Automated Linking Data with

Olivier Grisel
http://www.nuxeo.com

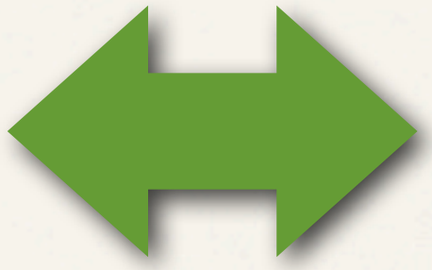
Rupert Westenthaler
http://www.salzburgresearch.at

19. April, 2012

http://www.iks-project.eu

Semantic Content Management with Apache Stanbol

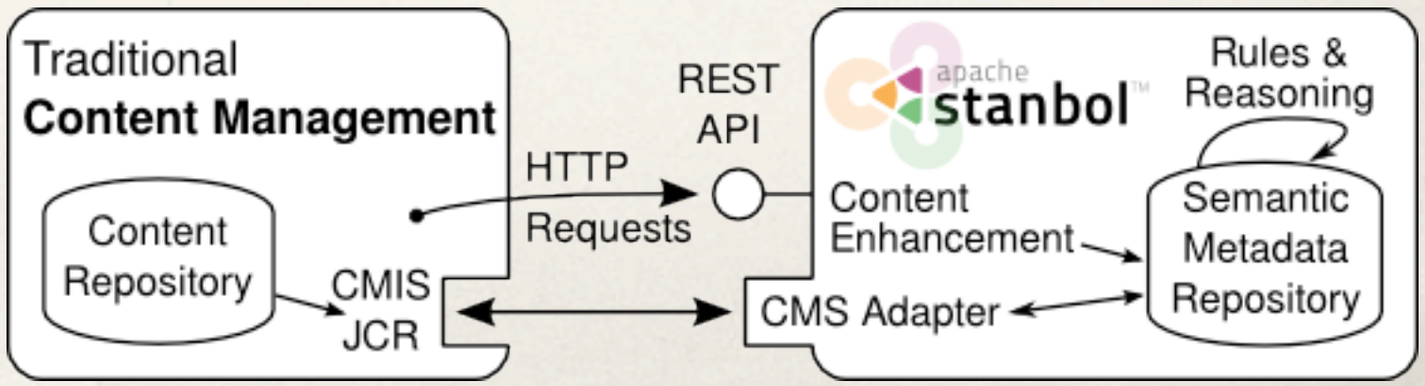
Traditional



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 Enhancer

Get to **know** our **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

Extracted entities

People

Bob Marley

Places

Paris

apache stanbol™

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

