



# Television Linked To The Web

---

## WP2

## Linking hypervideos to Web content

---

Raphael Troncy  
EURECOM

First Year Review Meeting – 6 February 2013

[www.linkedtv.eu](http://www.linkedtv.eu)

- Develop a **LinkedTV ontology** for representing video metadata
  - Legacy metadata, automatic analysis results, provenance information
  - Annotations are represented in RDF and stored in the **LinkedTV platform**
- Enrich a seed video (TV program) with relevant (structured) data and multimedia content from a curated list of sources or social media
- WP2 = a URI farm



http://www

# From a broadcast program ...



LINKEDTV

www.linkedtv.eu

MF URI

<http://data.linkedtv.eu/media/bdb0c0#t=2515&xywh=360,320,150,131>



MF URI

<http://data.linkedtv.eu/media/bdb0c0#t=2636&xywh=321,295,157,157>



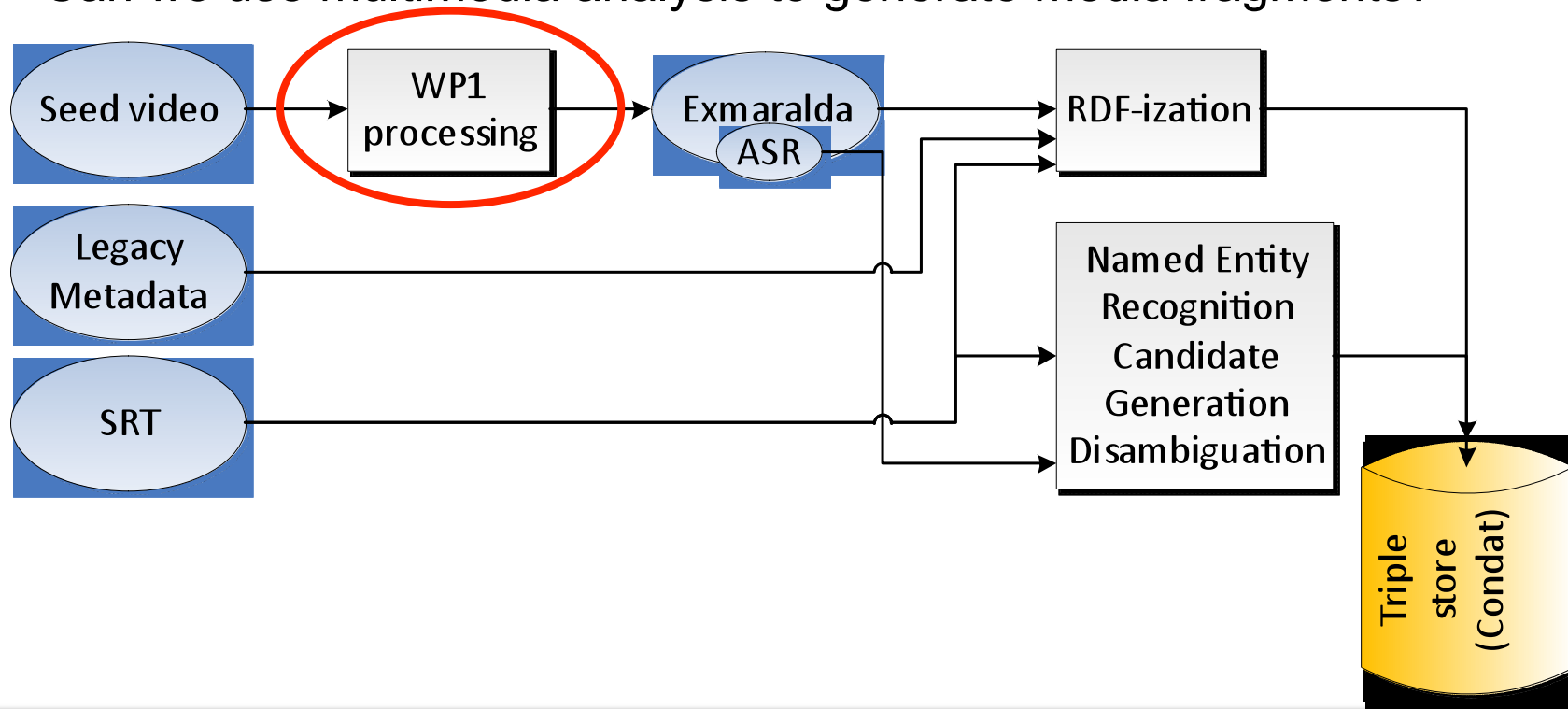
Program broadcasted by RBB on August 9<sup>th</sup> 2012, featuring the actor Klaus Maria Brandauer reading a book to an audience.

- Broadcasted (legacy) Metadata
- Subtitle
- WP1 Results:
  - Shot detection
  - Concept identification
  - Face Recognition

content

# Scientific and technological challenges (1)

- Which ontologies to use to represent broadcast information, subtitles or automatic multimedia analysis results while keeping track of provenance?
- How much results of multimedia analysis processes should be RDF-ized?
- Can we use multimedia analysis to generate media fragments?



