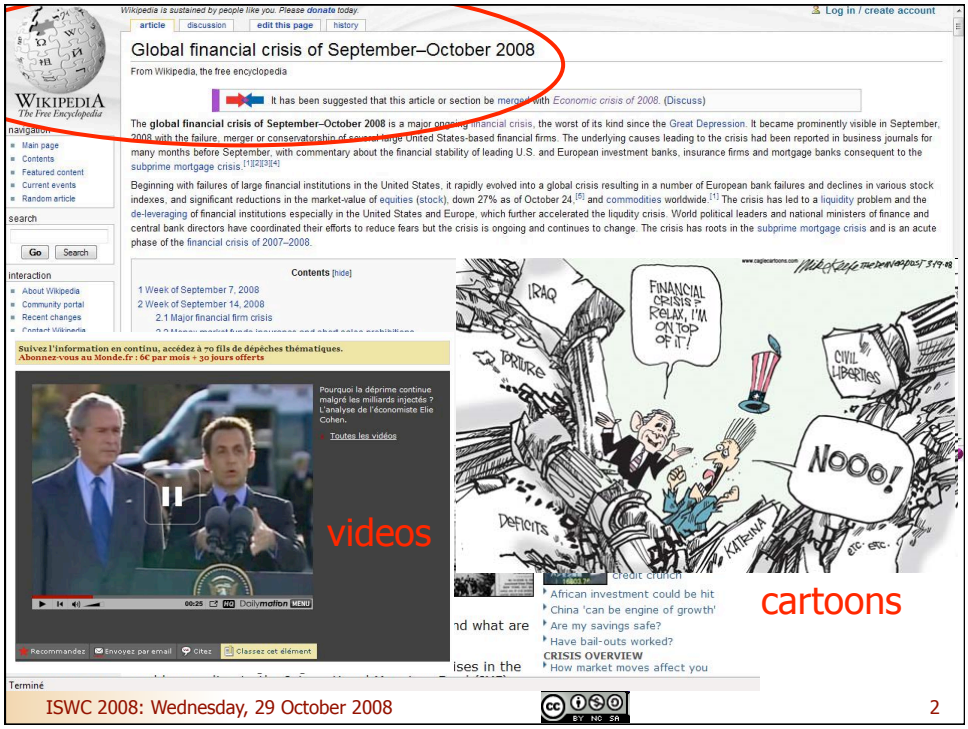
Bringing The IPTC News Architecture into the Semantic Web

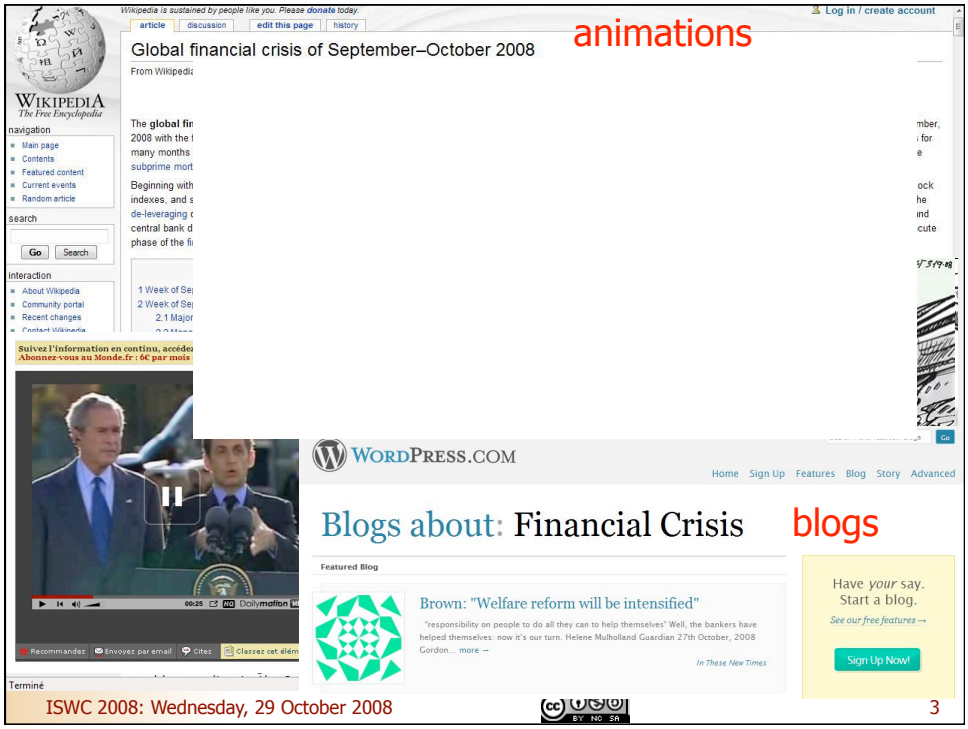
Raphaël Troncy, <Raphael.Troncy@cwi.nl>

CWI, Semantic Media Interfaces

cc BY NC SA

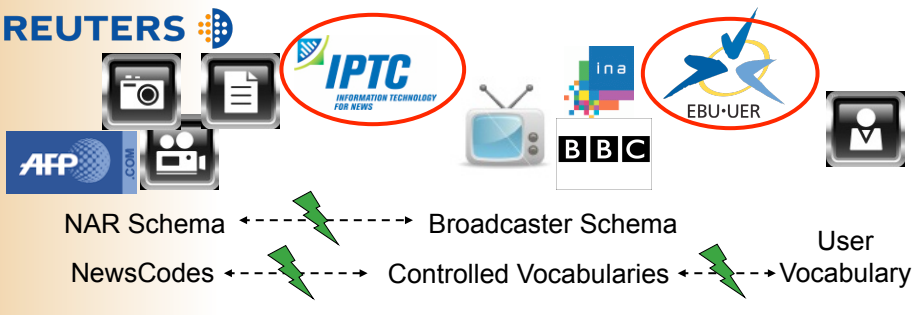


cc BY NC SA



News Workflow Interoperability

- No integration of media (stories, photo, animation, video)
- Little (or no) context in the news presentation
- Lack of interoperability in the current workflow