Domain Specific Enhancement

Bring our 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

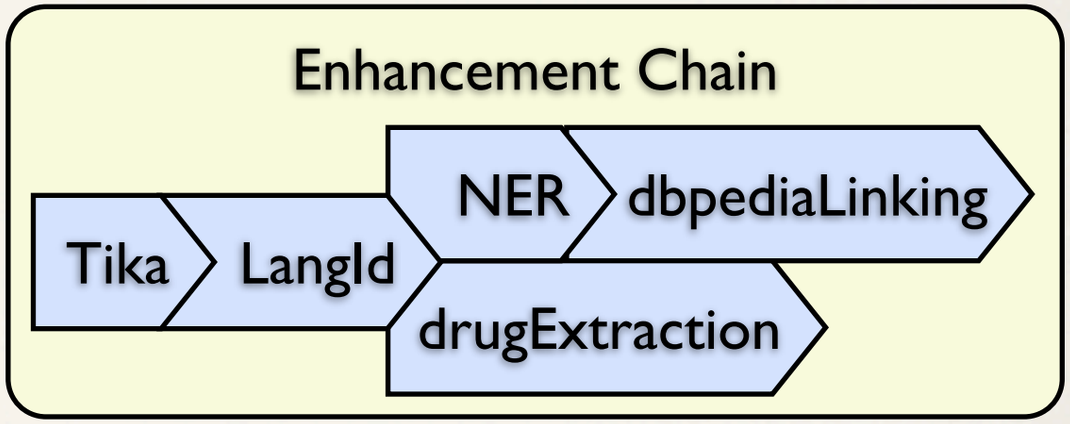
Extracted entities

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

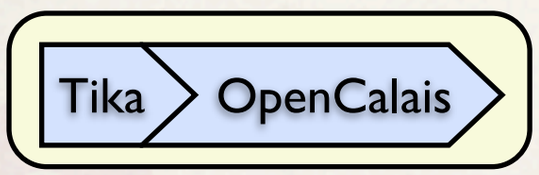
apache stanbol™

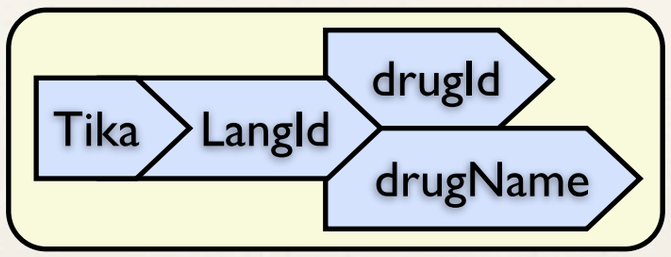
Enhancement Chains

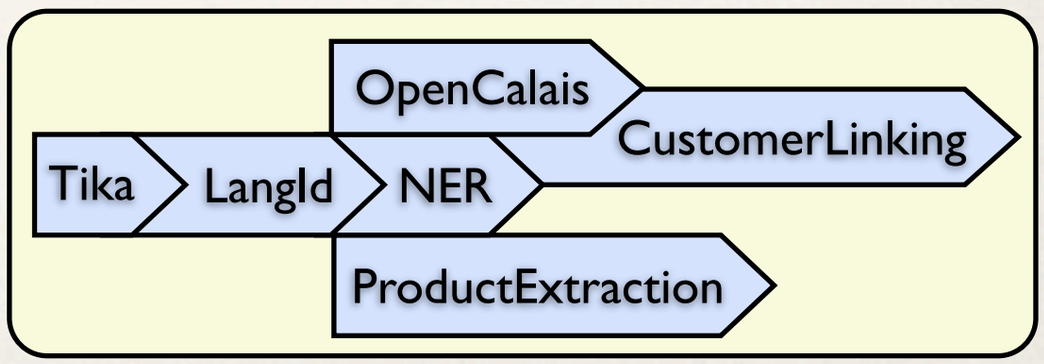
- Define how Content is processed by the Enhancer
  - `/enhancer` calls the default Chain
  - use multiple Chains `/enhancer/chain/{name}`
  - call single EnhancementEngines `/enhancer/engine/{name}`