# ... to enriched content with factual data



## Klaus Maria Brandauer

[Actor](#) | [Director](#)

Gave his debut as an actor in 1962. Member of the 'Wiener Burgtheater' in Vienna since 1972. Played at the 'Salzburger Festspiele' since 1975. Directed plays since 1977. Debuted as a movie actor in 1972. [See full bio »](#)

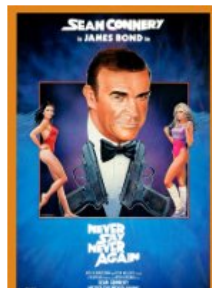
**Born:** [Klaus Georg Steng](#)  
[June 22, 1943](#) in [Bad Aussee, Styria, Austria](#)



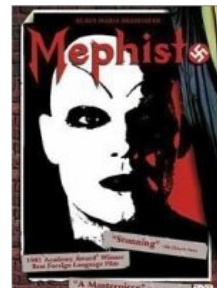
IMDb



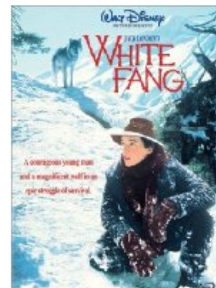
[Out of Africa - Souvenirs d'Afrique \(1985\)](#)



[Jamais plus jamais \(1983\)](#)



[Mephisto \(1981\)](#)



[Croc-Blanc \(1991\)](#)



[http://dbpedia.org/resource/Klaus\\_Maria\\_Brandauer](http://dbpedia.org/resource/Klaus_Maria_Brandauer)

## Klaus Maria Brandauer



[en.wikipedia.org](#)

Klaus Maria Brandauer is an Austrian actor, film director, and professor at the Max Reinhardt Seminar in Vienna. [Wikipedia](#)

**Born:** [June 22, 1943](#) (age 69), [Bad Aussee](#)

**Height:** 1.75 m

**Spouse:** [Natalie Krenn](#) (m. 2007), [Karin Brandauer](#) (m. 1963–1992)

**Children:** [Christian Brandauer](#)

**Parents:** [Maria Brandauer](#), [Georg Steng](#)

## Movies



[Out of Africa](#)  
1985



[Never Say Never Again](#)  
1983



[Mephisto](#)  
1981



[Tetro](#)  
2009



[Oberst Redl](#)  
1985

## People also search for



[Szabó](#)



[Carrera](#)



[Denys Finch Hatton](#)



[Sydney Pollack](#)



[Irvin Kershner](#)

