SYMPOSIUM ON BIAS AND DIVERSITY IN IR

A TESTBED FOR DIVERSIFICATION IN SEARCH

Koblenz, August 31, 2011
Michael Matthews, Barcelona Media/Yahoo! Research
OVERVIEW

• Introduction to LivingKnowledge Testbed – The Diversity Engine
• Getting started – Our first application!
• Adding text analysis
• Adding multimedia analysis
• Evaluation
• Indexing and search
• Developing applications
• Future work
DIVERSITY ENGINE

- Provide collections, annotation tools and an evaluation framework to allow for collaborative and comparable research
- Supports indexing and searching on a wide variety of document annotations including entities, bias, trust, polarity, and multimedia features
- Support development of bias and diversity aware applications
ARCHITECTURE

Document Collections → Analysis Pipeline → Index/Search → Application Development

- Prediction of Community Acceptance
- Sentiment in Comments
- Polarizing Videos
- Topic of Videos

Evaluation Framework

ARC Crawls

Yahoo! News

NYT
DESIGN DECISIONS

• Use Open Source tools when available
• Programming Language - Java 1.6
• Data format – LK XML
• Analysis tools Operating System – Linux (any software language)
• Indexing/Search - Solr
• GUI – JSP, HTML, JavaScript, CSS
LK-XML format.
DOCUMENT COLLECTIONS

• Supported Formats - ARC (Internet Memory Crawls), Text, HTML. Kyoto, BBN, NYT

• Collections
  – Testing Examples included with Diversity Engine
  – Large ARCs available from Internet Memory
  – Converters provided for other collections (MPQA, BBN, NYT) that have licensing restrictions
# ANALYSIS MODULES

## Image Processing
- Face Detection
- Naturalness
- Colourfulness
- SIFT Features
- Photo/Cartoon/CG
- Histogram
- EXIF
- City/Landscape
- Tone
- Photomontage
- Face Tampering
- Annotations
- Sentiment
- Image Clustering

## Text Processing
- OpenNLP Entities
- Wikipedia Places
- Disambiguated Entities
- Wikipedia People
- Document Layout
- URLs
- Readability4J

## Text Annotation Processing
- POS
- Syntax & Semantics
- Dictionary
- TimeML
- Sentence Subjectivity
- Phrases
- Statements
- SuperSense Tagger
- Quotes
- Subjective Expressions
- RDFa Injection

## Image Annotation Processing
- Vector Quantisation
INDEXING/SEARCH

• Solr
  – Enterprise search platform built on top of Lucene
  – Xml input and output allows for easy integration with Diversity Engine
  – Plug-in framework allows customization
  – Built-in facet capabilities support indexing and searching on annotations

• Integration
  – Converter from LK XML – Solr XML
  – Plug-in for facet ranking and speed improvements
APPLICATION DEVELOPMENT

• Basis for LivingKnowledge Applications
  – Future Predictor
  – Media Content Analysis

• Support development – coding required!

• Real World Problems
  – HTML Extraction
  – Scaling to Large Collections
  – Provenance
  – Some pluggable GUI components
  – Examples to ease learning curve
APPLICATION DEVELOPMENT
**APPLICATION DEVELOPMENT**

Alain Lamassoure, a French conservative MEP and confidante of French presidential front-runner Nicolas Sarkozy, will in plenary on Wednesday present his ideas on how the EU should raise more money directly in future - called "own resources" in Brussels jargon - instead of relying on handouts from stingy EU states.

The UN’s envoy on Kosovo, Martti Ahtisaari, will on Tuesday tell senior EU diplomats in the Political and Security Committee what he is proposing to the UN security council on the future status of Kosovo. EU foreign ministers will meet informally in Bremen, Germany on Friday to debate his ideas, while MEPs will discuss Kosovo in plenary on Wednesday, reacting to a report by Dutch green member Joos Lagendijk.

