
Entity Extraction and Consolidation for Social Web Content Preservation

Stefan Dietze¹, Diana Maynard², Elena Demidova¹,
Thomas Risse¹, Wim Peters²,
Katerina Doka³, Yannis Stavrakas³

¹ L3S Research Center, Hannover, Germany

² University Sheffield, UK

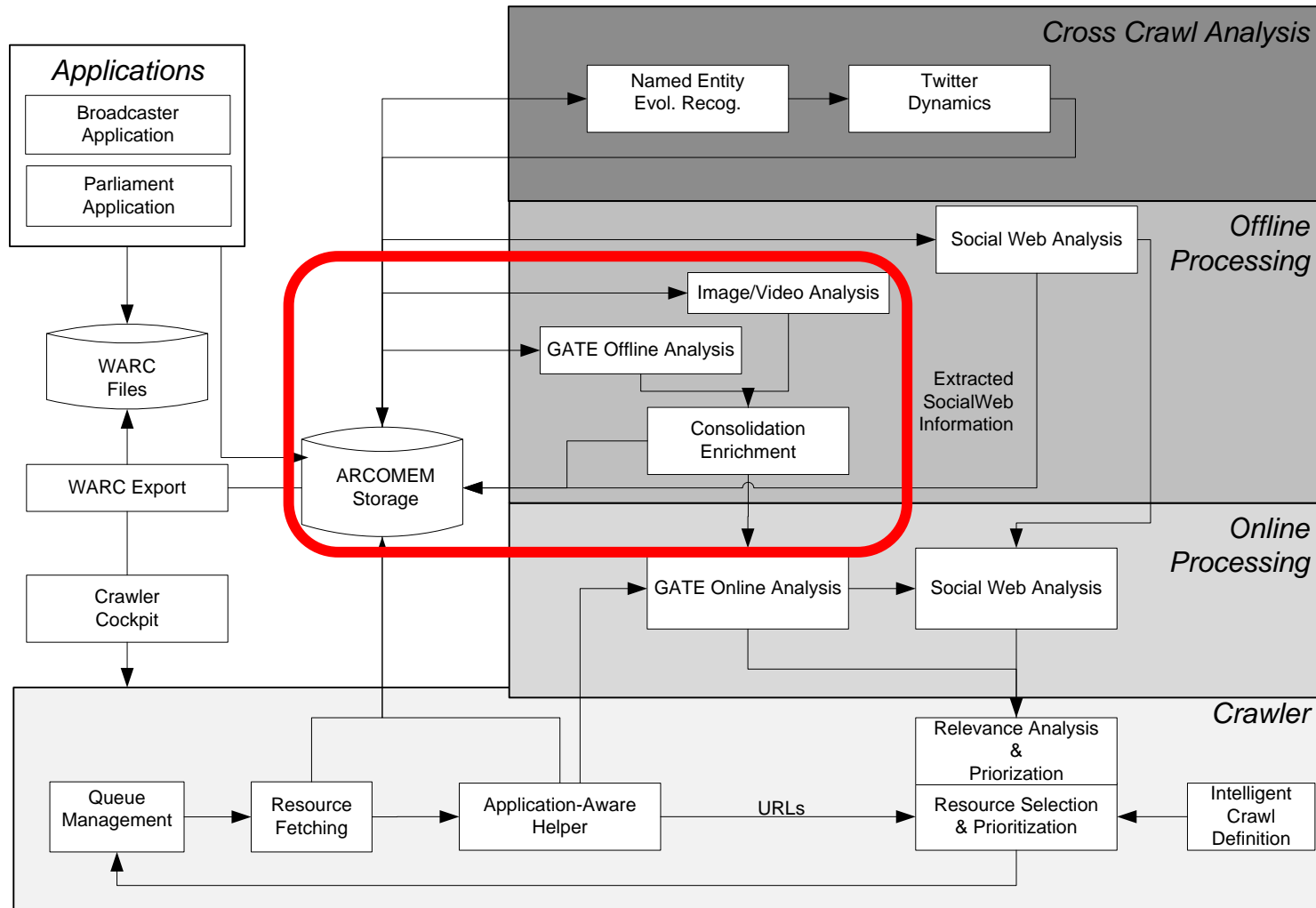
³ IMIS, RC ATHENA, Athens, Greece



SDA 2012, September 27, 2012



Architecture



The Extraction Components for Text

Aim

- **Extraction of Entities, Topics, Events and Opinions (ETOE)s from**
 - Web Pages
 - Social Web (Twitter, YouTube, Facebook, ...)