## ... and enriched content with media

MF URI

<http://data.linkedtv.eu/media/adbrf0#t=2237>

### SHMF: Das Eröffnungskonzert 2012

Sendedatum: 03.10.2012 09:10 Uhr

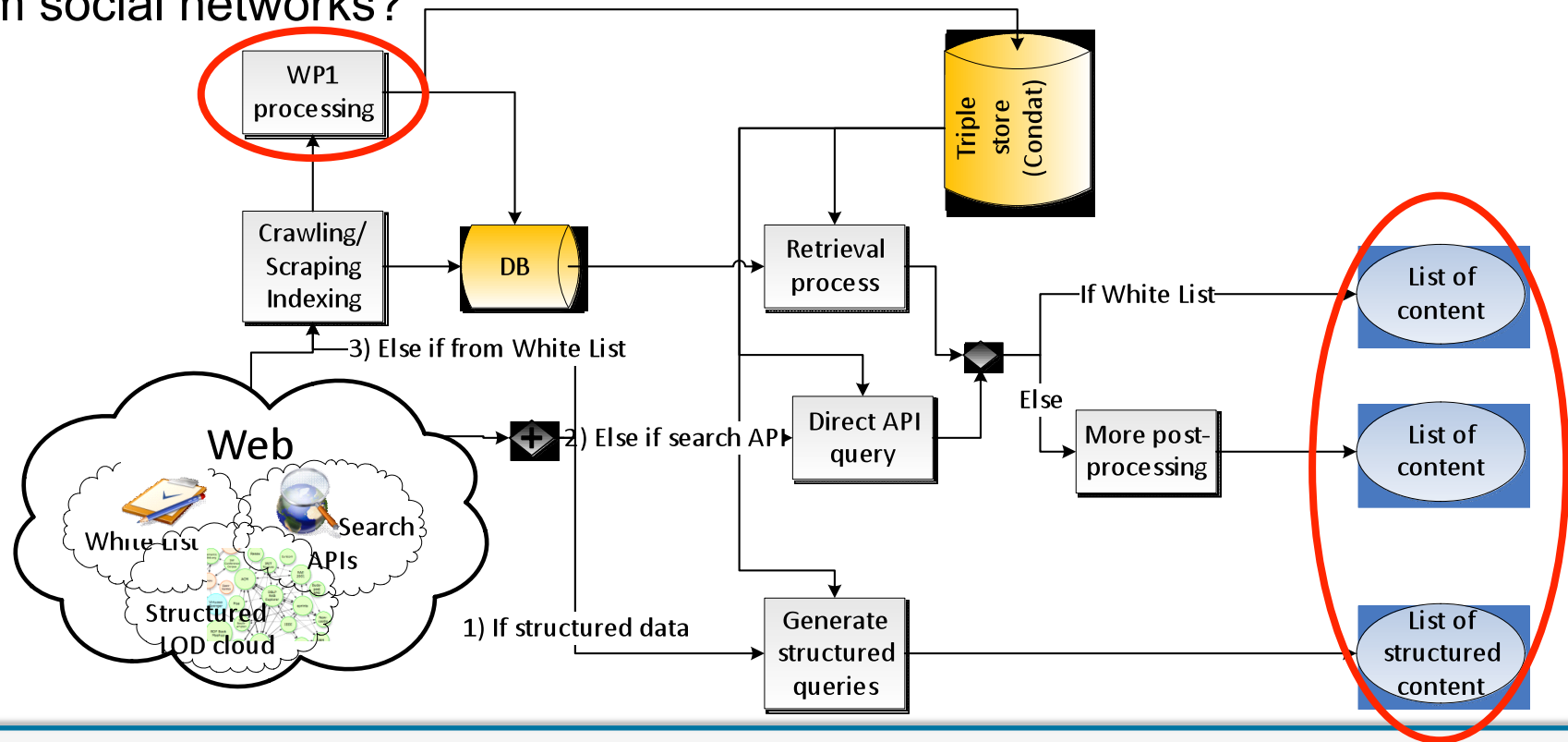


Klaus Maria Brandauer und Thomas Hengelbrock brachten zur Eröffnung des Schleswig-Holstein Musik Festivals Ibsens "Peer Gynt" mit der Schauspielmusik von Edvard Grieg auf die Bühne.

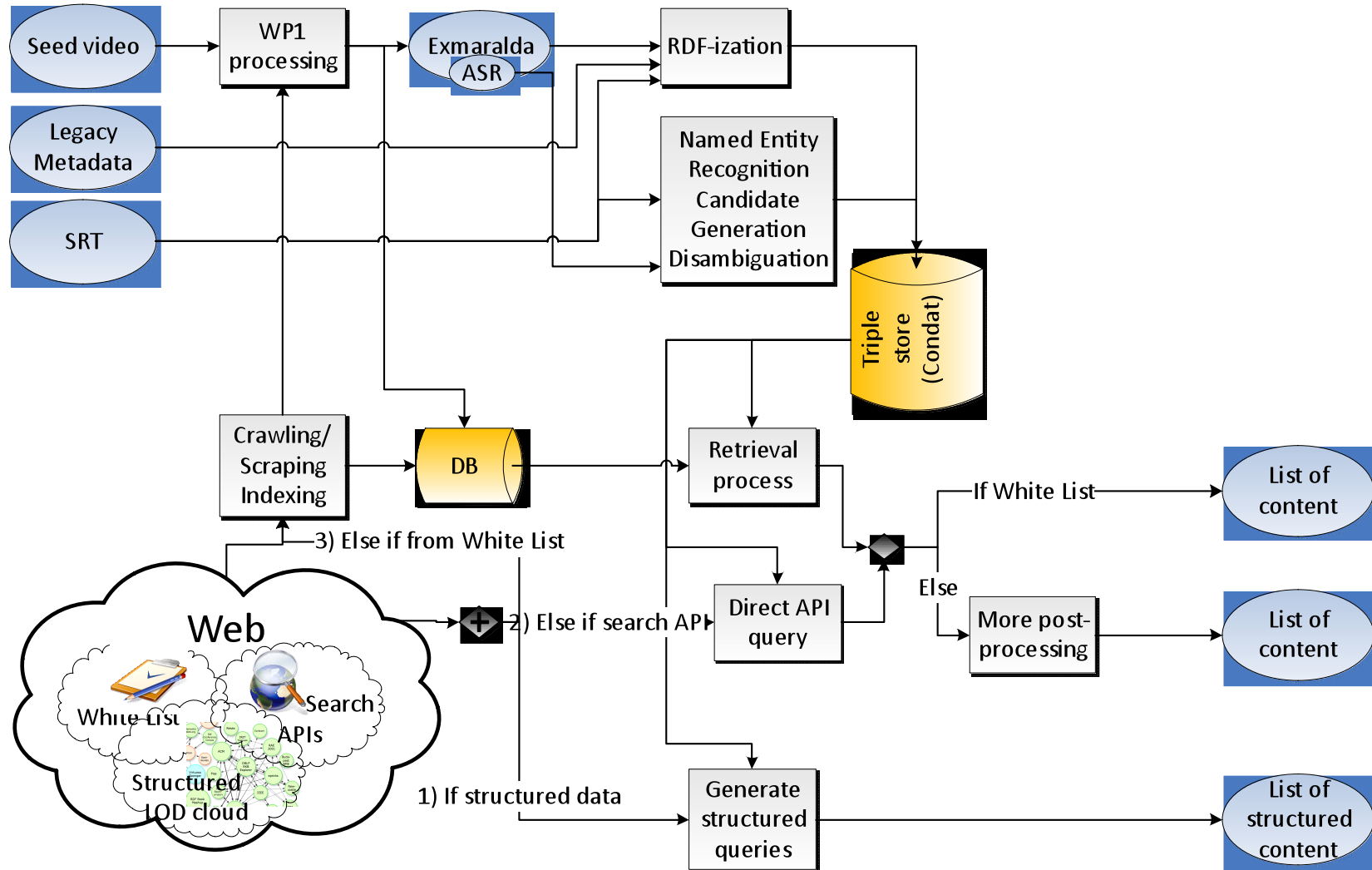
<http://www.ndr.de/fernsehen/sendungen/media/gynt101.html>

