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 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
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
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;

**Domain Specific Enhancement**

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>

**Life Sciences**

- `SIDER 2` Side Effect Resource
- `DRUGBANK` Open Data Drug & Drug Target Database
- `Diseasome`
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:
  * Tika $\rightarrow$ OpenCalais
  * Tika $\rightarrow$ LangId $\rightarrow$ drugId $\rightarrow$ drugName
  * Tika $\rightarrow$ LangId $\rightarrow$ NER $\rightarrow$ CustomerLinking
  * Tika $\rightarrow$ LangId $\rightarrow$ NER $\rightarrow$ ProductExtraction
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 allergy);
- you are allergic to any other ACE inhibitor, such as captopril (Capoten), fosinopril (Monopril), enalapril (Vasotec), 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 D: Lotrel can harm a nursing baby. Do not use Lotrel if you are pregnant. Lotrel can cause problems in your second or third trimester. Consult your doctor if you are pregnant or plan to get pregnant.

http://viejs.org

Demo: http://bergie.github.com/hallo/annotate.html
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
- Used for Entity-Linking by the Enhancer
- Support for Linked Data
- Support specific Sites via Extensions

Web Resources
- revyu.com
- linkedct.org

Clinical Trials
- Used to Retrieve additional Information for Entities

Specific Resources
- Reuters
- News about Entities

Local Cache
- dbpedia
- revyu cache
- linkedct cache
- reuters cache

Remote Sites
- referenced sites
Stanbol Entityhub

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

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

```python
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);
```

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

@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
- Store
- Semantic Indexer

- Provide annotated Content
- Stanbol will keep Indexes in Sync
- Support for different Semantic Indexes

- plain Content
- enhanced Content

Stanbol will keep Indexes in Sync
Support for different Semantic Indexes

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
SIDER
DrugBank
Dailymed
Diseasome
Life Science

 Scopes
LinkedMovie DB schema
Creative Commons
Music Ontology
LinkedMovie DB schema
Schema.org

 Media
Social Network

 Reasoner
Rules

 Enhancements
Tim’s Profile

 Enhancements

 revyu.com
Stanbol Ontology
Manager, Reasoning and Rules

- Stanbol Rules
  - Recipes: Manage a set of Rules that are executed together
  - Rules are converted to SWRL, Jena Rules or SPARQL CONSTRUCT depending on the available RuleEngine

- Typical Use Cases
  - integrity checks for imported Data
  - harmonize Vocabularies e.g. simple SEO by using schema.org

- Refactor Enhancement Engine
  - allows to execute Recipes on extracted Metadata
Contributions
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
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)

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