The US and Europe’s UN security council veto powers, France and the UK, are keen to put the UN-administered Serbian province firmly on the road to independence. But Russia has threatened to veto the move if Belgrade does not agree, with Belgrade so far expressing nothing but contempt for Mr Ahtisaari’s ideas. The situation could cause “choosers” in the Western Balkans if it is resolved badly, enlargement commissioner Olli Rehn has warned.

On Thursday the EU’s de facto foreign minister, Javier Solana, will brief MEPs on US plans to install interceptor missiles and radar bases in Poland and the Czech republic. Washington says the scheme will protect the whole continent, including Russia, from rogue states like Iran. But Russia says it is designed to undermine its nuclear deterrent, with Russia-wary states like Germany worried about prospects of a new Cold War.

German foreign minister Frank-Walter Steinmeier will on Tuesday in Astana.

<table>
<thead>
<tr>
<th>Question</th>
<th>Value</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. Media ID</td>
<td>Conf: Unknown</td>
<td>0</td>
</tr>
<tr>
<td>Q2. Date of Publication</td>
<td>DateTime</td>
<td>No options</td>
</tr>
<tr>
<td>Q3. Publication periodicity</td>
<td>Conf: n/a</td>
<td>Daily</td>
</tr>
<tr>
<td>Q4. Contains Advertisements</td>
<td>Conf: Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>Q5. Range of Distribution</td>
<td>Conf: n/a</td>
<td>Local</td>
</tr>
</tbody>
</table>
EVALUATION FRAMEWORK

• Framework for the evaluation of analysis tools
• Evaluates any possible annotation pipeline
• Measures correctness and quality
• Outputs Precision + Recall
• Compares annotation output of pipeline with ground truth data
OUR FIRST APPLICATION

- Download Diversity Engine release from SourceForge
- `tar xzvf [release file]`
- `cd testbed`
- `ant build`
- `apps/testbed/conf/testbed/tutorial-application.xml`
- What happened?
  - 197 text files and 127 images files converted from arc format to LK XML and stored in devapps/example/data/lkxml
  - 2 annotators were run over collection
    - OpenNLP for tokenization, sentence splitting, Pos tags
    - SST named entity recognizer
    - Results stored in devapps/example/data/lkxml
  - Files were converted to Solr xml format and indexed using solr
    - Solr XML stored to devapps/example/data/solr
  - HTML Visualization Files stored in devapps/example/data/html
- `ant deploy-testbed`
EXAMPLE SOLR OUTPUT

http://localhost:8983/solr/select/?q=putin

```xml
<response>
  <lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">19</int>
    <lst name="params"></lst>
  </lst>
  <result name="response" numFound="33" start="0">
    <doc>
      <str name="id">
        EA-EUElections2009-euobserver-0729-20090729085530-00000.arc.10283109
      </str>
      <arr name="keywords">
        <str>org:Putin</str>
        <str>loc:Moscow</str>
        <str>loc:Germany</str>
        <str>org:Nord Stream</str>
        <str>loc:Europe</str>
        <str>loc:Poland</str>
        <str>per:Mr Putin</str>
        <str>loc:Kaliningrad</str>
        <str>org:Gazprom</str>
        <str>loc:Russia</str>
      </arr>
      +<str name="lkttext"></str>
      +<str name="lkttitle">
        Putin questions Baltic pipeline, as oil and gas prices drop
      </str>
      <arr name="loc">
        <str>Germany</str>
        <str>Baltic Sea</str>
      </arr>
    </doc>
  </result>
</response>
```
EXAMPLE APPLICATION