# Scientific and technological challenges (2)

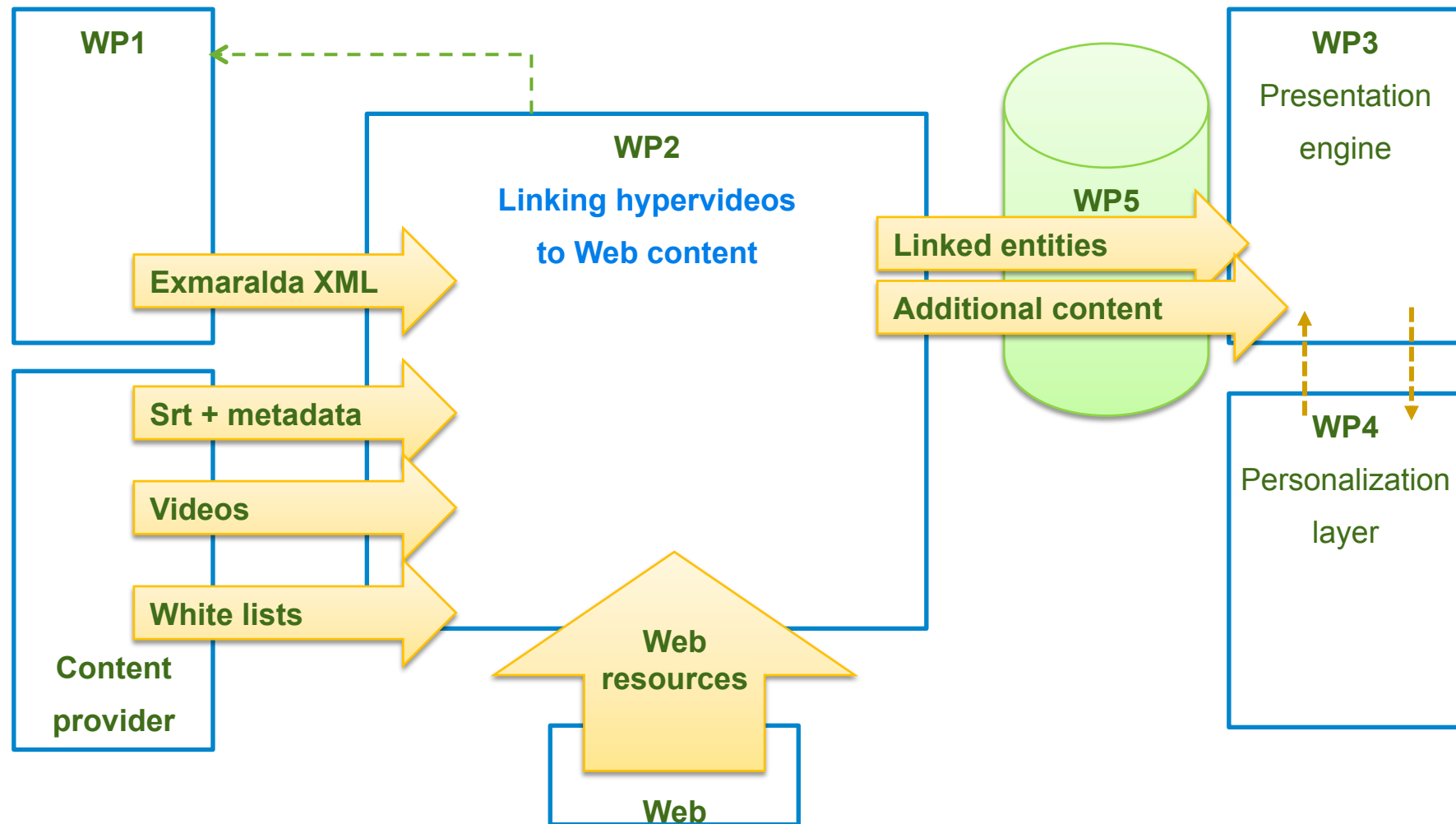
- How to crawl, index and analyze efficiently a curated list of web sites?
- How to enrich a seed video program with other images and videos available on the web and in broadcaster archives?
- How to enrich a seed video program with fresh media and sentiments from social networks?