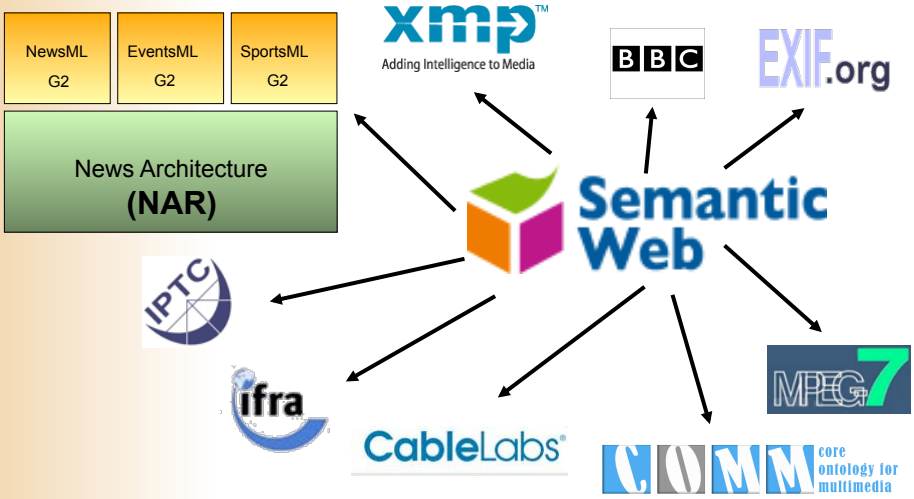


Metadata is Key

- (Ultimate) Goal:
  - Provide an environment for ***searching*** and ***browsing contextualized multimedia news*** information
- Required integration:
  - Data: various media, different forms, various sources
  - Metadata: schema integration, semantic models
- Influence and implications of UI:
  - How to **represent** semantic multimedia metadata to facilitate **presenting** information?
  - *in other words* ... What constraints do end-user interfaces put on the modeling of the metadata?

cc BY NC SA

News and Multimedia Formats



cc BY NC SA

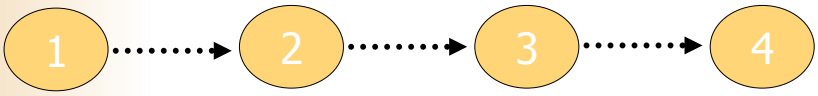
Porting Schemas and Thesauri to the Semantic Web

- Methodologies and tools for building ontologies:
  ... from scratch
- "SKOSification" of thesauri in the CH domain:
  - preparation, syntactic and semantic conversion, standardization

⇨ Lack of best practices for
modeling ontologies from UML diagrams,
integrating ontologies with various thesauri,
while taking the end-user interface into account

cc BY NC SA

Building a Semantic Web Infrastructure for News



Modeling the NAR ontology

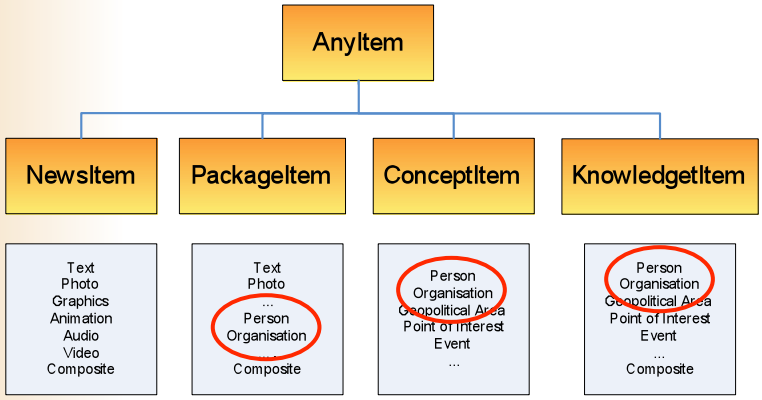
Linking with media ontologies

Building SKOS thesauri

Enriching the metadata

cc BY NC SA

Step 1: Modeling the NAR Ontology



⇨ focus on reuse of XML types leading to multiple repetition resulting in overly complex nested XML structures

cc BY NC SA

Step 1: Modeling the NAR Ontology

- Flattening the XML structure

```
<!--AFP NewsML2 text-photo profile-->
<!--Processed by Xafp1-4ToNewsML2 b11-->
- <newsMessage>
  + <header></header>
  - <itemSet>
    - <newsItem guid="urn:newsml:afp.com:20010101:.DV90996" schema="0.7" version="11" xml:lang="en">
        <catalogRef href="http://iptc.org/std-dev/NAR/1.0/specification/IPTC-TempCatalog-inc_3.xml"/>
      - <itemMeta>
          <contentClass code="ccls:photo"/>
          <provider literal="afp.com"/>
          <itemCreated>2006-07-09T21:20:00Z</itemCreated>
          <modified>2006-07-11T09:14:38Z</modified>
          <fileName>DV90996</fileName>
          <edNote>MOBILE SERVICES OUT</edNote>
        </itemMeta>
      + <contentMeta></contentMeta>
      + <contentSet></contentSet>
      </newsItem>
    </itemSet>
  </newsMessage>
```