http://localhost:8983/testbed/results.jsp?query=putin
Natural Language View

ENTITIES

<table>
<thead>
<tr>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Title: Russia and Ukraine face EU sanctions threat</td>
</tr>
<tr>
<td>2. The EU has raised the threat of sanctions, but it is unclear what it can do or who to blame (Photo: gazettefrance.com)</td>
</tr>
<tr>
<td>3. PHILIPPA RUNNER</td>
</tr>
<tr>
<td>4. 16.01.2009</td>
</tr>
<tr>
<td>5. @ 17:18 CET</td>
</tr>
<tr>
<td>6. EUOBSERVER / BRUSSELS - The European Commission on Friday (16 January) threatened to take unspecified sanctions against Russia and Ukraine unless gas flows to the EU resume after the weekend</td>
</tr>
<tr>
<td>7. &quot;It's a situation the seriousness of which goes beyond the specific issue of gas,&quot; Commission spokesman Johannes Laitenberger said</td>
</tr>
<tr>
<td>8. &quot;As of next week, if the gas does not flow again, we will have to look point by point at our relations with Russia and with Ukraine and assess in each case whether we can do business as usual.&quot;</td>
</tr>
<tr>
<td>9. &quot;The meetings in the coming days offer the best and last chance for Russia and Ukraine to demonstrate they are serious about resolving the dispute.&quot;</td>
</tr>
<tr>
<td>10. Russian President Dmitry Medvedev has invited Ukrainian President Viktor Yushchenko and EU leaders from the 18 countries affected by the dispute to hold a crisis &quot;summit&quot; in Moscow on Saturday.</td>
</tr>
<tr>
<td>11. It was uncertain on Friday if the Medvedev event would take place after Mr Yushchenko opted to meet senior politicians from Poland, Slovakia and Moldavia in Kiev instead, while arguing that any summit should be held on neutral ground in Brussels or Prague.</td>
</tr>
</tbody>
</table>
<lk-application logDir="log" appDir="devapps/example">
  <corpus dir="corpora/examples/smallarc" format="arc"/>
  <image-pipeline>
    <annotators>
      <annotator exec="./opennlp"/>
      <annotator exec="./sst"/>
    </annotators>
  </image-pipeline>
  <pipeline>
    <annotators>
      <annotator exec="./opennlp"/>
      <annotator exec="./sst"/>
    </annotators>
  </pipeline>
</lk-application>

<indexer solrHomeDir="solr/solr"
  solrDataDir="solr/solr/data"
  converter="conf/testbed/tutorial-lk2solr.xml"/>

<searcher appTitle="LivingKnowledge - Example Application"
  appShortTitle="Example Application"
  appUrl="http://localhost:8983/solr/">
  <facets>
    <facet field="per" description="Person"/>
    <facet field="loc" description="Location"/>
  </facets>
</searcher>
</lk-application>
TEXT ANALYSIS

<pipeline>
  <annotators>
    <annotator exec="./opennlp"/>
    <annotator exec="./sst"/>
  </annotators>
</pipeline>

<pipeline>
  <annotators>
    <annotator exec="./opennlp"/>
    <annotator exec="./sst"/>
    <annotator exec="./facts"/>
    <annotator exec="./unitn_tagger"/>
    <annotator exec="./unitn_subjexpr"/>
  </annotators>
</pipeline>

apps/testbed -run pipeline conf/testbed/tutorial-application.xml
apps/testbed -run visualization conf/testbed/tutorial-application.xml
TEXT ANALYSIS - FACTS

devapps/example/data/lkxml/EA-EUElections2009-euobserver-0729-20090729085530-00000.arc.15521713.facts.xml

- `<lk-annotation>`
- `<meta-info>`
  - `<tag name="base">` EA-EUElections2009-euobserver-0729-20090729085530-00000.arc.15521713.lktext.xml `</tag>`
  - `<tag name="annotator">Advanced-Facts-Extractor-v3</tag>`
- `<annotation scope="#EA-EUElections2009-euobserver-0729-20090729085530-00000.arc.15521713.lktext.xml" p`
### Natural Language View