# WP2 - Workflow

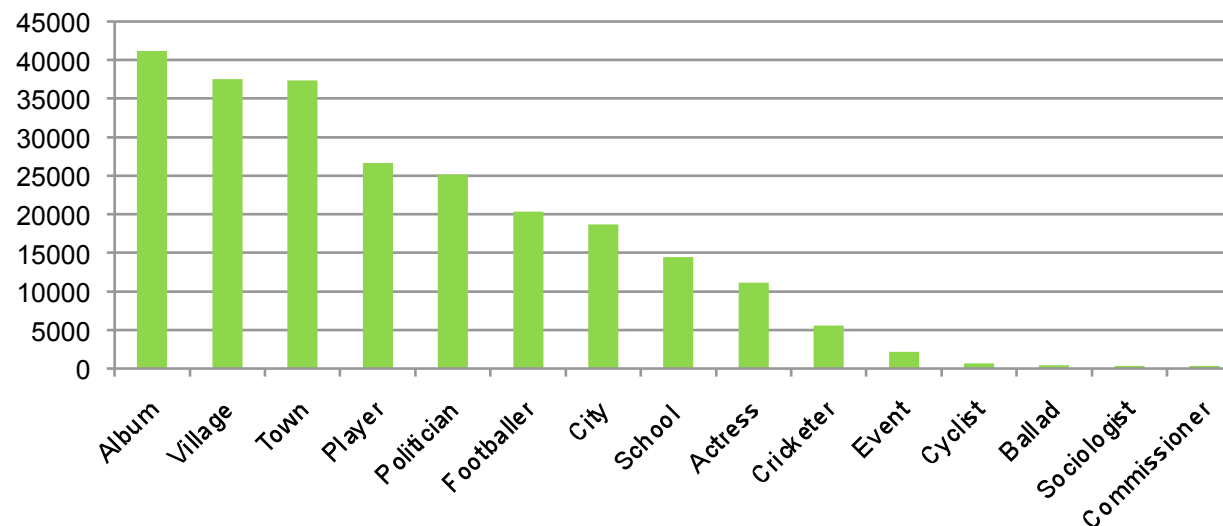


# WP2 – Dependencies with other WPs



# Enrichment for Personalization (WP2/WP4)

- Additional structured data and content is provided with a confidence score and/or soft classification used within WP4 for personalization
  - The BOA tool will provide soft entity classification to multiple entity types
- More fine-grained types is better for personalization
  - THD complements existing NER tools by providing additional fine-grained types (e.g. Angela Merkel is a “Chancellor”, etc.)



**A selection from 20.000 entity types assigned by THD, along with Wikipedia frequency**

1. Shot and scene segmentation used for generating Media Fragments
  - W3C Recommendation for temporal and spatial fragments
2. Re-use as much as possible common vocabularies in the semantic web
  - schema.org, Open Annotations, PROV-O, Ontology for Media Resources
  - FOAF, Dublin Core, NERD, LSCOM, DBpedia Ontology
3. Name Entity Recognition
  - Statistical based approaches
  - Knowledge Based Approaches (Wikipedia/DBpedia)
  - Web-based APIs
4. Enrichment based on textual and visual analysis
  - Structured data: LOD cloud accessible through structured queries (SPARQL)
  - Search API: REST based query
  - Online content repositories (curated list):
    - Crawling, Wrapping, Indexing and Searching (Lucene/Solr)
    - Web-based and content-based mining approaches

# LinkedTV model (1)

## ■ Broadcast Metadata:

- Title
- Description
- Format
- Series, Episodes, Brands
- Categories
- Subtitles

## ■ External Datasets:

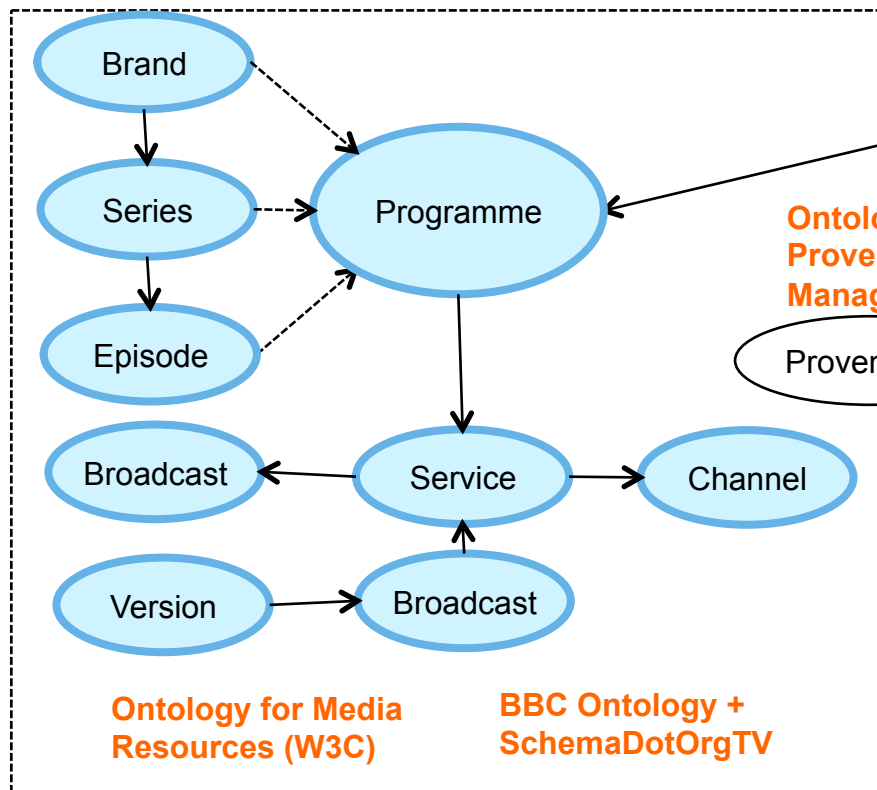
- LOD Cloud:
  - DBpedia, LinkedMDB, Geonames, BBC, etc.
- White List of data sources for every Scenario.
- Fresh information from Social Networks:
  - Facebook, MySpace, Twitter, ...

