Semantic CMS Community

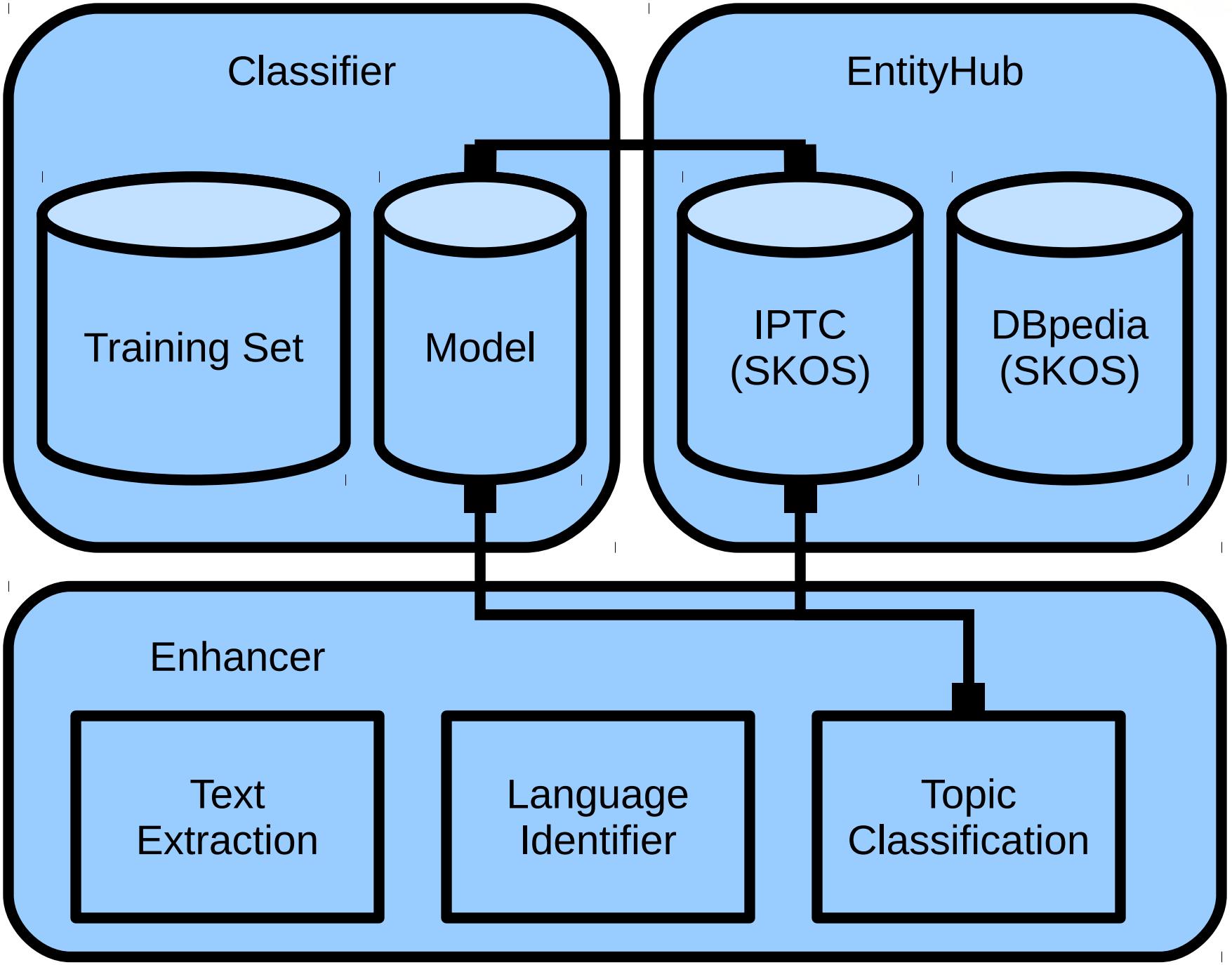
Topic Classification

Olivier Grisel
Nuxeo

March, 2012

The missing documentation ;)

Co-funded by the European Union


Classifier
EntityHub
Training Set
Model
IPTC
(SKOS)
DBpedia
(SKOS)
Enhancer
Text
Extraction
Language
Identifier
Topic
Classification