Challenges

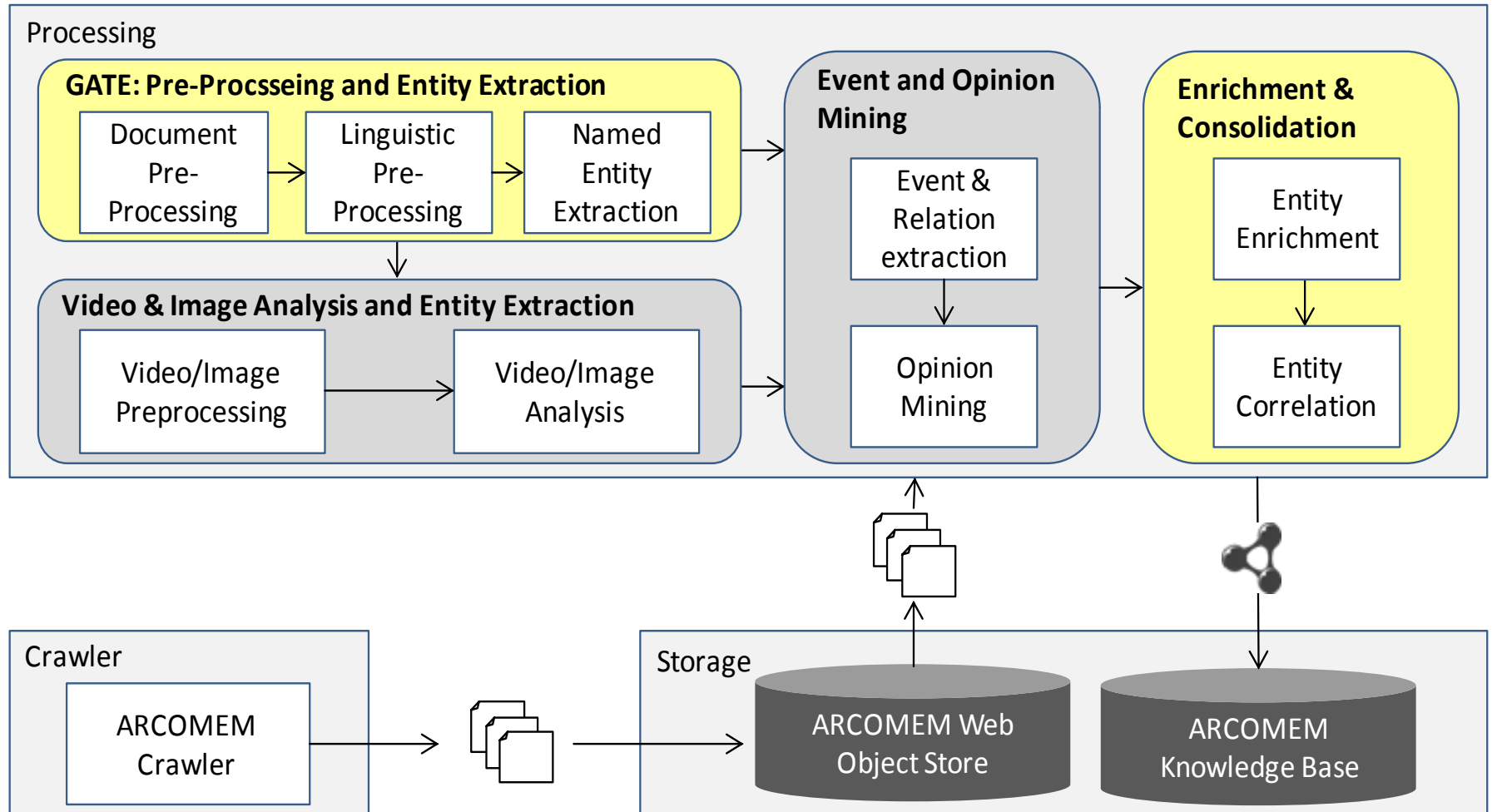
- **Entity recognition** from degraded input sources (tweets etc)
 - Advancing state of the art NLP and text mining
 - Dynamics detection: evolution of terms/entities
- **Semantic representation** of Web objects and entities
 - Appropriate RDF schemas for ETOE and Web objects
 - Exploiting (Linked Open) Web data to enrich extracted ETOE
- **Entity classification** (into events, locations, topics etc) & **consolidation**