## ■ Automatic Analysis Results:

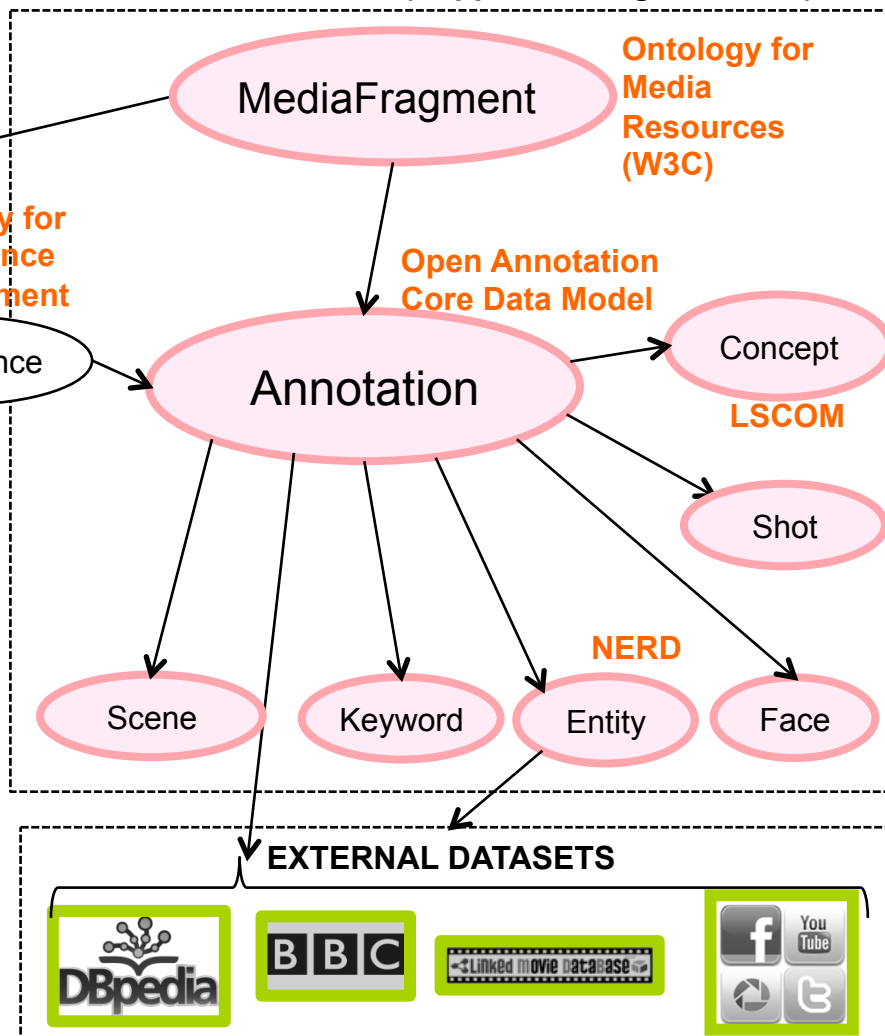
- Exmaralda XML files (WP1):
  - Spatiotemporal Segments
  - Scene segmentation
  - Shot segmentation
  - ASR
  - Concepts
  - Keywords
- Entities (NERD)
- Links between media and concepts/entities

# LinkedTV model (1)

## BROADCAST DATA



## ANALYSIS RESULTS (Support for segmentation)



**LinkedTV Ontology Datamodel  
for Representing Information  
about Television Content**

# LinkedTV metadata conversion (2)

[Home](#) [Get started](#) [About](#) [Contact](#)

LINKEDTV

# TV Metadata Generator

Automatic conversion of television content metadata into RDF.

<http://linkedtv.eurecom.fr/metadata/>



Instructions >

Select Video Resource >

Choose Metadata Files >

RDF Generation >

Upload Results to Triplestore >

Queries over the RDF Data >

## 1. Introduction

Here we present the tool "TV Metadata Generator", developed by [Eurecom](#) under the scope of the European Project [LinkedTV FP7](#). In a nutshell, this application takes a video and its corresponding metadata files, available in different formats, and produces a RDF representation of the whole information. The knowledge graph obtained as result is suitable to be interlinked with other data in the Web, and allow the execution of complex queries that can bring the viewers a new way of enjoy Television.

Demo available

# Named Entity Recognition Platform (3)

N-E-R-3

- <sup>1</sup> <http://nerd.eurecom.fr/ontology>
- <sup>2</sup> <http://nerd.eurecom.fr/api/application.wadl>
- <sup>3</sup> <http://nerd.eurecom.fr>

ontology<sup>1</sup>   REST API<sup>2</sup>  
UI<sup>3</sup>

## Web APIs



# Named Entity Recognition Platform (3)

- SemiTags - Named Entity Classification (German and Dutch)
  - Two independent algorithms for recognition evaluated
  - Currently we use the recognition based on Stanford Parser which outperformed other solutions
  - Disambiguation based on co-occurrences of entities
  - Web Service for integration SemiTags into NERD
- Targeted Hypernym Discovery Entity Classification (English, German, Dutch)
  - Provides closer types than most industry-grade NER systems
  - RDF output (NIF format)



# Demonstration (1) – NERD User Interface



You are logged in.

[Logout!](#)

[home](#)

[documentation](#)

[ontology](#)

[publication](#)

URI article

Analyze web resource

Raw text

Demo  
available

Analyze plain text



# Demonstration (2) – THD

## Targeted Hypernym Discovery v0.2<sup>Beta</sup>

Home Application Web API Contact

### Advanced Entity and Hypernym Extraction tool

Erst vor rund einer Woche hatte Airline-Chef Hartmut Mehdorn seinen Hut genommen. Unter seiner Leitung war das Streckennetz zusammengestrichen und der Verkauf von überflüssigen Flugzeugen angestoßen worden.