Building & Deploying

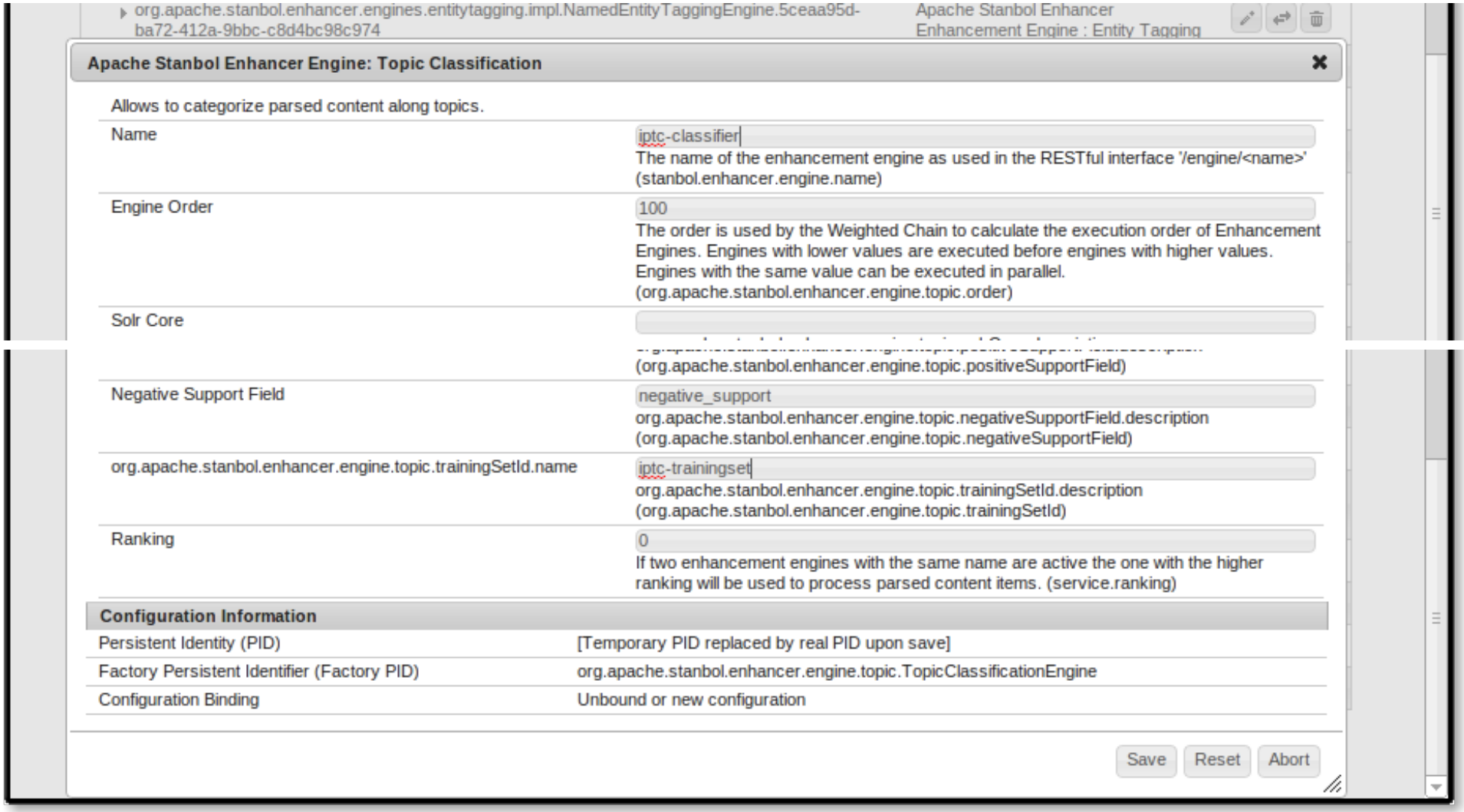
```
cd ~/stanbol/launchers/stable/target/
java -jar org.apache.stanbol.launchers.stable-
0.10.0-incubating-SNAPSHOT.jar -p 9090

cd ~/stanbol/enhancer/engines/topic
mvn install -DskipTests -PinstallBundle \
 -Dsling.url=http://localhost:9090/system/console

cd ~/stanbol/enhancer/topic-web
mvn install -DskipTests -PinstallBundle \
 -Dsling.url=http://localhost:9090/system/console
```