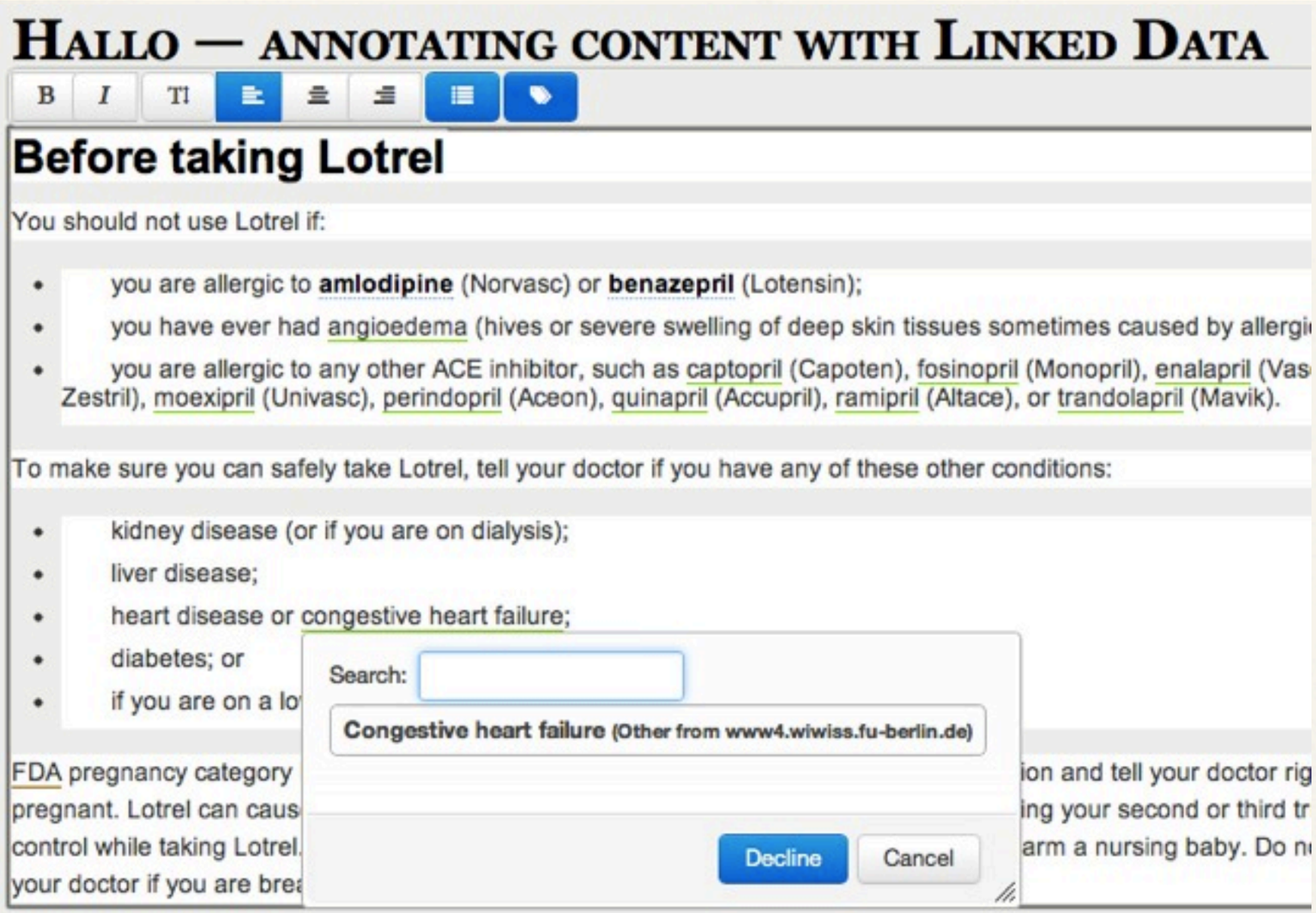
- Some Examples:







Using Stanbol in Web Applications



vie

http://viejs.org

apache stanbol™

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

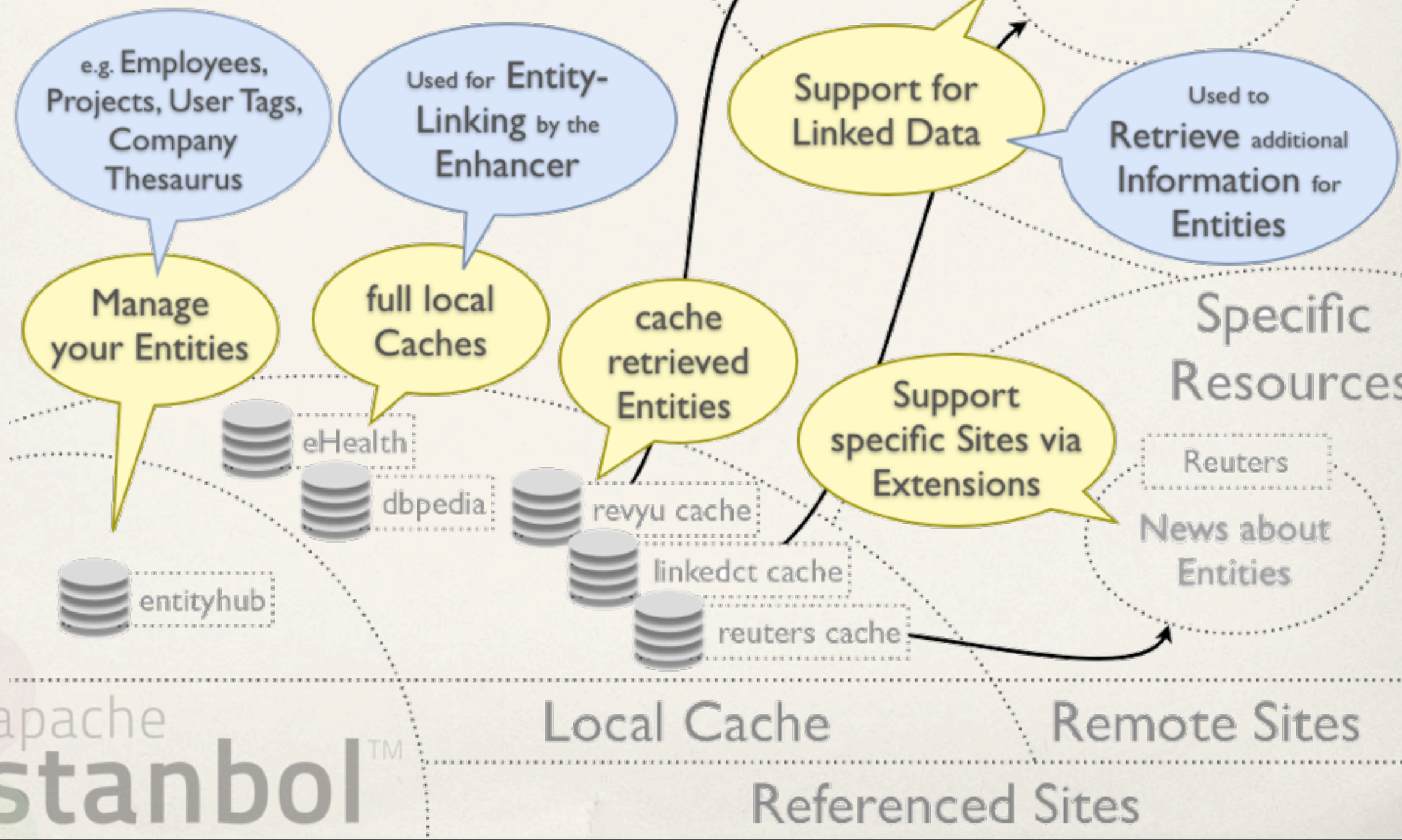
We are looking for

Work with the **Stanbol Community**

- 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

apache stanbol™

Stanbol Entityhub



apache stanbol™

Stanbol Entityhub

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

- Manage multiple Entity Source - Referenced Sites
  - Supports fast local Caches using Apache Solr or apache clerezza
- 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
```

```
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™

You can help by

Work with the Stanbol Community

- Integrate with Data Reconciliation Tools
  - Google Refine:

  - Silk: Entity Link discovery Framework

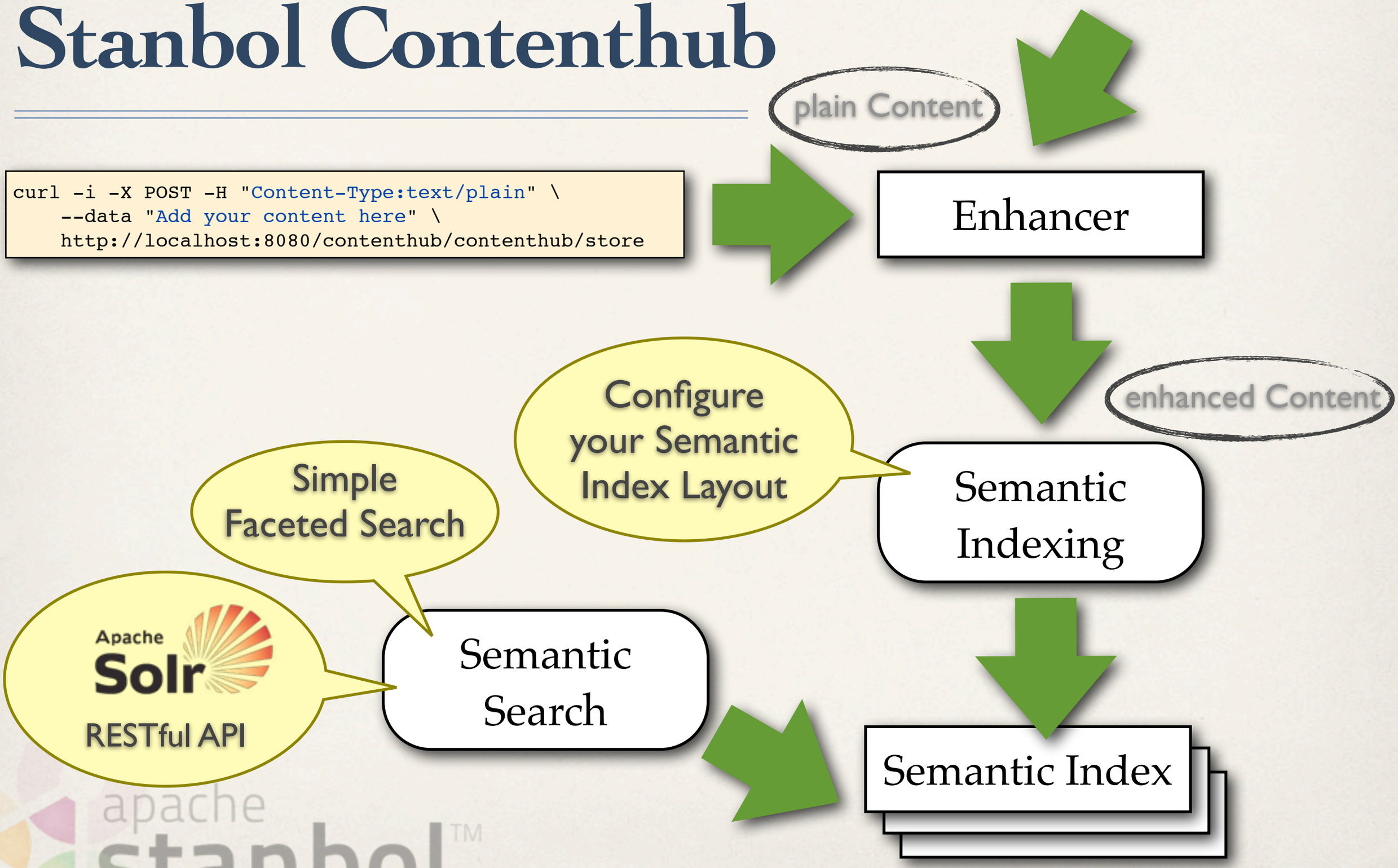
Freie Universität Berlin

- Support for <your> Dataset
  - direct access vie EntityDereferencer implementation
  - provide as Entityhub ReferencedSite (or RDF dump)

apache stanbol™

Stanbol Contenthub

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



apache stanbol™

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 you Search Engine
  - Use the Stanbol Contenthub for semantic indexing
  - Configure Semantic Indexes by using LDpath