<table>
<thead>
<tr>
<th>Sentence #</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Title: EUobserver</td>
</tr>
<tr>
<td>2</td>
<td>22.07.2009</td>
</tr>
<tr>
<td>3</td>
<td>@ 09:16 CET</td>
</tr>
<tr>
<td>4</td>
<td>Voices have been heard recently in European institutions conveying discontent with the recently adopted amendment to Slovakia’s State Language Act. But under a flood of criticism, arguments and counterarguments, the substance has been lost, writes Miroslav Lajčák, Slovak Minister for Foreign Affairs.</td>
</tr>
<tr>
<td>5</td>
<td>14.07.2009</td>
</tr>
<tr>
<td>6</td>
<td>@ 08:04 CET</td>
</tr>
<tr>
<td>7</td>
<td>Most Europeans have probably never heard of the European Defence Agency, which celebrates its fifth anniversary this month. Yet, this little known EU agency has slowly created a basis for the development of common military projects, writes Frank Slijper.</td>
</tr>
<tr>
<td>8</td>
<td>24.07.2009</td>
</tr>
<tr>
<td>9</td>
<td>@ 08:15 CET</td>
</tr>
<tr>
<td>10</td>
<td>The German Constitutional Court issued a remarkable verdict on 30 June. The press described it as an approval of the ratification of the Lisbon Treaty.</td>
</tr>
<tr>
<td>11</td>
<td>However, careful reading of the judgement shows that it is a fundamental rejection of the core constitutional content of the Treaty, writes former MEP Jens-Peter Bond</td>
</tr>
</tbody>
</table>
<image-pipeline>
  <annotators>
    <annotator exec="./soton haarfacedetector"/>
  </annotators>
</pipeline>

<pipeline>
  <annotators>
    <annotator exec="./opennlp"/>
    <annotator exec="./sst"/>
    <annotator exec="./facts"/>
    <annotator exec="./unitn_tagger"/>
    <annotator exec="./unitn_subjexpr"/>
    <annotator exec="./imageannots"/>
  </annotators>
</pipeline>

apps/testbed -run pipeline,image-pipeline -pipeline imageannots
conf/testbed/tutorial-application.xml

ls devapps/example/data/lkxml/img/*
ANALYSIS API

- Documents in LK XML format
- Annotators passed a single document directory – They should add annotations for each document in directory
- Files will have consistent naming convention
  - LkText file = id + “.lktext.xml”
  - LkMedia = id + “.lkmedia.xml”
  - LkAnnotation = id + “.” + annotatorId + “.xml”
- Annotators will be processed sequentially in the order listed in the XML file
- Annotators can be written in any language but must run on Linux – Helper classes will exist for Java, but there is no obligation to use them.
- Add application calling your new annotator to apps directory
- Add your application to the configuration file as before
ANALYSIS API – JAVA

- Extend class org.diversityengine.annotator.AbstractAnnotator
- Implement Methods
  - getName()
  - getType() - TEXT OR IMAGE
- For Image Analysis implement
  - LkAnnotation getLkAnnotation(ImageDocument document)
- For Text Analysis implement
  - LkAnnotation getLkAnnotation(TextDocument document)
- In main, instantiate and call annotator
  - NewAnnotator annotator = new NewAnnotator()
  - annotator.processDirectory(args[0]);
- Add application calling your new annotator to apps directory
- Add your application to the configuration file as before
EVALUATION

Evaluation works with same configuration file. Simply add evaluation element

```xml
<lk-application logDir="log" appDir="devapps/evaluation">
  <corpus dir="corpora/evaluation/sst/text/" format="bbn"/>
  <pipeline>
    <annotators>
      <annotator exec="./sst"/>
    </annotators>
  </pipeline>
  <evaluation evalDir="evaluation/sst/">
    <evaluator provides="ENTITIES"
                 goldDir="corpora/evaluation/sst/gold/"
                 goldAnnotator="sstgold"
                 annotator="sst"/>
  </evaluation>
</lk-application>
```