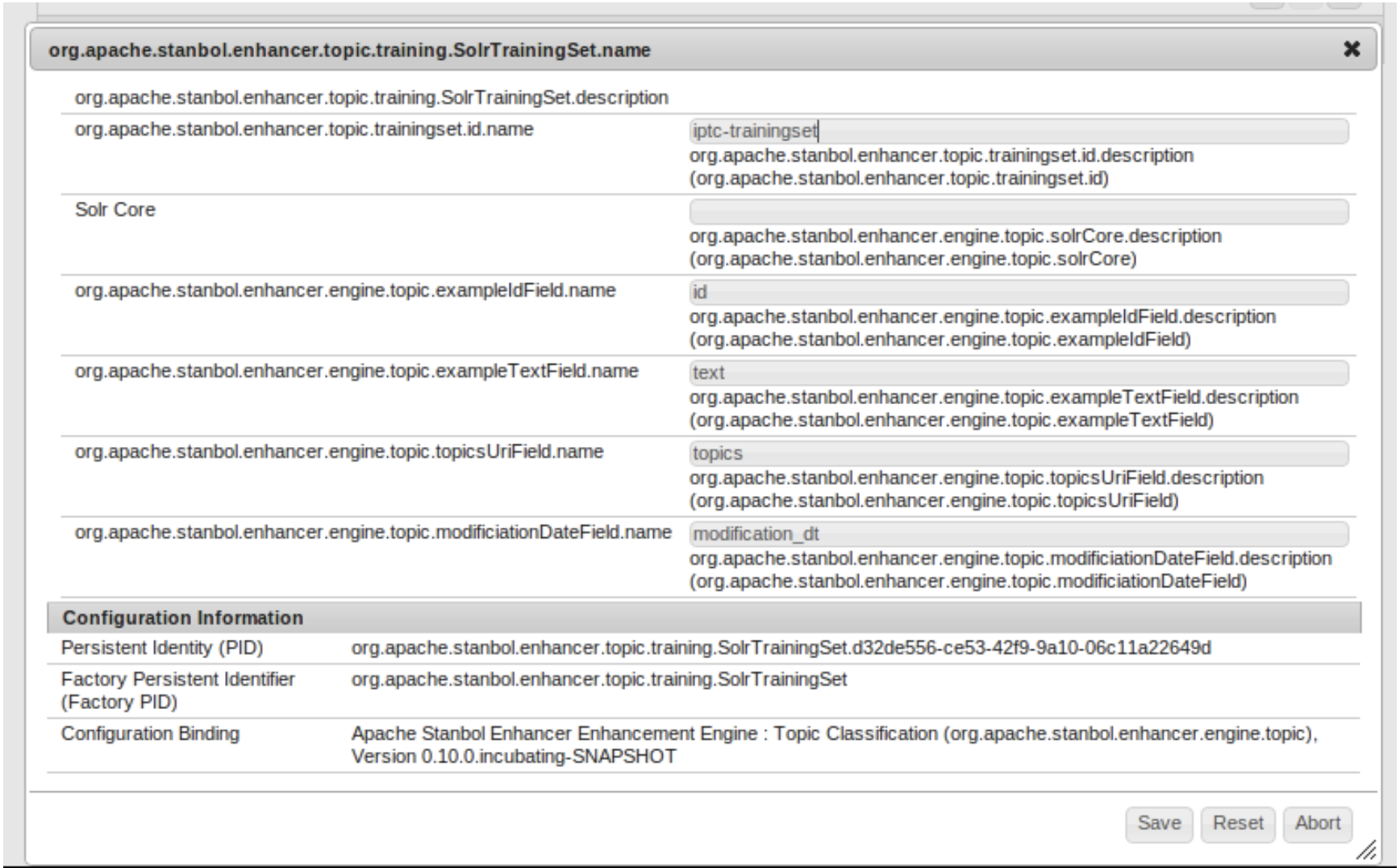
http://localhost:9090/system/console/configMgr

Create a new **TopicClassification** Engine configuration
Add values for **Name** and **Training Set** and hit “Save”