apache stanbol™

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;
```

apache stanbol™

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

currently in Development

coming with
Stanbol 0.10
follow STANBOL-471

plain Content

Enhancer

Stanbol will keep Indexes in Sync

Support for different Semantic Indexes

Provide annotated Content

enhanced Content

Store

Semantic Indexer

CMS Adapter

File System

Your CMS

Apache Jackrabbit

Apache Solr

apache clerezza

Jena

CouchDB relax

apache stanbol™

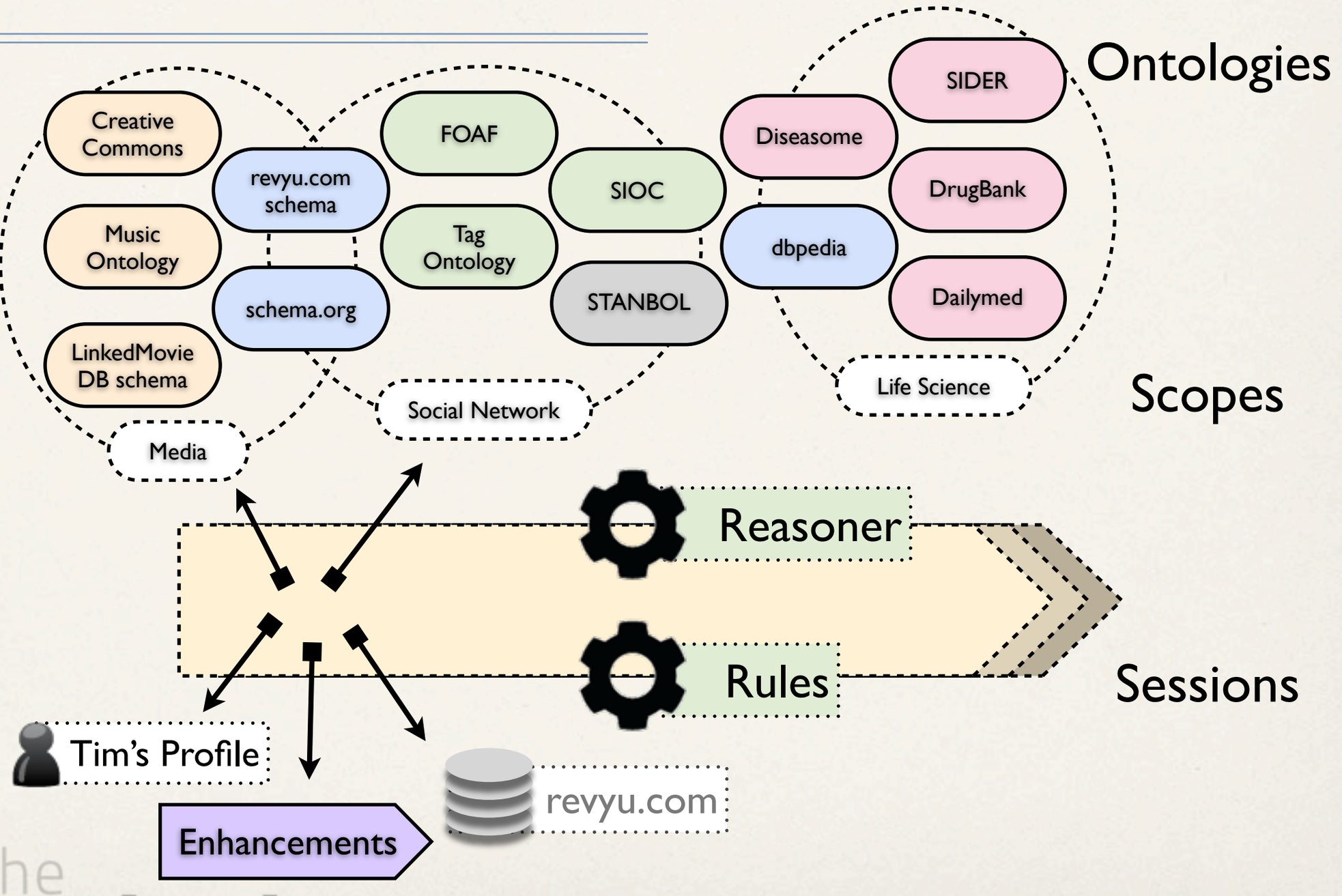
Stanbol Ontology

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



apache stanbol™

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

Work with the Stanbol Community

