Linking Related Web Content to Videos

Ivo Lašek
Faculty of Information Technology
Czech Technical University in Prague
lasekivo@fit.cvut.cz

14. 3. 2013 – Knowledge Engineering Group Seminar


Web and TV
seamlessly interlinked =
LinkedTV

Scenario 1:

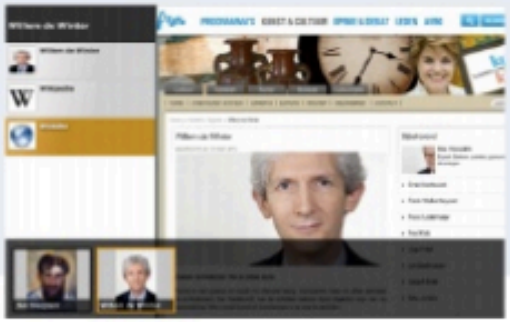
Interactive News Show

- Professional news content produced by RBB
- Seed content: local news show "rbb Aktuell"

Scenario 2:

Hyperlinked Documentary

- Cultural content from S&V (1700 hours of cultural heritage AV-content under CCL)
- Seed content: "Antique Roadshow"



Scenario 3:

Media Arts

- Content and Performance by NUMEDIART Institute for New Media Art Technology
- Mons: European Capital of Culture 2015.

CONCEPT IN PLAYER

Cubism

Expressionism

Fauvism

FACETS / PROPERTIES OF CONCEPT

CONTENT ENRICHMENT

+ ASR Transcripts

+ Extracted Keywords

Named Entity Recognition

Goldman Sachs Rises as Investors Bet on Comeback

Goldman Sachs Group Inc. (GS) rose 5.5 percent in N York trading as investo loss an revenue in underwriting and takeovers.

http://en.wikipedia.org/wiki/Goldman_Sachs

Industry	**Financial services**
LocationCity	**New York City**
date_founded	**1869**

Named Entity Recognition

Nach den anhaltenden Gewaltattacken in der Berliner S-Bahn fordert der Senat von der Deutsche Bahn Lösungen zur Vorbeugung und Bekämpfung.

Als Eigentümerin sei die Bahn in der Pflicht, ebenso wie die BVG Überwachungskameras zu installieren, sagte eine Sprecherin von Verkehrsenator Michael Müller SPD am Sonntag dem rbb.

Auch Innensenator Frank Henkel CDU sprach in der Berliner Morgenpost von einem längst überfälligen Schritt.

Der Betriebsrat der S-Bahn lehnt bisher eine Videoüberwachung ab. Die Arbeitnehmervertreter befürchten, dass die Kameras zur Überwachung der Mitarbeiter eingesetzt werden könnten.

Hintergrund der Debatte sind mehrere brutale Übergriffe in den vergangenen Tagen. Für Schlagzeilen sorgte vor allem der Fall eines geistig behinderten Fußballfans, der von Unbekannten beinahe erdrosselt wurde.

Der Überfall ereignete sich am S-Bahnhof Olympiastadion. Unbekannte hatten den am Down-Syndrom erkrankten Mann erst geschlagen und anschließend dessen Fanschal eng um den Hals geschnürt und das Ende des Schals an einem Geländer festgeknotet. Dann ließen sie ihn einfach auf dem Bahnsteig sitzen. Dabei wäre der 31-Jährige fast erstickt, da er sich aufgrund seiner Erkrankung nicht selbst befreien konnte. Als Polizisten den Mann später bemerkten, sei er schon stark benommen gewesen. Er wurde in ein Krankenhaus gebracht. Die Polizei ermittelt wegen versuchter Tötung.