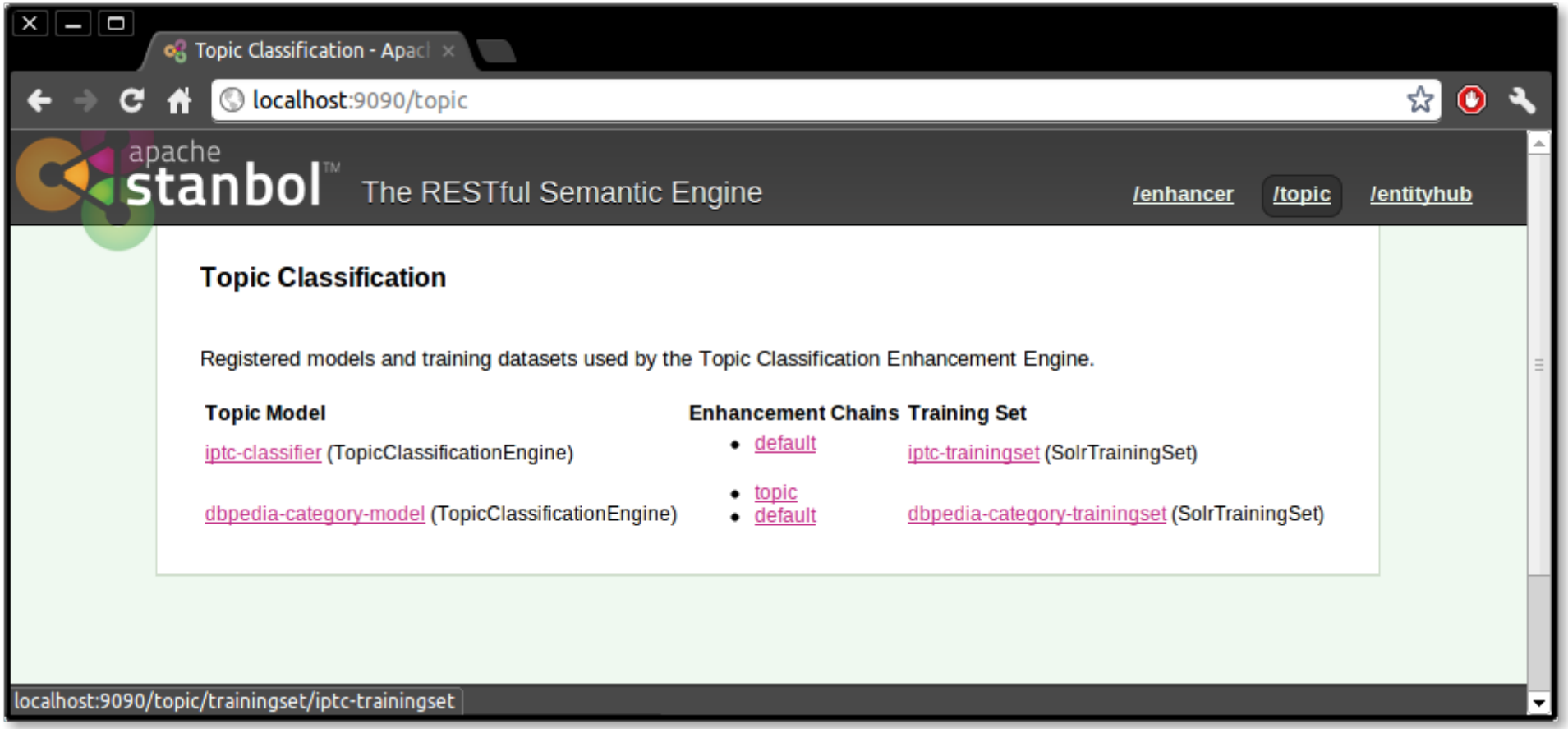


http://localhost:9090/system/console/configMgr

Create a new Topic **Solr Training Set** configuration
Add value for **Name** and hit “Save”



http://localhost:9090/topic



Registering concepts

```
$ curl -X POST http://localhost:8080/topic/model\
 /iptc-classifier/concept?id=concept_1

$ curl -X POST http://localhost:8080/topic/model\
 /iptc-classifier/concept?id=concept_2

$ curl -X POST http://localhost:8080/topic/model\
 /iptc-classifier/concept?id=concept_3\
 &broader=concept_1&broader=concept_2
```

Registering concepts from SKOS

```
$ curl -X POST –-data @iptc-skos.rdf.xml
 http://localhost:8080/topic/model/iptc-classifier
```

Loading a Training Set

```
$ curl -X POST –-data @file_1.txt \
 http://localhost:8080/topic/model\
 /iptc-classifier/trainingset\
 ?example_id=example_1\
 &concept=concept_3&concept=concept_42
```

Importing a NewsML archive

```
$ cd ~/stanbol/enhancer/topic-web/tools

$ python newsmlimporter.py /path/to/newsml/ 1000 \
 http://localhost:9090/topic/model/\
 iptc-classifier/trainingset
Processed news 100/1000 in 01.855s
Processed news 200/1000 in 01.893s
Processed news 300/1000 in 02.012s
Processed news 400/1000 in 01.771s
Processed news 500/1000 in 01.919s
Processed news 600/1000 in 01.823s
Processed news 700/1000 in 01.876s
Processed news 800/1000 in 01.793s
Processed news 900/1000 in 01.875s
Processed news 1000/1000 in 01.817s
```

Training the model

```
$ curl -X POST http://localhost:9090/topic/model/\
 iptc-classifier/trainer?incremental=true
Successfully updated the statistical model(s) of
200 concept(s).

$ curl -X POST http://localhost:9090/topic/model/\
 iptc-classifier/trainer?incremental=true
Successfully updated the statistical model(s) of 0
concept(s).
```

Future work

- Make it possible to plug alternative trainable classifier
  - The present Solr-based Approximate Rocchio Classifier that scales to large numbers of classes
  - OpenNLP Text Categorizer: Averaged Perceptron
  - Mahout's SGD Logistic Regression with the hashing trick for scaling to large number of classes