apps/testbed/conf/evaluation/sst.xml
<evaluation goldDir="/home/mikemat/code/livingknowledge/WP6/testbed/corpora/evaluation/sst/gold/" lkDir="/home/mikemat/code/livingknowledge/WP6/testbed/devapps/evaluation/data/lkxml" annotation="sst" goldAnnotation="sstgold" provides="ENTITIES">
  <docs>
    <doc id="WSJ0375" N="19" tp="18" fp="1" fn="1" />
    <doc id="WSJ0380" N="19" tp="15" fp="4" fn="1" />
    <doc id="WSJ0376" N="72" tp="61" fp="11" fn="7" />
    <doc id="WSJ0377" N="26" tp="17" fp="9" fn="6" />
    <doc id="WSJ0378" N="10" tp="10" fp="0" fn="0" />
    <doc id="WSJ0379" N="24" tp="19" fp="5" fn="2" />
  </docs>
  <totals N="170" tp="140" fp="30" fn="17" p="0.8235294117647058" r="0.89171974522293" f="0.8562691131498471" />
</evaluation>

`cat evaluation/sst/sst.ENTITIES.xml`
INDEXING AND SEARCH

• Search Engines - Traditional
  – Bag-of-words representation
  – Inverted index (words -> documents) for efficiency
  – 10 docs ranked according tf-idf similarity with query

• Search Engines – Today
  – Much metadata associated with documents
  – Ranking based on 100s of features (date, location, pagerank, click data, etc, personalization)
  – Richer display
    • Facets for exploratory search
    • Answers when appropriate
    • etc..
  – Many open source options - Lucene/Solr most widely used
APACHE LUCENE/SOLR

Lucene/Solr

Diagram by Yonik Seeley
FACETED SEARCH

Diagram by Yonik Seeley
FACETED SEARCH

• Summarize query results aggregation properties of returned pages
  – price ranges for product query
  – related people or locations for news query

• Exploratory Search
  – Show documents that matching the query term and a selected facet
  – Make inferences not clear from simple document list

• Living Knowledge Analysis is modeled very well by facets
  – Topics as determined by entity and fact extraction
  – Location and Time diversity dimensions
  – Opinions as determined by opinion extraction
LK XML TO SOLR

- Solr has well defined XML input format for adding new documents
- Diversity Engine provides a simple language to map LX XML to Solr XML
LKV2SOLR CONVERSION

```xml
<indexer
  solrHomeDir="solr/solr"
  solrDataDir="solr/solr/data"
  converter="conf/testbed/tutorial-lk2solr.xml"/>

<lktosolr>
  <field
    solr="per"
    annotation="ENTITIES_CLEAN"
    value="$text"
    filter="org.diversityengine.solr.converter.filters.PerValueFilter"/>
  <field
    solr="loc"
    annotation="ENTITIES_CLEAN"
    value="$text"
    filter="org.diversityengine.solr.converter.filters.LocValueFilter"/>
  <field
    solr="keywords"
    annotation="TOP_ENTITIES"
    value="$text"
  />
  <field
    solr="pubdate"
    annotation="metainfo:lktext"
    value="date"
    type="date"/>
</lktosolr>
```

- **solr** - Name of the field in solr
- **annotation** - Name of the LKXML Annotation
- **value** - Value of annotation
- **filter** - Allows post processing on annotation
- **type** - Only Date supported currently
ADDING FACTS TO INDEX

```xml
<lktosolr>
  <field solr="per" annotation="ENTITIES_CLEAN" value="$text"
        filter="org.diversityengine.solr.converter.filters.PerValueFilter"/>
  <field solr="loc" annotation="ENTITIES_CLEAN" value="$text"
        filter="org.diversityengine.solr.converter.filters.LocValueFilter"/>
  <field solr="keywords" annotation="TOP_ENTITIES" value="$text" />
  <field solr="pubdate" annotation="metainfo:lktext" value="date"
        type="date"/>
  <field solr="yago" annotation="yago-entities" value="$text" />
  <field solr="yago-country" annotation="facts"
        value="xpath://entity-information[facts/type/text()=
              'wordnet_country_108544813']/id/text()" />
</lktosolr>
```

```
apps/testbed -run convert-solr conf/testbed/tutorial-application.xml
ls devapps/example/data/solr/*

apps/testbed -run index conf/testbed/tutorial-application.xml
```
FACTS TO SOLR