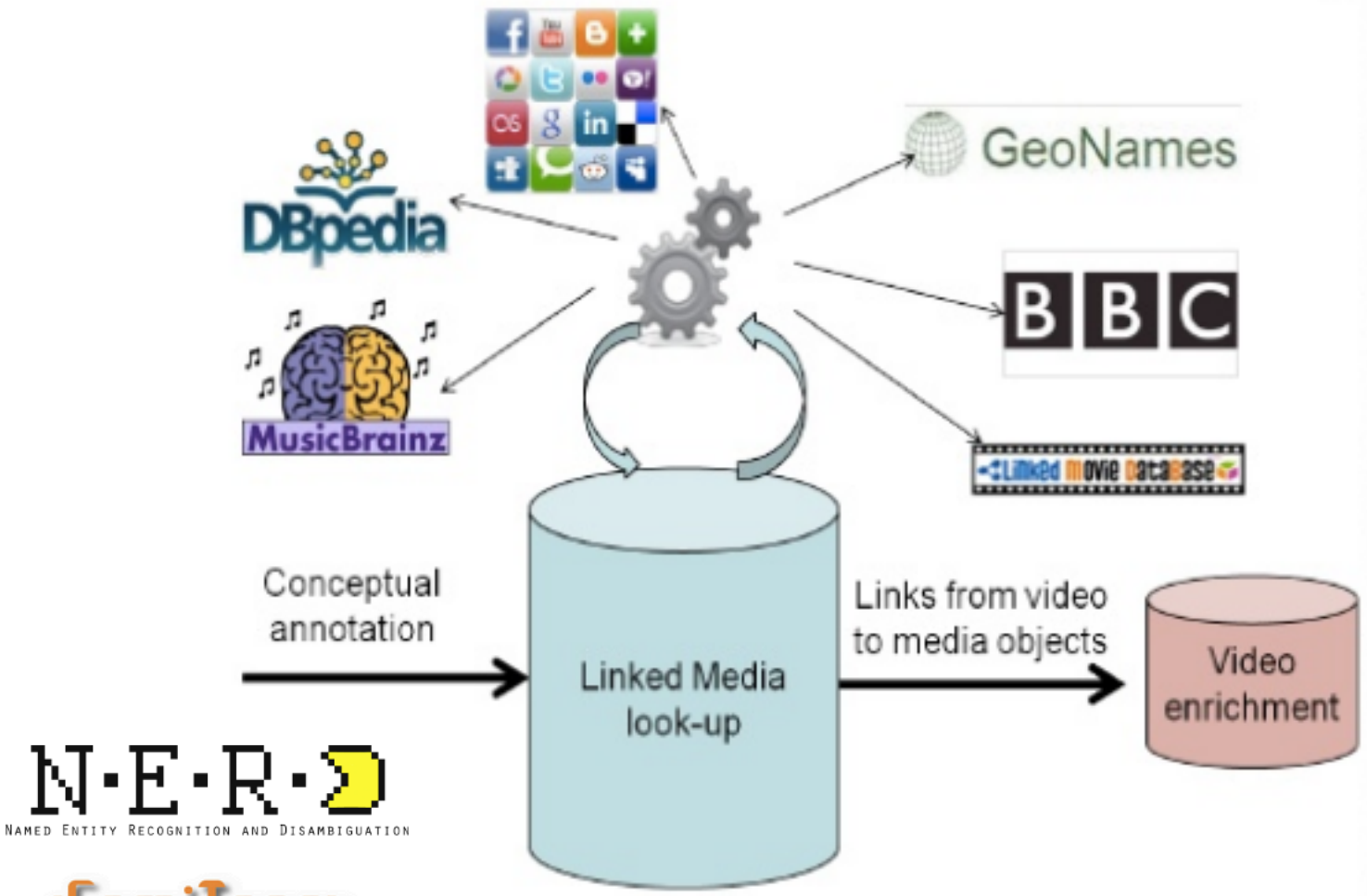
de Recognize

Identified Named Entities

- Berliner (I-MISC)
- Deutsche (I-MISC)
- BVG (I-ORG)
- Michael Müller (I-PER)
- SPD (I-ORG)
- Frank Henkel (I-PER)
- CDU (I-ORG)
- Berliner Morgenpost (I-ORG)
- Olympiastadion (I-LOC)
- Down-Syndrom (I-MISC)

Disambiguated Named Entities

- Berliner ... http://de.wikipedia.org/wiki/Berlin
- BVG ... http://de.wikipedia.org/wiki/Berliner_Verkehrsbetriebe
- Michael Müller ... http://de.wikipedia.org/wiki/Michael_M%C3%BCller_(Berlin)
- SPD ... http://de.wikipedia.org/wiki/Sozialdemokratische_Partei_Deutschlands
- Frank Henkel ... http://de.wikipedia.org/wiki/Frank_Henkel
- CDU ... http://de.wikipedia.org/wiki/Christlich_Demokratische_Union_Deutschlands
- Berliner Morgenpost ... http://de.wikipedia.org/wiki/Berliner_Morgenpost
- Olympiastadion ... http://de.wikipedia.org/wiki/Olympiastadion_Berlin
- Down-Syndrom ... http://de.wikipedia.org/wiki/Down-Syndrom