SDA 2012, September 27, 2012



ETOE Processing Chain

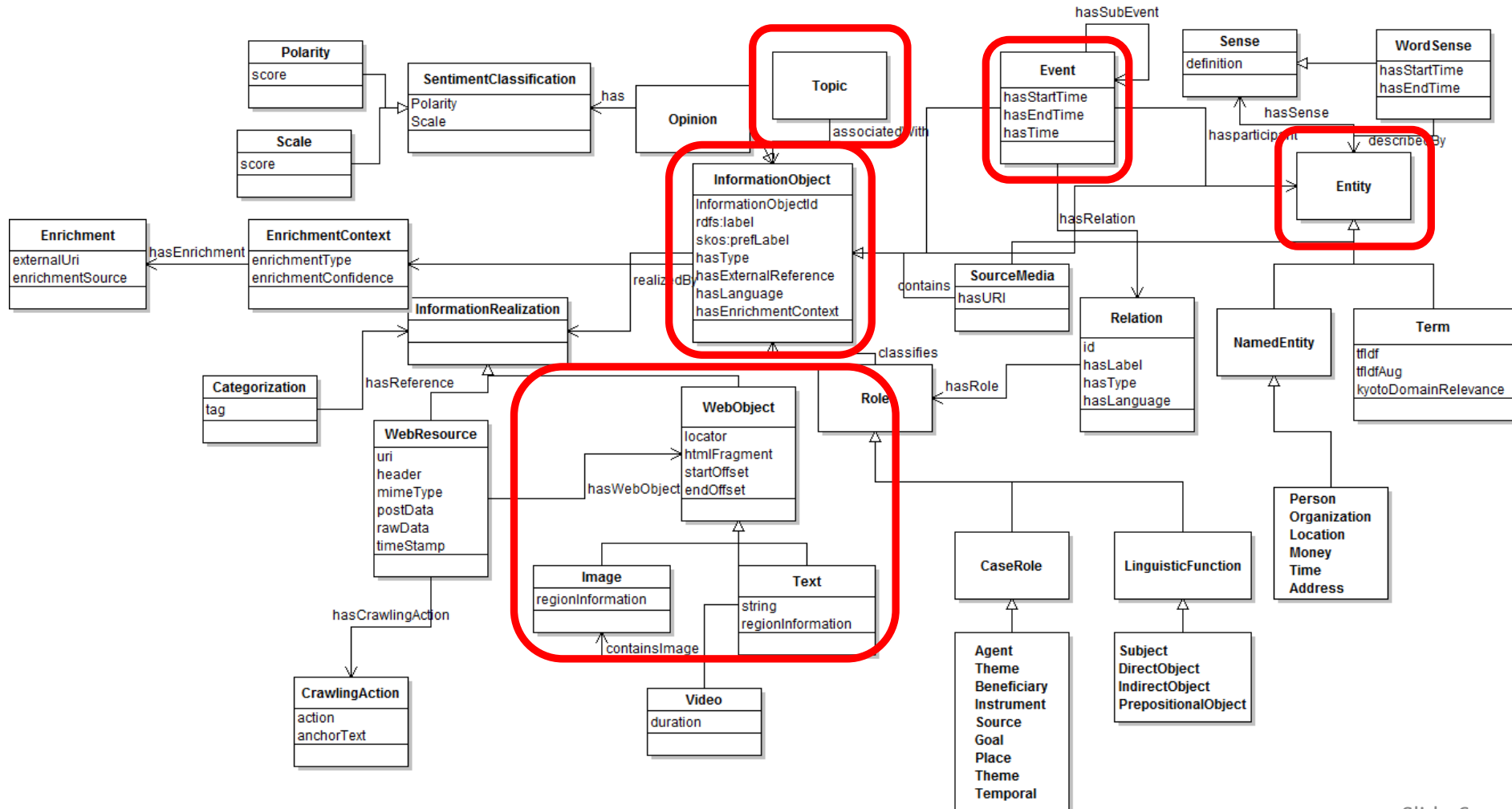


SDA 2012, September 27, 2012



RDF Schema for ARCOMEM Knowledge Base

- Relationships between ARCOMEM entities (ETOE etc) and information objects
- RDF schema: <http://www.gate.ac.uk/ns/ontologies/arcomem-data-model.rdf>



ETOE Extraction with GATE

ARCOMEM research challenges:

- Text processing in multiple languages (automated language detection)
- Language processing & entity recognition on social media/degraded texts (e.g. tweets)
- Entity classification (particularly wrt ETOE)

Progress so far:

- 3 adopted components for (a) term recognition, (b) entity recognition, and (c) event detection
- Languages: English & German (automated language detection)
- Applied to ARCOMEM use case data:
 - Greek financial crisis dataset: 84 Web documents from news sites, 32 Facebook posts, 41,000 tweets and 800 user comments
 - SWR Rock am Ring festival: 51 HTML documents (>3000 user comments)
 - Austrian Parliament crawl: ca 326 HTML and PDF documents



ETOE Extraction with GATE

The Greek Crisis: Trichet rejects ECB role as lender of last resort

The Greek Crisis

Tuesday, October 4, 2011

Trichet rejects ECB role as lender of last resort

Financial Times
October 4, 2011

Jean-Claude Trichet has dashed hopes that the European Central Bank will ride to the rescue of the eurozone by pledging to backstop crisis-hit member states.

In one of his last appearances as ECB president, Mr Trichet rejected the idea of the ECB acting as lender of last resort to governments. It was up to eurozone political leaders to restore investor confidence in Europe's monetary union, he told the European Parliament.

"It is their responsibility, individually and collectively, to ensure financial stability. It is the way Europe has been constructed and it is the way it seems to all of us, we must proceed. If it is said Mr Trichet, whose non-renewable

- Date
- Head
- Location
- Lookup
- Money
- MultiWord
- Organization
- Person
- Sentence
- SpaceToken
- Split
- TermCandidate
- Token
- deleted_MultiWord
- deleted_SelectedToken

MultiWord			
aug.tf.idf	4.392317422778761		X
canonical	monetary union		X
count	1.0		X
head	union		X
idf	4.392317422778761		X
lang	eng		X
local.tf.idf	4.392317422778761		X
tf.idf	4.392317422778761		X

candidate multi-word term

ETOE extraction results so far

- **Example entities (types):**
 - **ECB** (Organisation),
 - **Athens** (Location),
 - **Jean Claude Trichet** (Person)
- **Example queries:**
 - (1) Simple: Get **Web Objects** about **events** of type “**industrial action**”
=> <http://tinyurl.com/78ny7p5>
 - (2) Correlated: Get **Web objects** about **events** (arco:Event) in **Athens** (arco:Location) (involving the **IMF** (arco:Organisation))
=> <http://tinyurl.com/78uj5at>

Type	#Entities
arco:Time	51416
arco:Money	6335
arco:Event	759
arco:Organisation	15376
arco:Location	21218
arco:Person	4465
Total	99569

(+ large number of terms)



ETOE extraction results: evaluation

- **Manually created gold standard:**
Facebook posts, Financial Crisis Crawl
315 entities, 221 selected by at least two annotators
- **NE evaluation:** comparison of system results with gold standard
- **„Adjusted“:** exclusion of terms which were outside of annotated sentences (as system only considered terms as part of detected sentences) => increase of recall