<field solr="yago" annotation="yago-entities" value="$text" />

---

EUFacts 2009-07-29 07:58:55.103764

<doc boost="1.0">  
  <field name="id" boost="1.0">  
    EU-EUElections2009-euobserver-0729-20090729085530-00000-arc.15521713.1ktext.xml</field>
  <field name="ktext" boost="1.0">EUobserver 22.07.2009 @ 09:18 CET Voices have been heard recently in European inst
guments and counterarguments, and the substance has been lost, write MMiroslav Lajca
Agency, which celebrates its fifth anniversary this month. Yet, this little known EU agen
Constitutional Court issued a remarkable verdict on 30 June. The press described it as a
cor constitutional treaty, writes former MEP Jens-Peter Bonde.
  </field>
</doc>
FACTS TO SOLR

```xml
<field solr="yago-country" annotation="facts"
value="xpath:/entity-information[facts/type/text()=
'wordnet_country_108544813']/id/text()" />
```

```xml
<doc boost="1.0">
  <field name="id" boost="1.0">EA-EUElections2009-euobserver-0729-20090729085530-00000.arc.15521713</field>
  <field name="ltxt" boost="1.0">EUobserver 22.07.2009 @ 09:18 CET Voices have been heard recently in European inst
arguments and counterarguments, the substance has been lost, writes MIroslav Lajca
Agency, which celebrates its fifth anniversary this month. Yet, this little known EU agent
Constitutional Court issued a remarkable verdict on 30 June. The press described it as a
core concept of the Treaty, writes former MEP Jens-Peter Bonde.</field>
  <field name="url" boost="1.0">http://euobserver.com/7</field>
  <field name="ltxt" boost="1.0">EUobserver</field>
  <field name="yago-country" boost="1.0">Slovakia</field>
  <field name="per" boost="1.0">Frank Slipper</field>
  <field name="pubdate" boost="1.0">20110826</field>
  <field name="yago" boost="1.0">Federal Constitutional Court of Germany</field>
</doc>
```
ADDING IMAGES TO INDEX

```
<lktsolr>
  <field solr="per" annotation="ENTITIES_CLEAN" value="$text"
        filter="org.diversityengine.solr.converter.filters.PerValueFilter"/>
  <field solr="loc" annotation="ENTITIES_CLEAN" value="$text"
        filter="org.diversityengine.solr.converter.filters.LocValueFilter"/>
  <field solr="keywords" annotation="TOP_ENTITIES" value="$text" />
  <field solr="yago" annotation="yago-entities" value="$text" />
  <field solr="yago-country" annotation="facts"
         value="xpath:/entityinformation[facts/type/text() = 'wordnet_country_108544813']/id/text()" />
  <field solr="pubdate" annotation="metainfo:lktext" value="date"
         type="date"/>
  <field solr="image" annotation="IMAGE_ANNOTS" value="$text" />
  <field solr="bestimage" annotation="BEST_IMAGES" value="$text" />
</lktsolr>
```

```
apps/testbed -run convert-solr conf/testbed/tutorial-application.xml
ls devapps/example/data/solr/*

apps/testbed -run index conf/testbed/tutorial-application.xml
```
APPLICATION DEVELOPMENT

- Examples
- HTML Extraction
- Scaling to Large Collections
- Provenance
- Some pluggable GUI components
FACT/IMAGE APPLICATION

```xml
<searcher appTitle="LivingKnowledge - Example Application"
  appShortTitle="Example Application"
  appUrl="http://localhost:8983/solr/">
  <facets>
    <facet field="yago" description="Yago"/>
    <facet field="yago-country" description="Country"/>
    <facet field="per" description="Person"/>
    <facet field="loc" description="Location"/>
    <facet field="image" description="Images"/>
  </facets>
</searcher>

ant deploy-testbed
```
FACT/IMAGE APPLICATION

http://localhost:8983/testbed/results.jsp?query=putin
OPINION APPLICATION

Opinions are at sentence level, not document level – same analysis, but different indexing