NewsItem

↑

PhotoNewsItem

cc BY NC SA

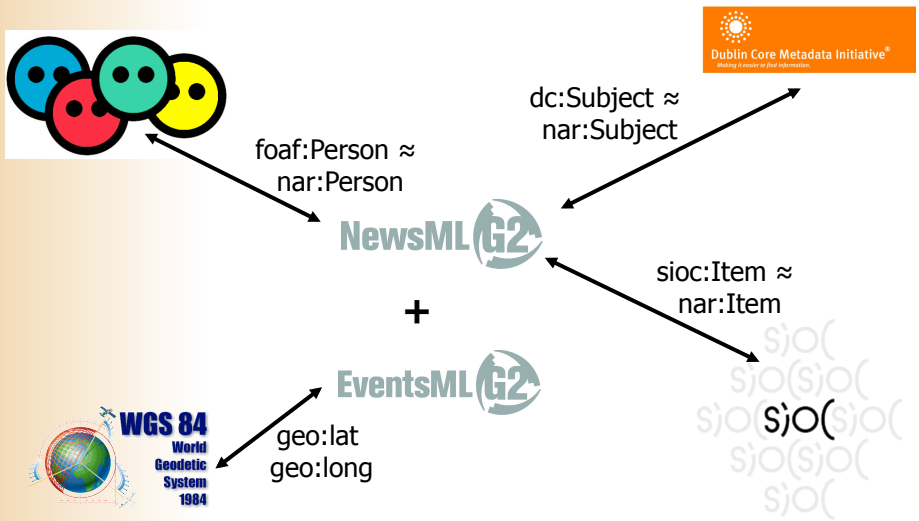
Step 1: Modeling the NAR Ontology

- Modeling unique identifiers
  - Use of dereferencable URIs for any resources (news items + vocabularies)
  - Future: Use of URIs for resource fragments http://www.youtube.com/watch?v=1bibCui3lFM#t=1m45s
- Modeling the provenance of the information
  - Reification
  - Named (and Networked) Graphs

```
{<> nar:subject cat:11002000}
        dc:creator team:md ;
        dc:modified ''2005-11-11T08:00:00Z''.
```