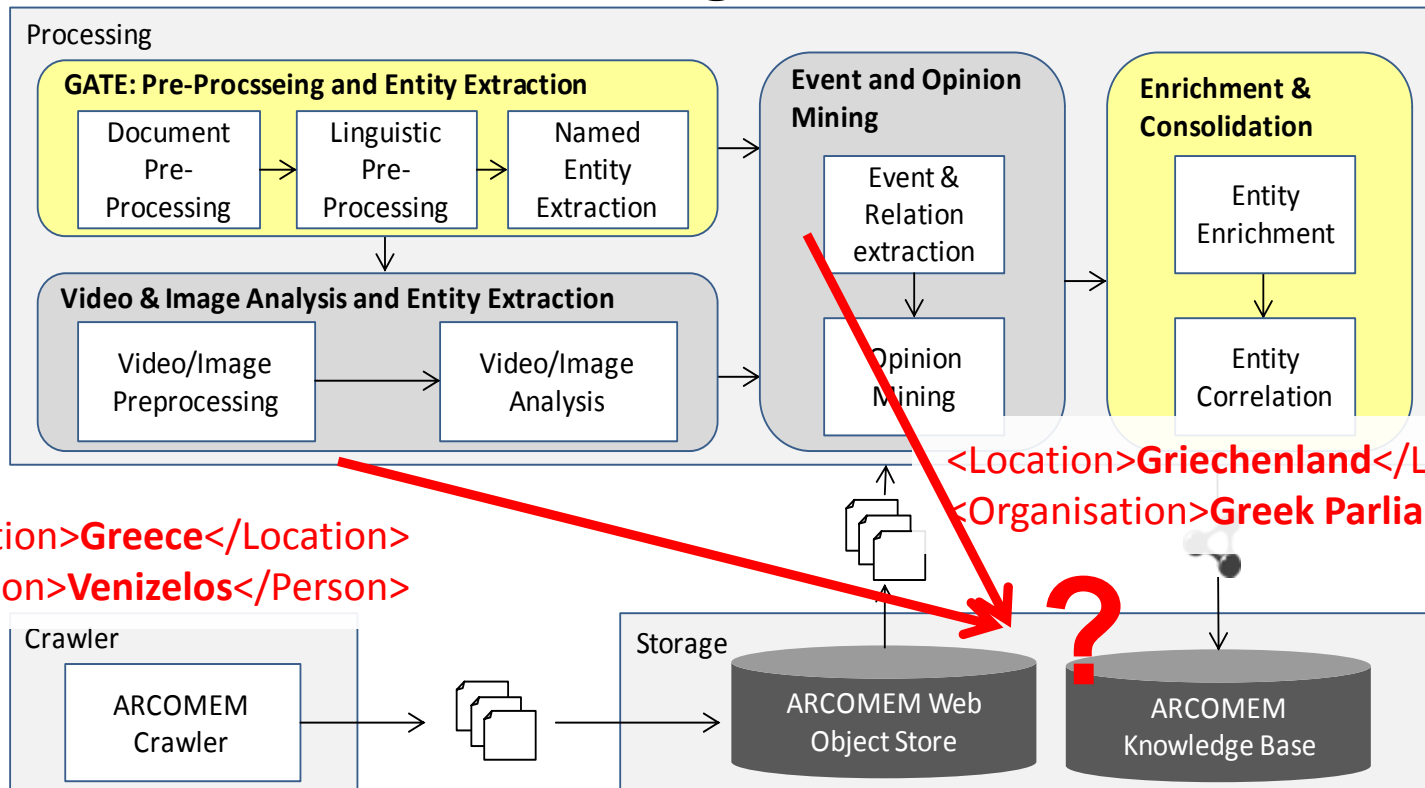
Task	Precision	Recall	F1
NE detection	80%	68%	74%
NE detection (adjusted)	80%	83.9%	81,9%
Type determination	98.8%	98.5%	98.6%
Full NE recognition	79%	67%	72.5%
Full NE recognition (adjusted)	79%	82.1%	80.5%



Data consolidation and integration problem

Data extracted from **different components** or during **different processing cycles** not aligned