Num. of hypernyms:

3

Request timeout (in seconds):

60

☐ English ☒ German ☐ Dutch

☐ Extract only Named Entities (or also common nouns)

Run!

Hypernyms:

#1: Industriemanager for entity disambiguated as Hartmut\_Mehdorn

#2: Unternehmer for entity disambiguated as Joachim\_Hunold

#3: Verkehrsunternehmen for entity disambiguated as Deutsche\_Bahn

#### Results

Erst vor rund einer Woche hatte Airline-Chef Hartmut Mehdorn seinen Hut genommen. Unter seiner Leitung war das Streckennetz zusammengestrichen und der Verkauf von überflüssigen Flugzeugen angestoßen worden.

Results processed in 0.771 seconds.

Demo available at <http://ner.vse.cz/thd/application/>

# Demonstration (2) – SemiTags

← → ↻ ner.vse.cz/SemiTags/app/index/

## SemiTags

### Named Entity Recognition

Berlins Bildungssenatorin Sandra Scheeres sieht Berlin gut für das neue Schuljahr gerüstet. Es seien genügend Lehrer an Bord, sagte Scheeres dem rbb. Mur an einzelnen Schulen könne keine hunderprozentige Abdeckung mit Lehrern ermöglicht werden. Diese Schulen habe sie aber genau im Auge, so Scheeres. Eine starke Abwanderungstendenz junger Lehrer in andere Bundesländer könne sie nicht bestätigen. Berlin sei attraktiv für Lehrer aus dem übrigen Deutschland, auch wenn hier nicht mehr verbeamtet werden könne, so die SPD-Politikerin.

Scheeres widersprach der Kritik, wonach Grundschulen mit hoher Migrantenquote mit weniger Lehrern auskommen müssten, weil sie weniger Förderstunden bekommen als bisher. Die Grundschulen würden entlastet, so Scheeres, denn die Konrektoren müssen weniger unterrichten als bisher. Außerdem werden in Berlin 60 Lehrer zusätzlich an Problemschulen eingesetzt. Im kommenden Schuljahr will Scheeres ohnedies besonders die Grundschulen entlasten. Die Konrektoren sollen sich durch weniger Unterrichtsstunden besser den administrativen Aufgaben widmen zu können.

Sorgen bereiten Scheeres nach wie vor die vielen Dauerkranken. An den 704 Öffentlichen Schulen Berlins sind 1448 Lehrer ständig im Krankenstand. Diese Zahl ist zu hoch, sagte Scheeres.

Brandenburg Der Trend geht in Richtung Verjüngung  
Auch an den mehr als 700 Schulen in Brandenburg hat nach gut sechs Wochen Ferien am Montag das neue Schuljahr begonnen. Die Erstklässer wurden bereits am 4. August eingeschult. Auf den Schulstart sind Schulen und Lehrer gut vorbereitet, so Brandenburgs Bildungsministerin Martina Münch SPD. Mit 450 neuen Lehrern stellt Brandenburg so viel ein wie noch nie, so Münch.

Rund 100 Lehrerinnen und Lehrer mit sonderpädagogischer Ausbildung sollen beispielsweise das Modellprojekt Inklusive Schule zu einem Erfolg machen. Gemeinsam lernen Schülerinnen und Schüler mit und ohne Förderbedarf hier von Montag an in insgesamt 85 Grundschulen so an der Anne-Frank-Grundschule Teltow, der Mosaik-Grundschule Döbern und der Freien Landschule Wittstock. Eine 1.Klasse in dem Modellversuch soll nicht mehr als 23 Schüler haben.

de Recognize

#### Identified Named Entities

- Berlins (I-LOC)
- Sandra Scheeres (I-PER)
- Berlin (I-LOC)
- Berlin (I-LOC)
- Deutschland (I-LOC)
- SPD-Politikerin (I-MISC)
- Berlin (I-LOC)
- Berlins (I-LOC)
- Scheeres (I-PER)
- Brandenburg (I-LOC)
- Brandenburg (I-LOC)
- Brandenburgs (I-LOC)
- Martina Münch (I-PER)
- SPD (I-ORG)
- Brandenburg (I-LOC)
- Anne-Frank-Grundschule Teltow (I-ORG)
- Freien Landschule Wittstock (I-ORG)

#### Disambiguated Named Entities

- Berlins ... <http://de.wikipedia.org/wiki/Berlin>
- Sandra Scheeres ... [http://de.wikipedia.org/wiki/Sandra\\_Scheeres](http://de.wikipedia.org/wiki/Sandra_Scheeres)
- Berlin ... <http://de.wikipedia.org/wiki/Berlin>
- Berlin ... <http://de.wikipedia.org/wiki/Berlin>
- Deutschland ... <http://de.wikipedia.org/wiki/Deutschland>
- Berlin ... <http://de.wikipedia.org/wiki/Berlin>
- Berlins ... <http://de.wikipedia.org/wiki/Berlin>
- Brandenburg ... <http://de.wikipedia.org/wiki/Brandenburg>
- Brandenburg ... <http://de.wikipedia.org/wiki/Brandenburg>
- Brandenburgs ... <http://de.wikipedia.org/wiki/Brandenburg>
- Martina Münch ... [http://de.wikipedia.org/wiki/Martina\\_M%C3%BCnch](http://de.wikipedia.org/wiki/Martina_M%C3%BCnch)
- SPD ... [http://de.wikipedia.org/wiki/Sozialdemokratische\\_Partei\\_Deutschlands](http://de.wikipedia.org/wiki/Sozialdemokratische_Partei_Deutschlands)
- Brandenburg ... <http://de.wikipedia.org/wiki/Brandenburg>