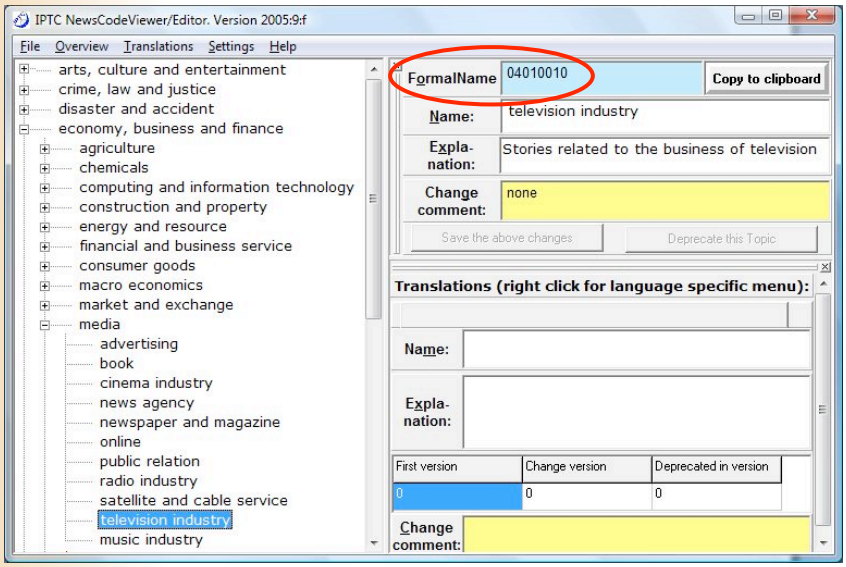

cc BY NC SA

Step 2: Linking with Media Ontologies



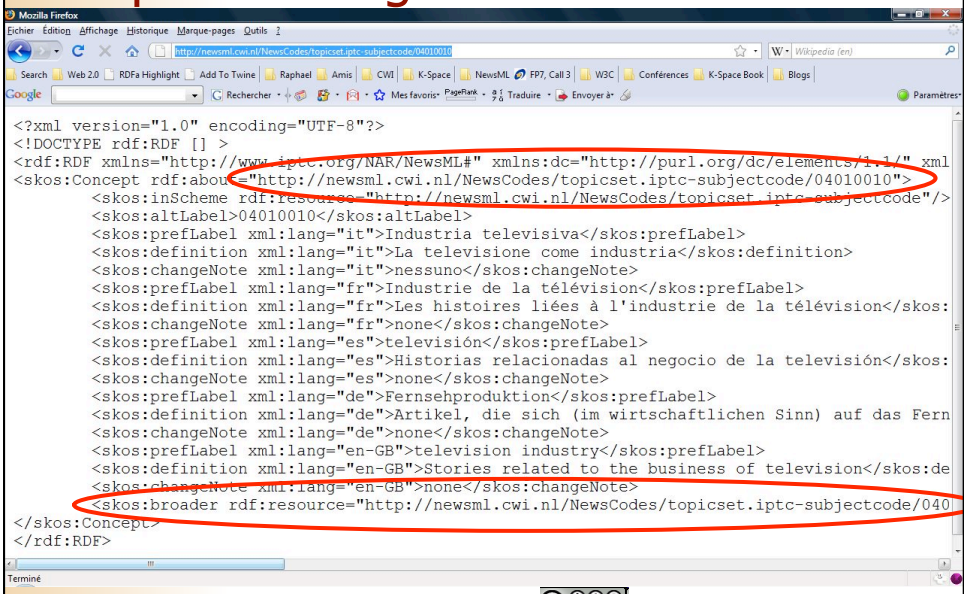
cc BY NC SA

Step 3: Getting SKOS Vocabularies



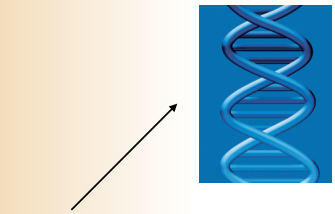
cc BY NC SA

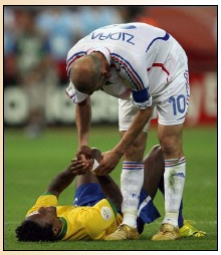
Step 3: Getting SKOS Vocabularies



cc BY NC SA

Step 4: Enriching the News Metadata

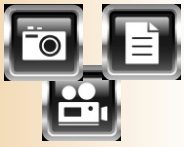




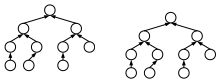
- Concepts/Entities that are subject of news
  - Thematic categories
  - People
  - Organizations
  - Geopolitical Areas
  - Points of Interest
  - Events
  - Products or artefacts