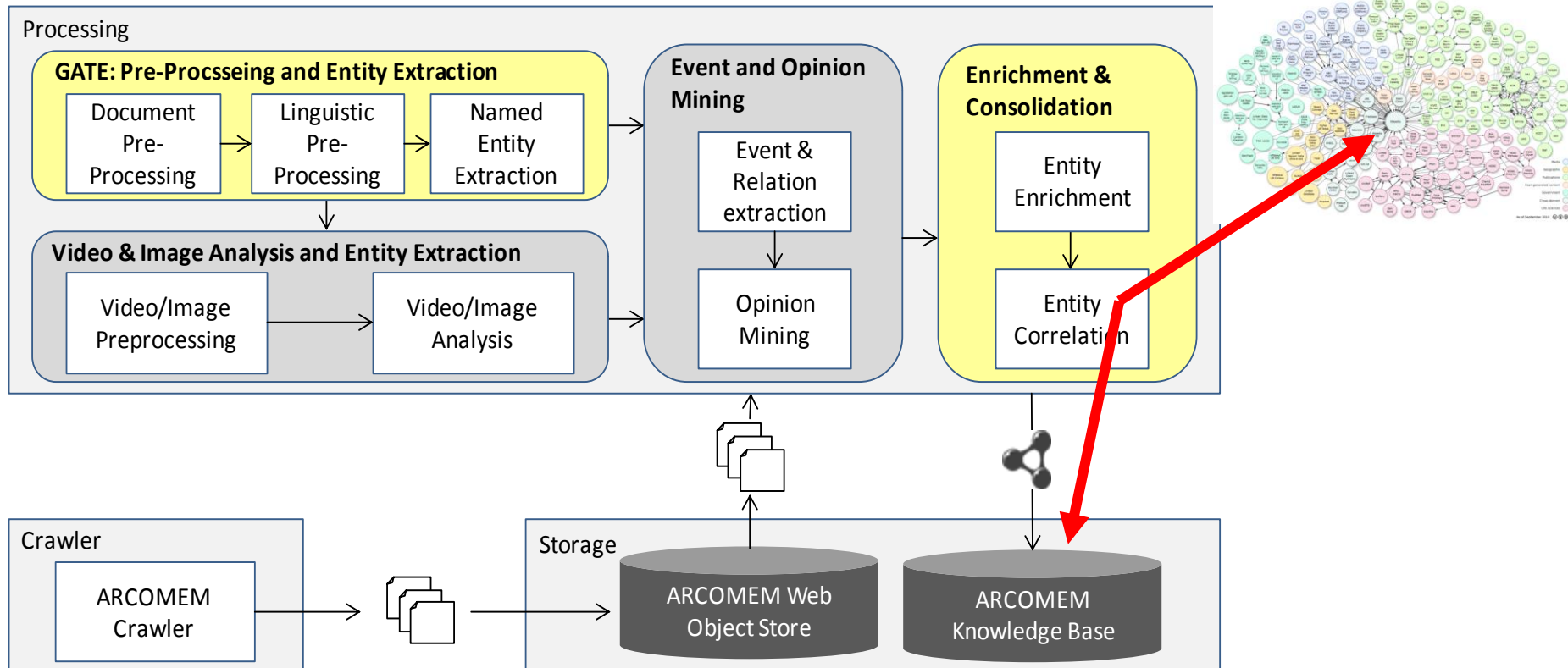
=> **consolidation, disambiguation & correlation** required.



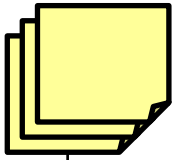
Data clustering & enrichment

Enrichment of entities with related references to Linked Data, particularly reference datasets (DBpedia, Freebase, ...)

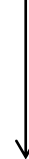
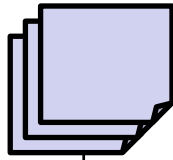
=> use enrichments for correlation/clustering/consolidation



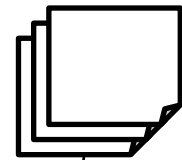
Enrichment for clustering and correlation: example



<Person>**Jean Claude Trichet**</Person>



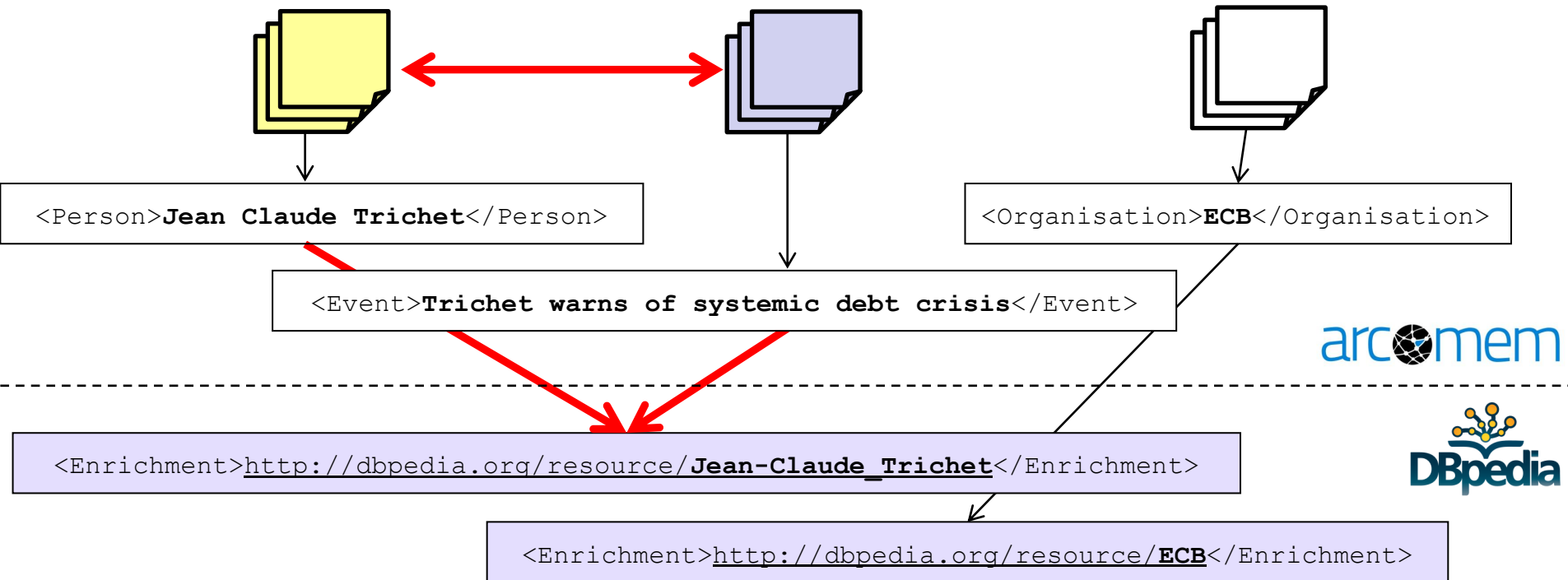
<Event>**Trichet warns of systemic debt crisis**</Event>