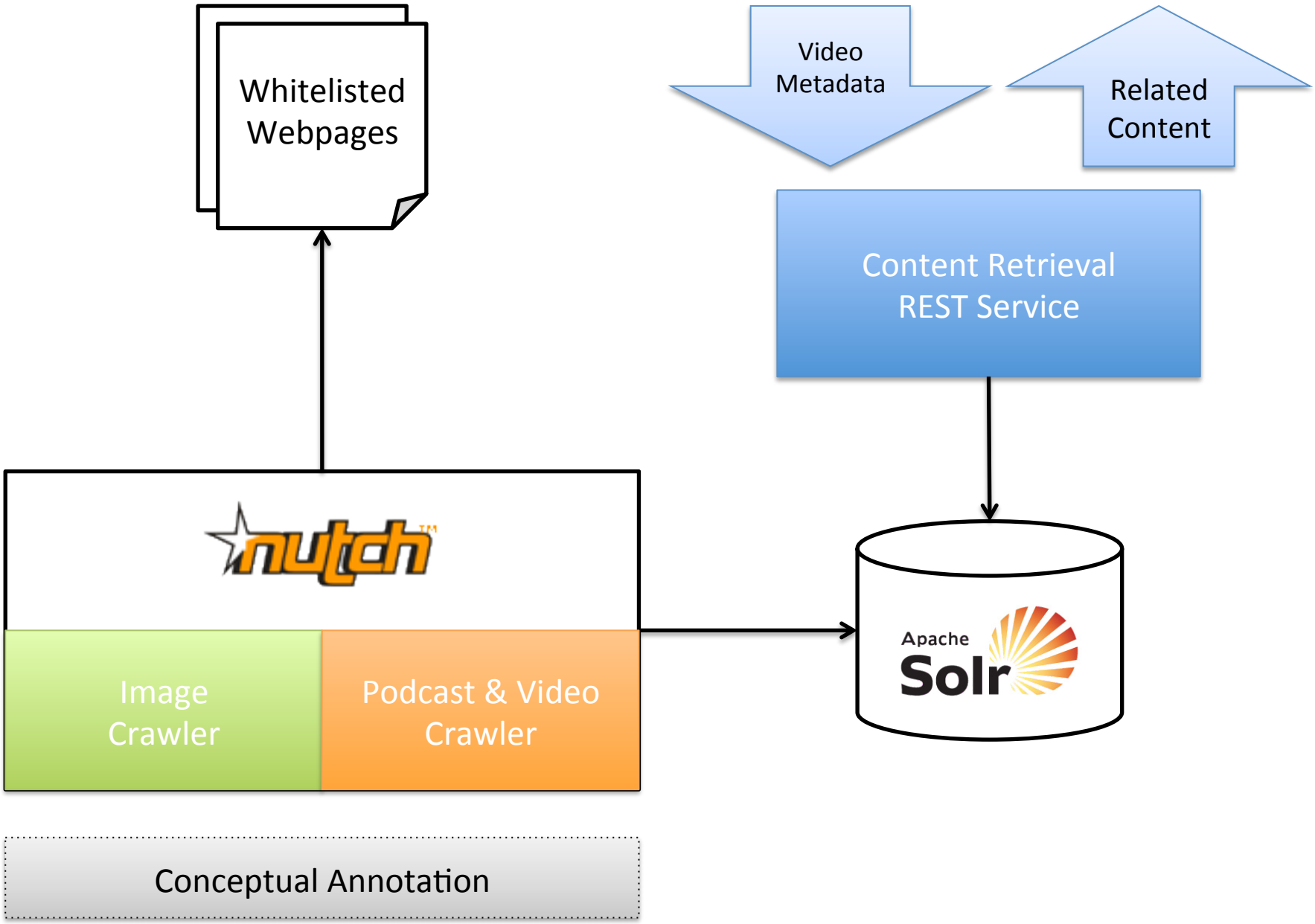
N·E·R·D

Named Entity Recognition and Disambiguation

«SemiTags»

Targeted Hypernym Discovery

Linking Related Content from the Web


Whitelisted Webpages
Video Metadata
Related Content
Content Retrieval REST Service
nutch
Image Crawler
Podcast & Video Crawler
Apache Solr
Conceptual Annotation

Related Content Retrieval

- Keywords
- Names of Named Entities – later identifiers of named entities
- Concepts identified in WP1

Future Work

Keyword Extraction

- Implement web service
- Support multiple users in the system (personalized keywords)
- Building the golden standard using German articles annotated on Amazon Machanical Turk
- Possible clustering of videos based on keywords

Named Entity Recognition

- Improve performance (Solr indexes instead of RDBMS)
- Enable fully featured web service (including DBpedia and Wikipedia identifiers)
- Enable English disambiguation
- Building the golden standard using German articles annotated on Amazon Machanical Turk
- Retrieve additional entity properties from DBpedia

Related Content Retrieval

- Implement multimedia content crawling plugins
- Implement querying web service
  - Need to be discussed with partners, what will be the exact input and in which format (we need available video metadata)
- Add concept detection (NER) also in the crawled web content
- Linking based on relations between contained entities

Conclusion

March

- Confidence in KW extraction
- Web service for KW extraction
- SemiTags applied on Frauenhofer Data
- WebService for SemiTags
- Podcats crawler
- Video crawler

April

- Multiple users for KW extraction
- Keyword Extraction evaluation on a standard dataset
- KW extraction golden standard creation
- SemiTags performance improvements
- Related texts attached to podcasts, images and videos

May

- Golden standard for NER
- NER evaluation
- NER applied on web resources for linking