Search using Entityhub, Ontonet or Wordnet

- Improve Search by Semantic Indexing
  - Keep using Solr as your Search Engine
  - Use the Stanbol Contenthub for semantic indexing
  - Configure Semantic Indexes by using LDpath
Customize Semantic Index
e.g. for the Life Science Domain

* Index Definition using LDpath [1]

```plaintext
@prefix dailymed: <http://www4.wiwiss.fu-berlin.de/dailymed/resource/dailymed/> ;
@prefix drugbank: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugbank/> ;
@prefix diseasome: <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseasome/> ;
@prefix sider: <http://www4.wiwiss.fu-berlin.de/sider/resource/sider/> ;

drug = .[rdf:type is dailymed:drugs | rdf:type is drugbank:drugs] :: xsd:anyURI;
drug_name = .[rdf:type is dailymed:drugs | rdf:type is drugbank:drugs]
  /skos:prefLabel :: xsd:string;

disease = .[rdf:type is diseasome:diseases] :: xsd:anyURI;
disease_name = .[rdf:type is diseasome:diseases]/skos:prefLabel :: xsd:string;

ingredient = .[rdf:type is dailymed:ingredients] :: xsd:anyURI;
ingredient_name = .[rdf:type is dailymed:ingredients]/rdfs:label :: xsd:string;

side_effect = .[rdf:type is sider:side_effects] :: xsd:anyURI;
side_effect_name = .[rdf:type is sider:side_effects]/rdfs:label :: xsd:string;
```

currently in Development

Enhancer

Provide annotated Content

plain Content

enhanced Content

Store

Stanbol will keep Indexes in Sync

Support for different Semantic Indexes

Semantic Indexer

CMS Adapter

Your CMS

File System

Apache Solr

apache clerezza™

Apache Jena

CouchDB

coming with Stanbol 0.1.0
follow STANBOL-471
Stanbol Ontology
Manager, Reasoning and Rules

- Manage your Ontologies
  - and use/combine them in Scopes

- Reasoning
  - on volatile Data loaded into a Sessions
  - consistency check / classification / enrichment
  - RDFS, OWL and OWL - 2

- Support for background Jobs
  - for long running reasoning tasks
Stanbol Ontology
Manager, Reasoning and Rules

Ontologies:
- SIOC
- FOAF
- STANBOL
- dbpedia
- SIDER
- DrugBank
- Dailymed
- Diseasome
- Life Science

Scopes:
- Creative Commons
- Music Ontology
- LinkedMovie DB schema
- schema.org
- Tag Ontology
- Social Network
- Media

Sessions:
- Reasoner
- Rules

Enhancements:
- Tim's Profile
- revyu.com

Enhancements:
- revyu.com

Ontologies:
- revyu.com schema

Tim's Profile:
- revyu.com
Welcome

- Share alignment rules across multiple domains
  - Especially with schema.org.
- Benchmarking:
  - how large are the scopes you are managing?
  - Sessions you use in your applications
- Wrap <your> Reasoner/Rule Engine as a Stanbol service
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
Apache Stanbol Facts

* Web:  http://incubator.apache.org/stanbol/

* Mailing List: stanbol-dev@incubator.apache.org

* Release:  in progress (currently: 0.9.0-incubation RC6)

  ![Apache Incubator](http://www.iks-project.eu)

* Incubation to Apache November 2010
  * based on code developed by the iKS project [1]

Rupert Westenthaler
rwesten@apache.org

http://www.salzburgresearch.at

Co-funded by the European Union

http://www.iks-project.eu