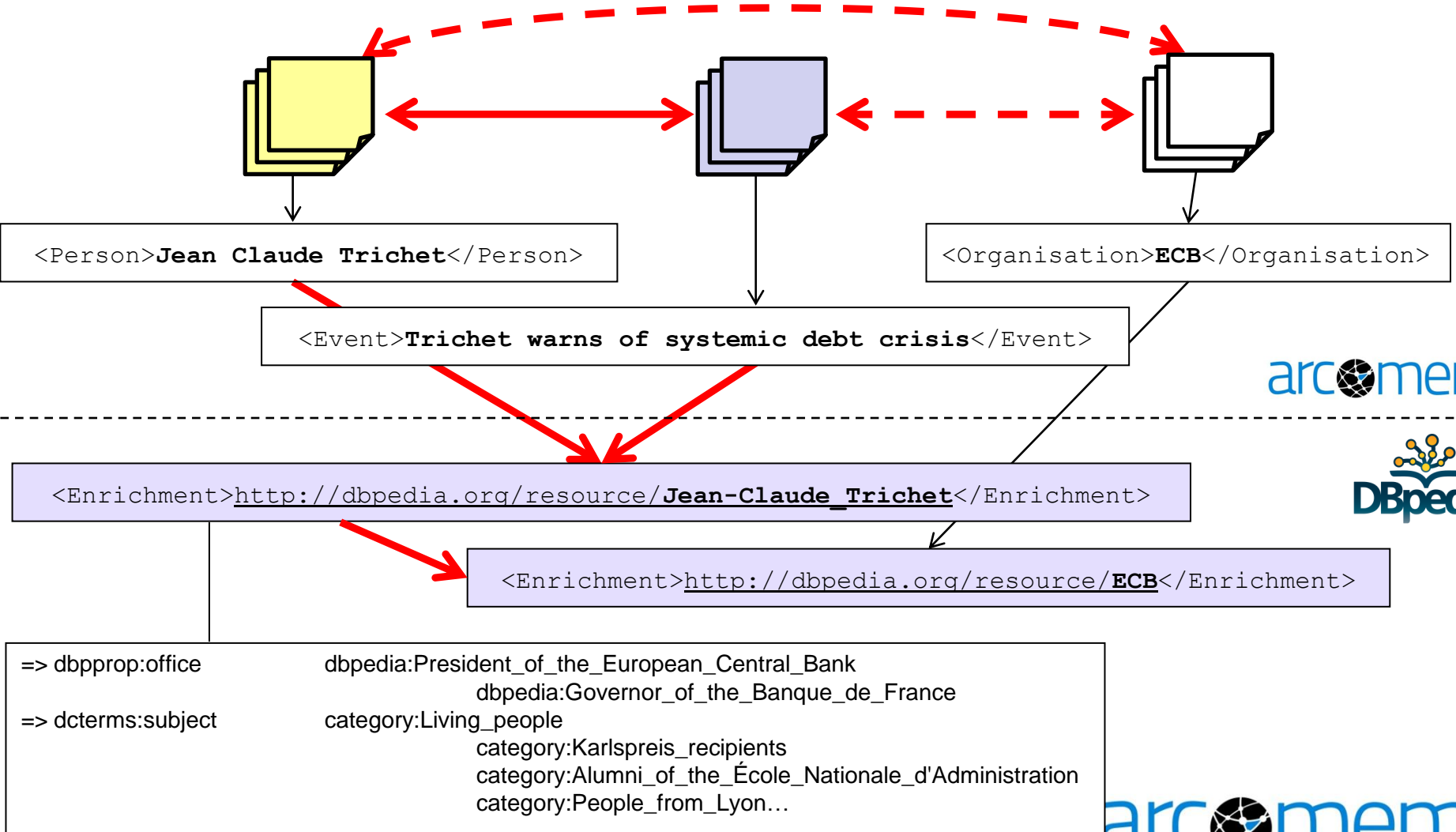
<Organisation>**ECB**</Organisation>



Enrichment for clustering and correlation: example



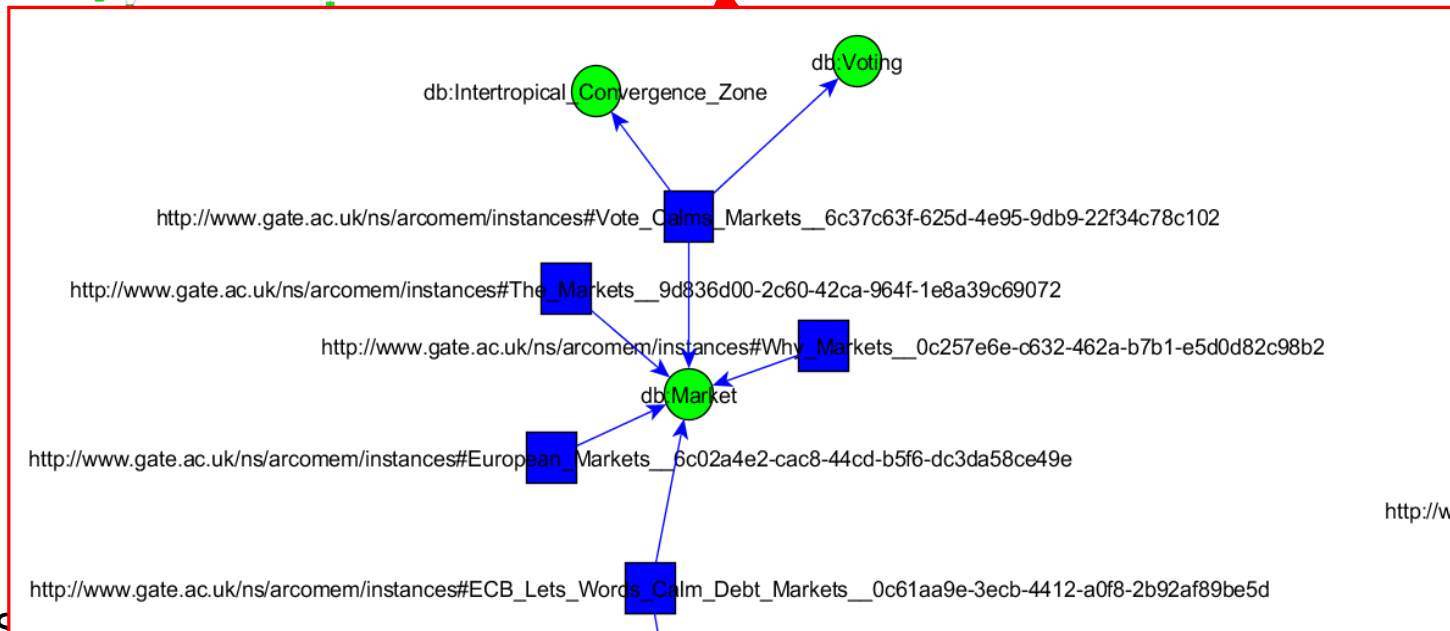
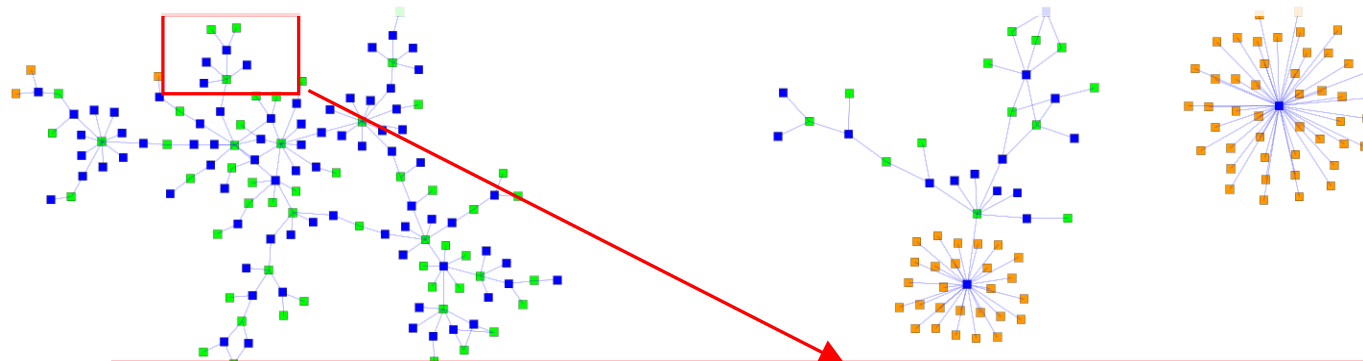
Enrichment for clustering and correlation: example



SDA 2012, September 27, 2012

ARCOMEM entities and enrichments - graph

- Nodes: entities/events (blue), enrichments DBpedia (green), Freebase (orange)
- 1013 clusters of correlated entities/events