© IPTC – www.iptc.org

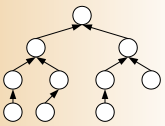
Step 4: Enriching the News Metadata



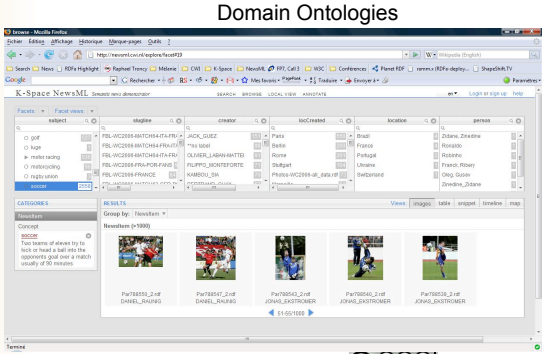
Named Entity Recognition

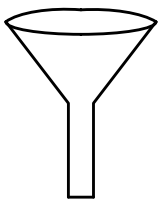


Domain Ontologies

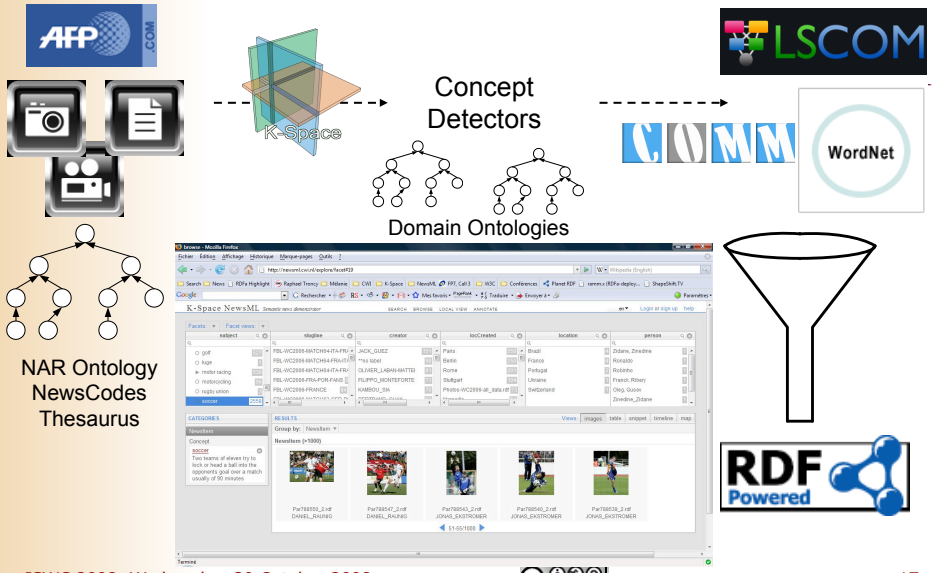


NAR Ontology
NewsCodes
Thesaurus



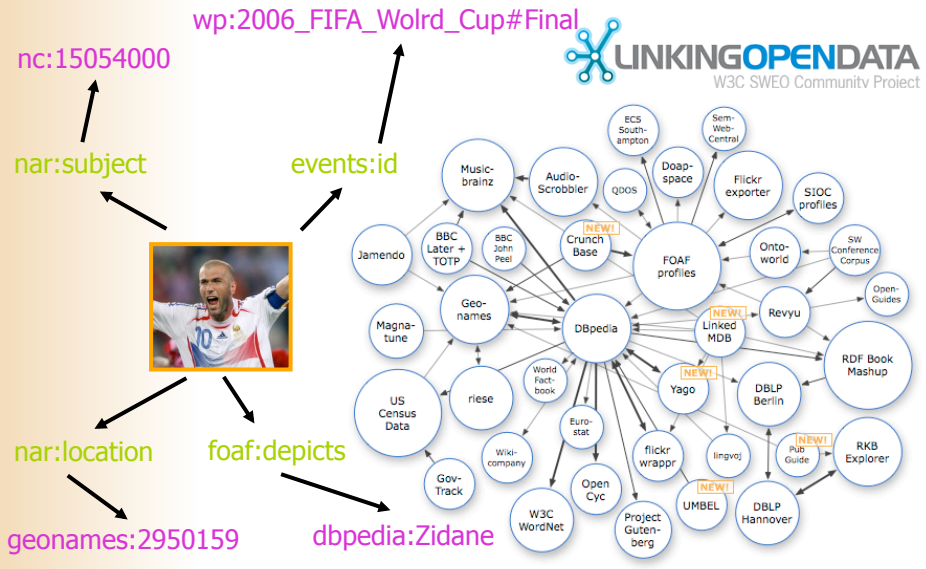


Step 4: Enriching the News Metadata



cc BY NC SA

Web of Data and Linked Data



cc BY NC SA

Presenting News Information

- Dimensions used for searching news items
  - When *time* 10/07/2006
  - Where *location* Paris
  - What *is depicted* J. Chirac, Z. Zidane
  - Why *event* WC 2006
  - Who *photographer* Bertrand Guay, AFP

} **Metadata**

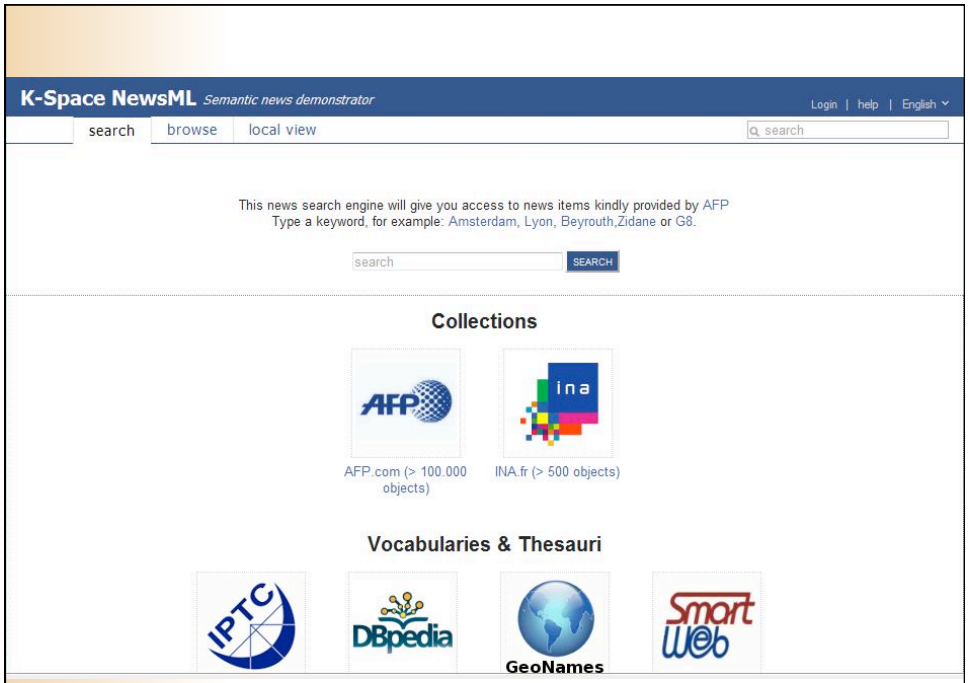
cc BY NC SA

Semantic Search of Multimedia News

Description	Number of RDF Triples
General Ontologies: NAR, DC, FOAF	7,336
Domain Specific Ontologies: football	104,358
Thesauri: newscodes	34,903
DBpedia, Geonames	53,468
AFP News Feed (June/July 2006)	804,446
AFP Photos (June/July 2006)	61,311
INA Broadcast Video (June/July 2006)	1,932
Total	1,067,754