cat conf/testbed/tutorial-lk2solr-sentence.xml

```
<lktsolr solDoc="SENTENCES" contextSize="1">
  <field solr="per" annotation="ENTITIES_CLEAN" value="$text"
    filter="org.diversityengine.solr.converter.filters.PerValueFilter" source="solrdoc" />
  <field solr="loc" annotation="ENTITIES_CLEAN" value="$text"
    filter="org.diversityengine.solr.converter.filters.LocValueFilter" source="solrdoc" />
  <field solr="keywords" annotation="TOP_ENTITIES" value="$text" source="solrdoc" />
  <field solr="yago" annotation="yago-entities" value="$text" source="solrdoc" />
  <field solr="image" annotation="IMAGE_ANNOTS" value="$text" source="solrdoc" />
  <field solr="bestimage" annotation="BEST_IMAGES" value="$text" source="solrdoc" />
  <field solr="pubdate" annotation="metainfo:lktext" value="date" type="date" />
  <field solr="polarity"
    annotation="MPQA-expressive-subjectivity,MPQA-direct-subjective"
    value="xpath:/node()[@pol]/@pol" source="solrdoc"
    filter="org.diversityengine.solr.converter.filters.PolarityValueFilter" />
  <field solr="pol-int"
    annotation="MPQA-expressive-subjectivity,MPQA-direct-subjective"
    value="xpath:concat(/node()@[pol and @int]/@pol,/node()@[int and @pol]/@int)"
    source="solrdoc" />
</lktsolr>
```

apps/testbed -run convert-solr,index
conf/testbed/tutorial-application-sentence.xml
ls devapps/example/data/solr/*
Voices have been heard recently in European institutions conveying discontent with the recently adopted amendment to Slovakia’s State Language Act. 14.07.2009

But under a flood of criticism, arguments and counterarguments, the substance has been lost, writes MMiroslav Lajcak, Slovak Minister for Foreign Affairs.
modify_webapp\WEB-INF\web.xml


  <description>
    LivingKnowledge Testbed Example Application
  </description>
  <display-name>Testbed Examples</display-name>

  <context-param>
    <param-name>applicationDef</param-name>
    <param-value>conf\testbed\tutorial-application-sentence.xml</param-value>
    <description>The Living Knowledge application description XML file</description>
  </context-param>

</web-app>

ant deploy-testbed
OPINION APPLICATION

http://localhost:8983/testbed/results.jsp?query=putin
HTML EXTRACTION

• Boilerplate can lead to false positive results and inaccurate facet aggregation
  – Real example – before extraction developed, most common person for most queries was in a top story title (on all pages) the day of the crawl!
• Titles, Authors and Dates are important for bias and diversity aware search
PROVENANCE

• How an annotation is derived is often as important as the annotation itself
  – Users want to verify results
  – Developers need to validate results

• Open Provenance provides an open source solution

• Testbed annotations can be extended with Open Provenance chains
Provenance Diagram
SCALING TO LARGE COLLECTIONS

• In the real world, even “small” datasets have million of documents
• NLP/Image processing is expensive – 1 doc/sec = 11 days for 1 million docs!
• Hadoop Mapper allows for scaling – scales linearly with number of machines
• ZipCollection writer allows partitioning data into subsets for processing
COMPONENTS- OPINIONS
COMPONENTS - TIME

Help Yugoslavia Look Forward
Help Yugoslavia Look Forward How many countries will there be in Europe after 1992? Seven, according to the bitter gibe circulating in Croatia: the European Community and the six republics that now make up Yugoslavia. It's no joke. Things could turn out that way if Slobodan Milosevic and his Communist cohorts in Belgrade continue to hold back change.
COMPONENTS - GEO
FUTURE WORK

• More components
• Maven to manage dependencies
• Better integration of Timeline and Geo visualization components
• Integration of ranking algorithms
• Better Documentation 😊
Thanks!

• LivingKnowledge Partners!
• You for coming!!
• Questions?