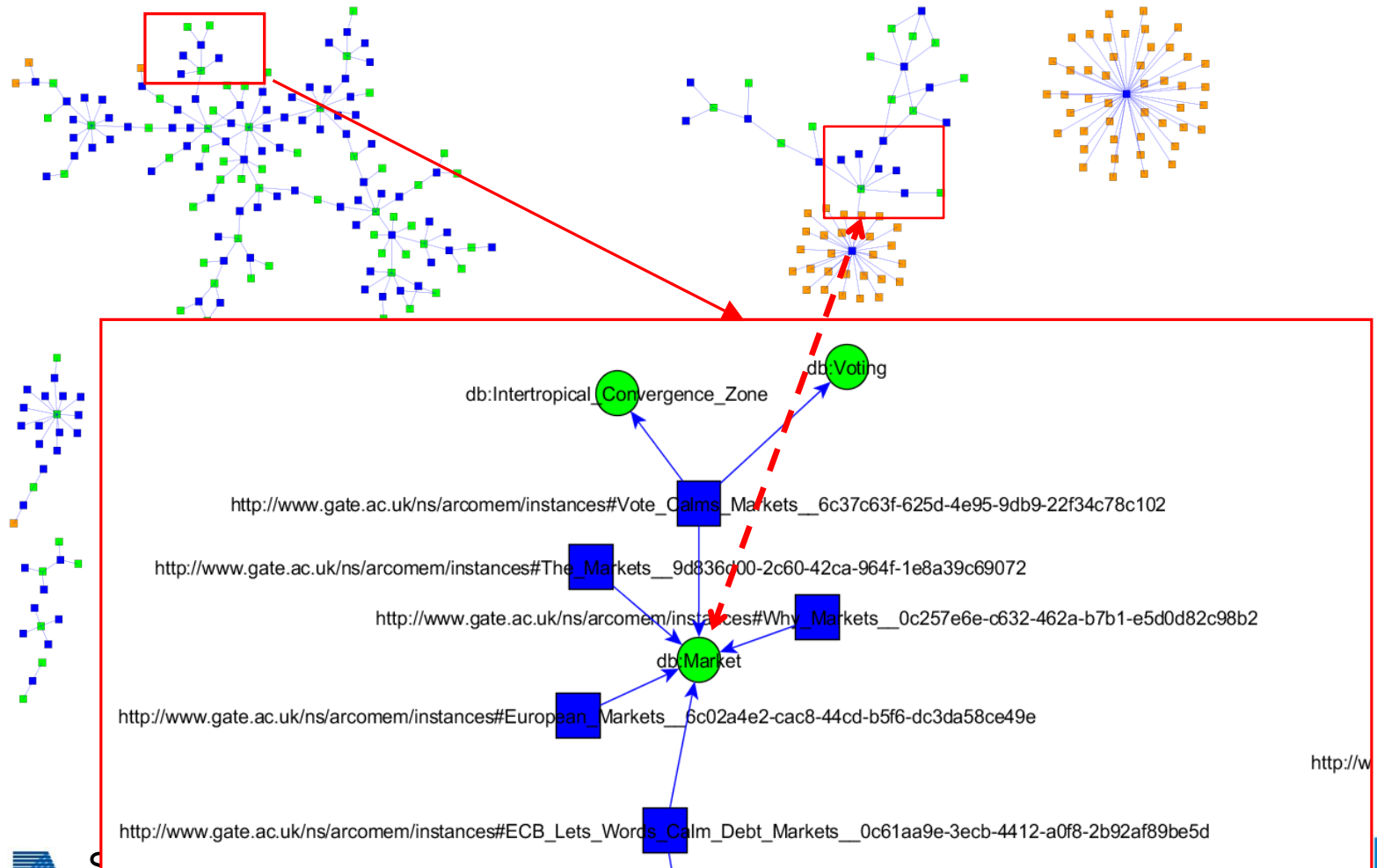
http://w

SDP 2012, September 27, 2012



ARCOMEM entities and enrichments - graph

- Nodes: entities/events (blue), enrichments DBpedia (green), Freebase (orange)
- 1013 clusters of correlated entities/events => **cluster expansion by considering related enrichments**



SDA 2012, September 27, 2012



Clustering of entities via enrichment relatedness

Discovery of “related” entities by discovering related enrichments

- (a) Retrieving possible paths between 2 enrichments (eg via RelFinder <http://www.visualdataweb.org/relfinder.php>)
- (b) Computation of relatedness measure (considering variables such as shortest path, number of paths, relationship types, number of directly connected edges of both enrichments...)
- (c) Clustering enrichments (entities) which are above certain threshold

The screenshot shows the RelFinder web application interface. The browser address bar displays www.visualdataweb.org/relfinder/relfinder.php. The application has two tabs: 'between' and 'examples'. In the 'between' tab, two input fields contain 'Jean-Claude Trichet' and 'ECB'. Below these fields are 'add', 'clear', and 'Find Relations' buttons. A 'Filter by:' section shows 'relations: (3/3)' and a table with columns 'length', 'class', 'link', and 'conne...'. The table has one row with '2' in the 'length' column, 'num' and 'vi' in the 'class' column, and '3/3' in the 'link' column. Below the table, there is a section for 'President of the E...' with a dropdown menu set to 'en' and a small image of Jean-Claude Trichet. The main area of the application displays a network graph with red arrows connecting nodes. The nodes are 'Jean-Claude Tric...', 'President of the E...', 'office', 'title', 'incumbent', 'leaderTitle', and 'European Central Ba...'. The graph shows a path from 'Jean-Claude Tric...' to 'President of the E...' via 'office', 'title', and 'incumbent', and another path from 'President of the E...' to 'European Central Ba...' via 'leaderTitle'.

Enrichment evaluation results

- Manual evaluation of 240 enrichment-entity pairs
- Available scores: 1 (correct), 0 (incorrect), 0.5 (vague or ambiguous relationship)

Entity Type	Average score DBPedia	Average score Freebase	Average Score Total
arco:Event	0.71		0.71
arco:Location	0.81	0.94	0.88
arco:Money	0.67		0.67
arco:Organization	0.93	1	0.97
arco:Person	0.9	0.89	0.89
arco:Time	0.74		0.74
Total	0.79	0.94	0.87



Outlook

Short term

- Investigation of reasons for enrichment noise
 - Ambiguous entities with no context (e.g. Athens in Greece vs. Athens in Greene County, New York).
 - Flaws in DBpedia Spotlight results, e.g. “Greek strategy on debt crisis” vs. “strategy games”
- Data quality in general
- Better support for degraded languages

Longer term

- Publication of ARCOMEM ETOE dataset
- Release of ETOE detection and clustering methods as general purpose tools

Related Workshop

- KECSM 2012: “Knowledge Extraction and Consolidation from Social Media”; related workshop at ISWC2012 => <http://blogs.ecs.soton.ac.uk/knowledgeextraction/>



SDA 2012, September 27, 2012



THANK YOU

CONTACT DETAILS

Dr. Thomas Risse
L3S Research Center
+49 511 762 17764
risse@L3S.de
www.arcomem.eu



SDA 2012, September 27, 2012