Demo available at <http://ner.vse.cz/SemiTags>

Media Fragment Enricher

Home Documentation

## Schkopi Show N°3 - Part 1 (With CC)



### Subtitle

00:02:54 to 00:02:57  
It's true that it's an instrument somewhat underestimated  
00:02:57 to 00:02:59  
and which isn't as easy to handle as it seems.  
00:03:00 to 00:03:03  
Most of the time the audience respects more guitarists  
00:03:03 to 00:03:08  
such as Carlos Santana, Jimi Hendrix, Keith Richards and Prince of course.  
00:03:08 to 00:03:11  
But who were your inspirations ?  
00:03:12 to 00:03:14  
Who were the people who influenced you to keep playing ?  
00:03:14 to 00:03:19  
I was influenced by some of the most renowned harmonicists,  
00:03:20 to 00:03:22  
such as Stevie Wonder of course,

### NERD

#### Thing (13 entities)

Olimpia  
diatonic harmonica  
instrument  
chromatic and  
rock music  
Tonight Show  
high school  
jam session  
Canal  
country music

#### Jimi Hendrix

##### label

Jimi Hendrix

##### abstract

This article is about the guitarist. For the band, see The Jimi Hendrix Experience.  
"James Hendrix" redirects here. For the Ame...

##### depiction



Demo  
available

1. Giuseppe Rizzo, Thomas Steiner, Raphaël Troncy, Ruben Verborgh, Josè Luis Redondo Garcia and Rik Van de Walle. **What Fresh Media Are You Looking For? Extracting Media Items from Multiple Social Networks.** In *Proc. International Workshop on Socially-Aware Multimedia (SAM'12)*, October 29-November 2, 2012, Nara, Japan.
2. Milan Dojchinovski, Tomáš Kliegr. **Recognizing, Classifying and Linking Entities with Wikipedia and DBpedia.** In *Proc. 7th Workshop on Intelligent and Knowledge Oriented Technologies (WIKT 2012)*, November 2012, Bratislava.
3. Yunjia Li, Giuseppe Rizzo, Raphaël Troncy, Mike Wald and Gary Wills. **Creating Enriched YouTube Media Fragments With NERD Using Timed-Text.** In *Proc. 11th International Semantic Web Conference (ISWC'12), Demo Session*, November 11-15, 2012, Boston, USA.
4. Sven Buschbeck, Anthony Jameson, Raphaël Troncy, Houda Khrouf, Osma Suominen and Adrian Spirescu. **A Demonstrator for Parallel Faceted Browsing.** In *Proc. Intelligent Exploration of Semantic Data Workshop (IESD'12)*, October 8-12, 2012, Galway, Ireland. [Winner of the IESD challenge](#)
5. Radek Škrabal, Milan Šimůnek, Stanislav Vojíř, Andrej Hazucha, Tomáš Marek, David Chudán, Tomáš Kliegr. **Association Rule Mining Following the Web Search Paradigm.** In *Proc. of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML-PKDD 2012*, Bristol, UK, 24-28 September 2012
6. Giuseppe Rizzo, Raphaël Troncy, Sebastian Hellmann and Martin Bruemmer. **NERD meets NIF: Lifting NLP Extraction Results to the Linked Data Cloud.** In *Proc. 5th Workshop on Linked Data on the Web (LDOW'12)*, April 16, 2012, Lyon, France.



# Named Entity Recognition Platform (3)

## ■ ETAPE 2012 Benchmark



| genre         | train   | dev    | test   | sources                                               |
|---------------|---------|--------|--------|-------------------------------------------------------|
| TV news       | 7h 40m  | 1h 40m | 1h 40m | BFM Story, Top QUestions (LCP)                        |
| TV debates    | 10h 30m | 5h 10m | 5h 10m | Pile et Face, Ca vous regarde, Entre les lignes (LCP) |
| TV amusements | -       | 1h 05m | 1h 05m | La place du village (TV8)                             |

|             | SLR           | Precision     | Recall | F-measure     | %correct |
|-------------|---------------|---------------|--------|---------------|----------|
| alchemyapi  | 37.71%        | <b>47.95%</b> | 5.45%  | 9.68%         | 5.45%    |
| lupedia     | 39.49%        | 22.87%        | 1.56%  | 2.91%         | 1.56%    |
| opencalais  | <b>37.47%</b> | 41.69%        | 3.53%  | 6.49%         | 3.53%    |
| wikimeta    | 36.67%        | 19.40%        | 4.25%  | 6.95%         | 4.25%    |
| <b>NERD</b> | 86.85%        | 35.31%        | 17.69% | <b>23.44%</b> | 17.69%   |

# Named Entity Recognition Platform (3)

- NERD in ETAPE
  - The NERD combined strategy outperforms any single extractor
  - NERD performs as good on perfect transcripts as on ASR
    - Not sensible to the grammar due to the use of black boxes in contrast to all participants
- SemiTags - Named Entity Classification
  - Tested on RBB data
- Targeted Hypernym Discovery Entity Classification
  - Tested on English dataset biased towards uncommon named entities

# Demonstration (1) – NERD Dashboard