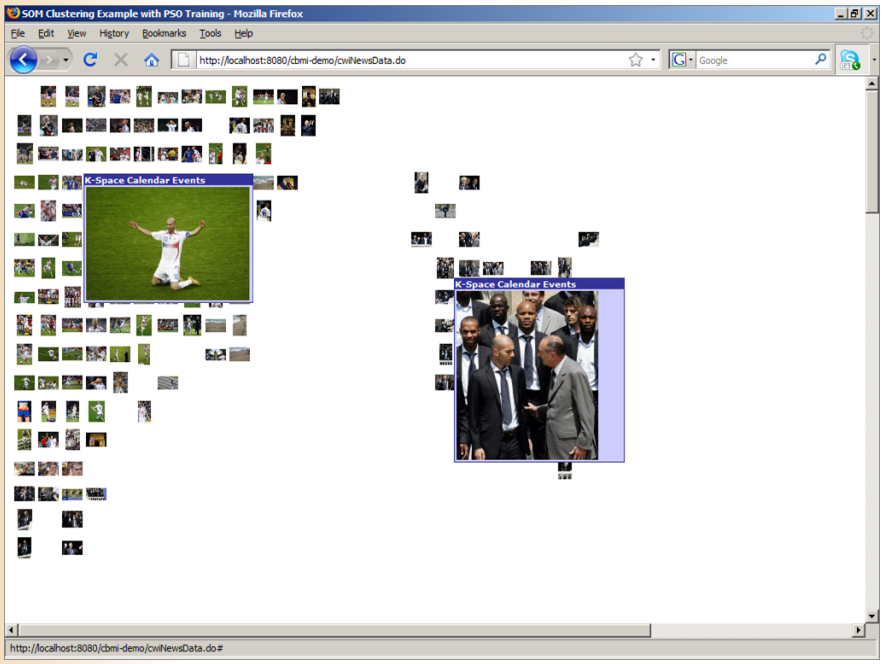
Powered by ClioPatria 1.0 alpha 3

cc BY NC SA



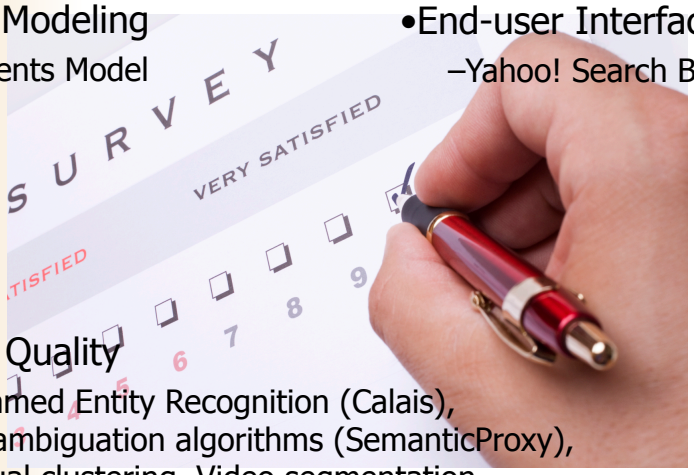
Conclusion

- 4-Steps methodology for building an ontology-based news infrastructure
  - UML-2-OWL: Flatten XML structure, Identify all resources
  - SKOS-ify existing thesauri and use the Web of Data
  - Reuse what is there ... Expose what you make
- Enrich metadata with text and visual analysis
  - Provide new dimensions (facets) for browsing the data
    - Ex: distinguish field images vs stadium and street images with a grass detector for the World Cup dataset



Future Work

- Data Modeling
  - –Events Model
- End-user Interfaces
  - –Yahoo! Search BOSS

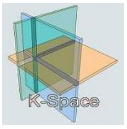


- Data Quality
  - –Named Entity Recognition (Calais), Disambiguation algorithms (SemanticProxy), Visual clustering, Video segmentation

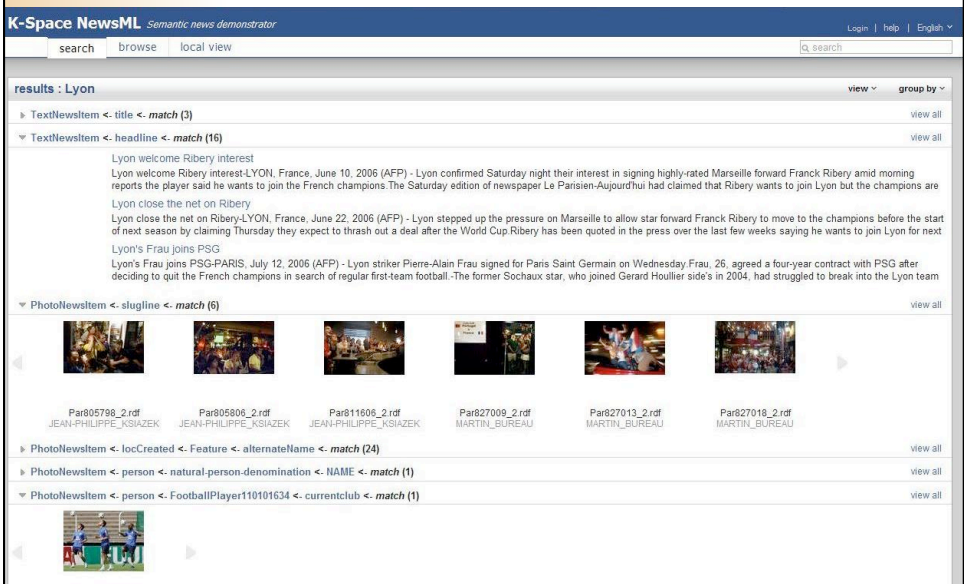
Credits

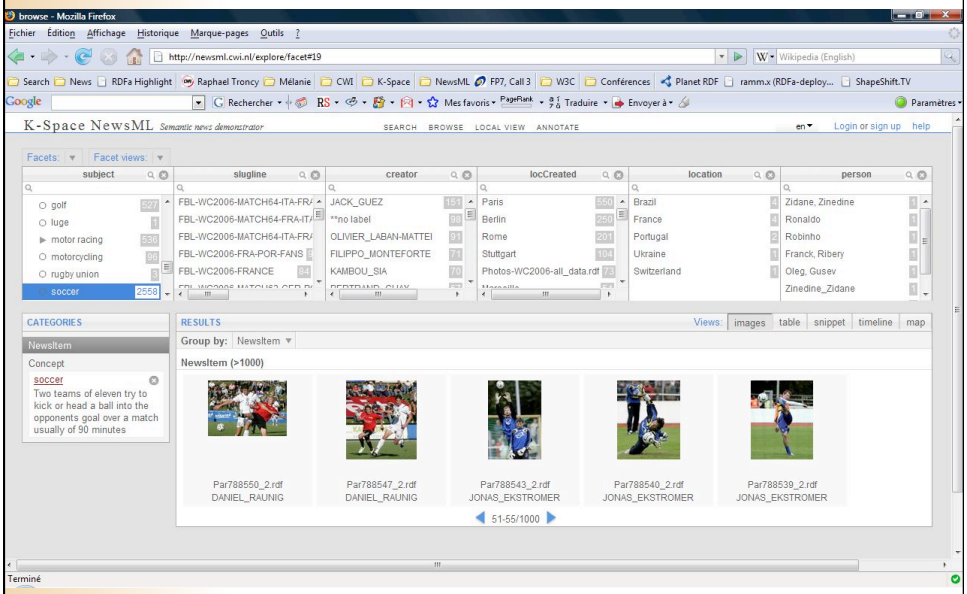
- Datasets:

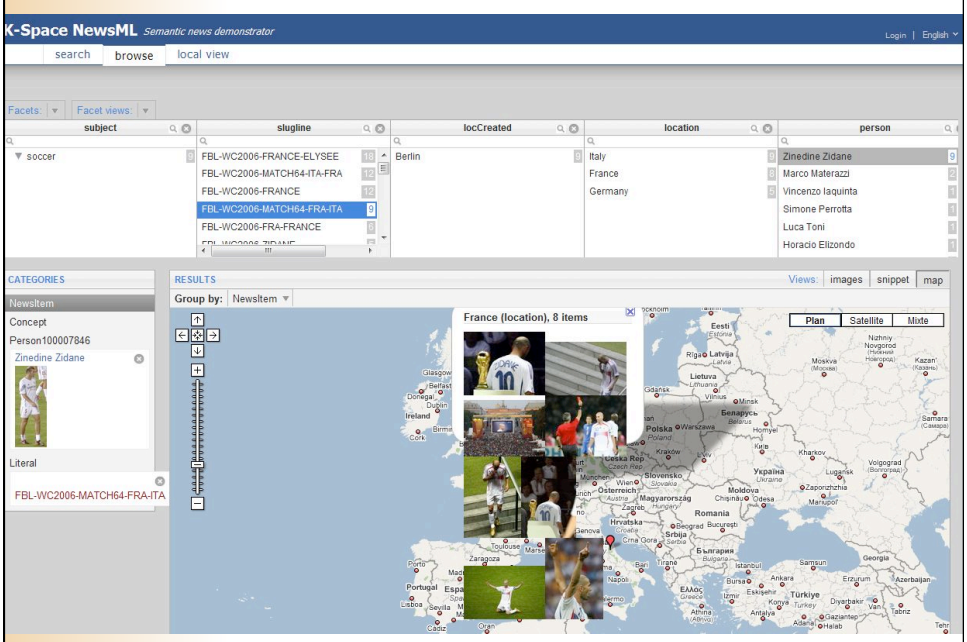
- People:

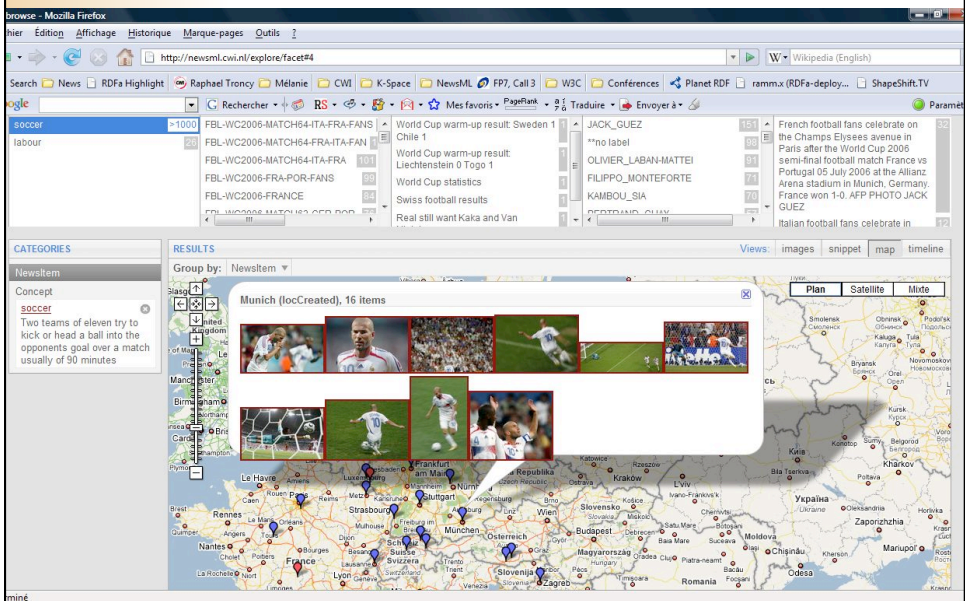


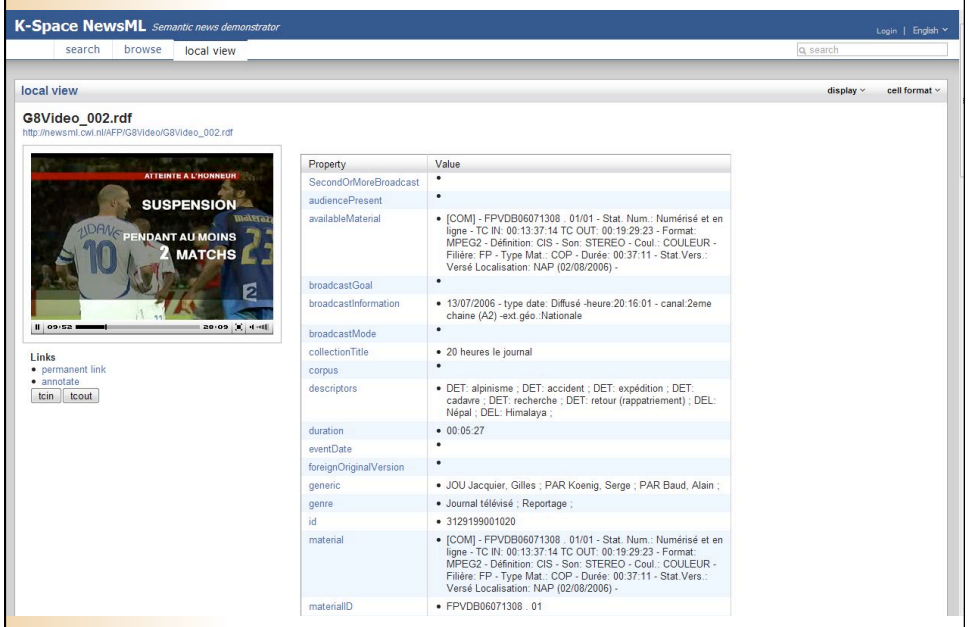
- More info: **http://newsml.cwi.nl**




browse - Mozilla Firefox
http://newsml.cwi.nl/explore/facet#19
K-Space NewsML Semantic news demonstrator
SEARCH BROWSE LOCAL VIEW ANNOTATE
Facets: Facet views:
subject
slugline
creator
locCreated
location
person
golf
luge
motor racing
motorcycling
rugby union
soccer
FBL-WC2006-MATCH64-ITA-FRA
FBL-WC2006-MATCH64-FRA-ITA
FBL-WC2006-MATCH64-ITA-FRA
FBL-WC2006-FRA-POR-FANS
FBL-WC2006-FRANCE
JACK_GUEZ
**no label
OLIVIER_LABAN-MATTEI
FILIPPO_MONTEFORTE
KAMBOU_SIA
Paris
Berlin
Rome
Stuttgart
Photos-WC2006-all_data.rdf
Brazil
France
Portugal
Ukraine
Switzerland
Zidane, Zinedine
Ronaldo
Robinho
Franck, Ribery
Oleg, Gusev
Zinedine_Zidane
CATEGORIES
NewsItem
Concept
soccer
Two teams of eleven try to kick or head a ball into the opponents goal over a match usually of 90 minutes
RESULTS
Views: images table snippet timeline map
Group by: NewsItem
NewsItem (>1000)
Par788550_2.rdf DANIEL_RAUNIG
Par788547_2.rdf DANIEL_RAUNIG
Par788543_2.rdf JONAS_EKSTROMER
Par788540_2.rdf JONAS_EKSTROMER
Par788539_2.rdf JONAS_EKSTROMER
51-55/1000
Terminé


K-Space NewsML Semantic news demonstrator
search browse local view
Facets: Facet views:
subject
slugline
locCreated
location
person
soccer
FBL-WC2006-FRANCE-ELYSEE
FBL-WC2006-MATCH64-ITA-FRA
FBL-WC2006-FRANCE
FBL-WC2006-MATCH64-FRA-ITA
FBL-WC2006-FRA-FRANCE
Berlin
Italy
France
Germany
Zinedine Zidane
Marco Materazzi
Vincenzo Iaquinta
Simone Perrotta
Luca Toni
Horacio Elizondo
CATEGORIES
NewsItem
Concept
Person100007846
Zinedine Zidane
Literal
FBL-WC2006-MATCH64-FRA-ITA
RESULTS
Views: images snippet map
Group by: NewsItem
France (location), 8 items
Plan Satellite Mixte


browse - Mozilla Firefox
CATEGORIES
RESULTS
Views: images snippet map timeline
Group by: NewsItem
Munich (locCreated), 16 items
Plan
Satellite
Mixte


K-Space NewsML Semantic news demonstrator
search
browse
local view
local view
G8Video_002.rdf
http://newsml.cwi.nl/AFP/G8Video/G8Video_002.rdf
ATTEINTE A L'HONNEUR
SUSPENSION
PENDANT AU MOINS
2 MATCHS
Links
permanent link
annotate
tcin
tcout
Property
Value
SecondOrMoreBroadcast
audiencePresent
availableMaterial
[COM] - FPVDB06071308 . 01/01 - Stat. Num.: Numérisé et en ligne - TC IN: 00:13:37:14 TC OUT: 00:19:29:23 - Format: MPEG2 - Définition: CIS - Son: STEREO - Coul.: COULEUR - Filière: FP - Type Mat.: COP - Durée: 00:37:11 - Stat.Vers.: Versé Localisation: NAP (02/08/2006) -
broadcastGoal
broadcastInformation
13/07/2006 - type date: Diffusé -heure:20:16:01 - canal:2eme chaine (A2) -ext.géo.:Nationale
broadcastMode
collectionTitle
20 heures le journal
corpus
descriptors
DET: alpinisme ; DET: accident ; DET: expédition ; DET: cadavre ; DET: recherche ; DET: retour (rappatriement) ; DEL: Népal ; DEL: Himalaya ;
duration
00:05:27
eventDate
foreignOriginalVersion
generic
JOU Jacquier, Gilles ; PAR Koenig, Serge ; PAR Baud, Alain ;
genre
Journal télévisé ; Reportage ;
id
3129199001020
material
[COM] - FPVDB06071308 . 01/01 - Stat. Num.: Numérisé et en ligne - TC IN: 00:13:37:14 TC OUT: 00:19:29:23 - Format: MPEG2 - Définition: CIS - Son: STEREO - Coul.: COULEUR - Filière: FP - Type Mat.: COP - Durée: 00:37:11 - Stat.Vers.: Versé Localisation: NAP (02/08/2006) -
materialID
FPVDB06071308 . 01