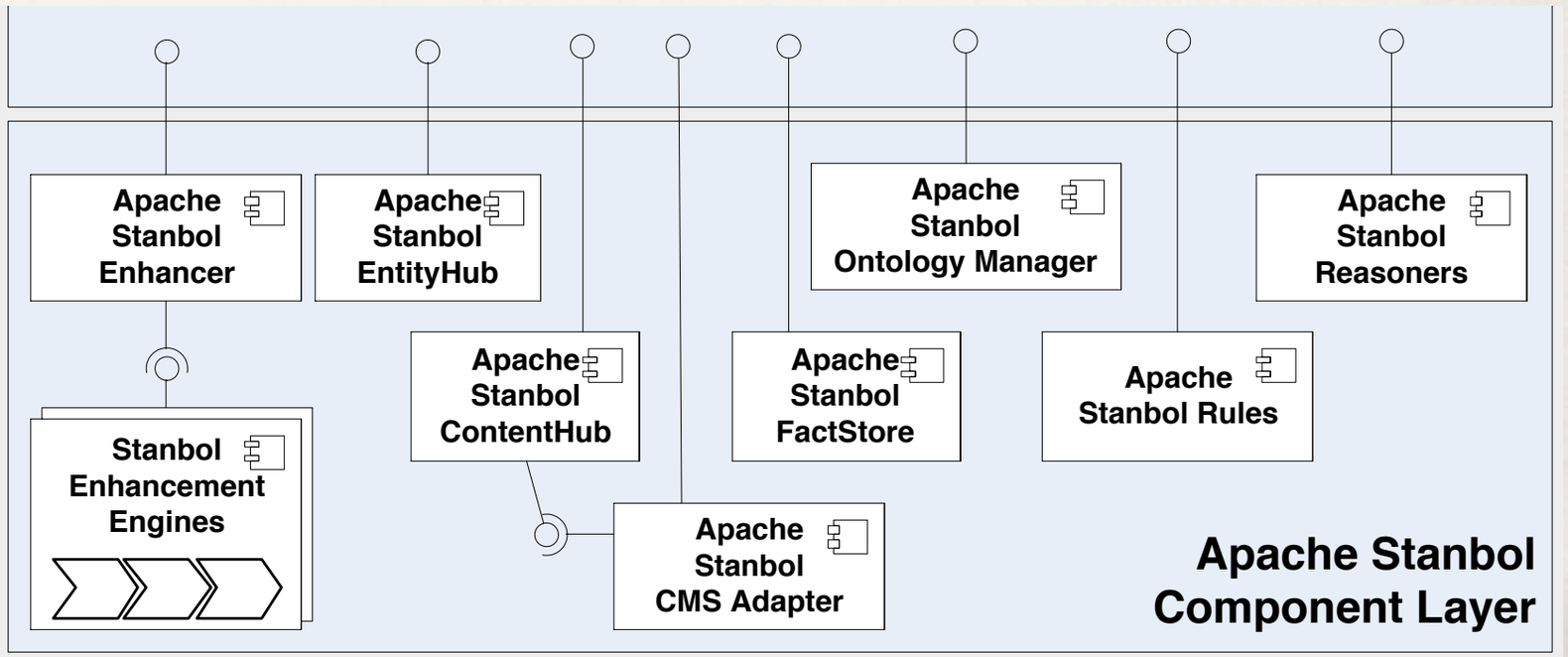
- Share alignment rules across multiple domains
  - Especially with schema.org.
- Benchmarking:
  - how large are the scopes you are managing?
  - Sessions you use in your applications
- Wrap <your> Reasoner/Rule Engine as a Stanbol service

Stanbol Design and Integration Patterns

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



apache stanbol™

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]

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

http://incubator.apache.org/stanbol

stanbol-dev@incubator.apache.org

@westei

Rupert Westenthaler

rwesten@apache.org

salzburgresearch

http://www.salzburgresearch.at

IKS

http://www.iks-project.eu

Co-funded by the European Union