Entity Extraction and Consolidation for Social Web Content Preservation

Stefan Dietze[1], Diana Maynard[2], Elena Demidova[1],
Thomas Risse[1], Wim Peters[2],
Katerina Doka[3], Yannis Stavrakas[3]

[1] L3S Research Center, Hannover, Germany

[2] University Sheffield, UK

[3] IMIS, RC ATHENA, Athens, Greece

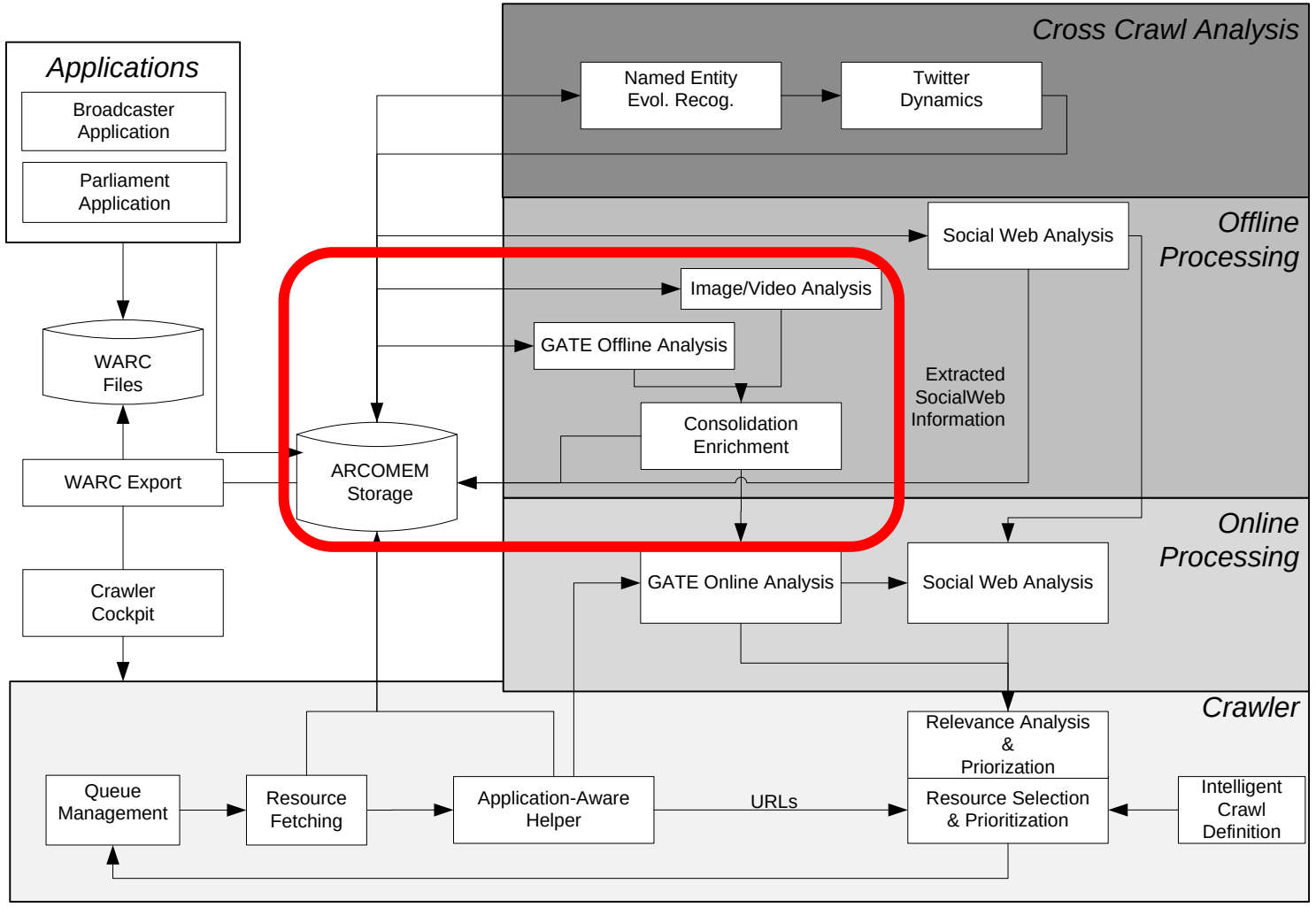
The ARCOMEM Approach

- **Make use of the Social Web**
  - Huge source of user generated content
  - Wide range of articulation methods
    From simple „I like it“-Buttons to complete articles
  - Represents the diversity of opinions of the public

- **User activities often triggered by**
  - Events and related entities
    (e.g. Sport Events, Celebrations, Crises, News Articles, Persons, Locations)
  - Topics (e.g. Global Warming, Financial Crisis, Swine Flu)

→ **A semantic-aware and socially-driven preservation model is a natural way to go**

Architecture



The Extraction Components for Text

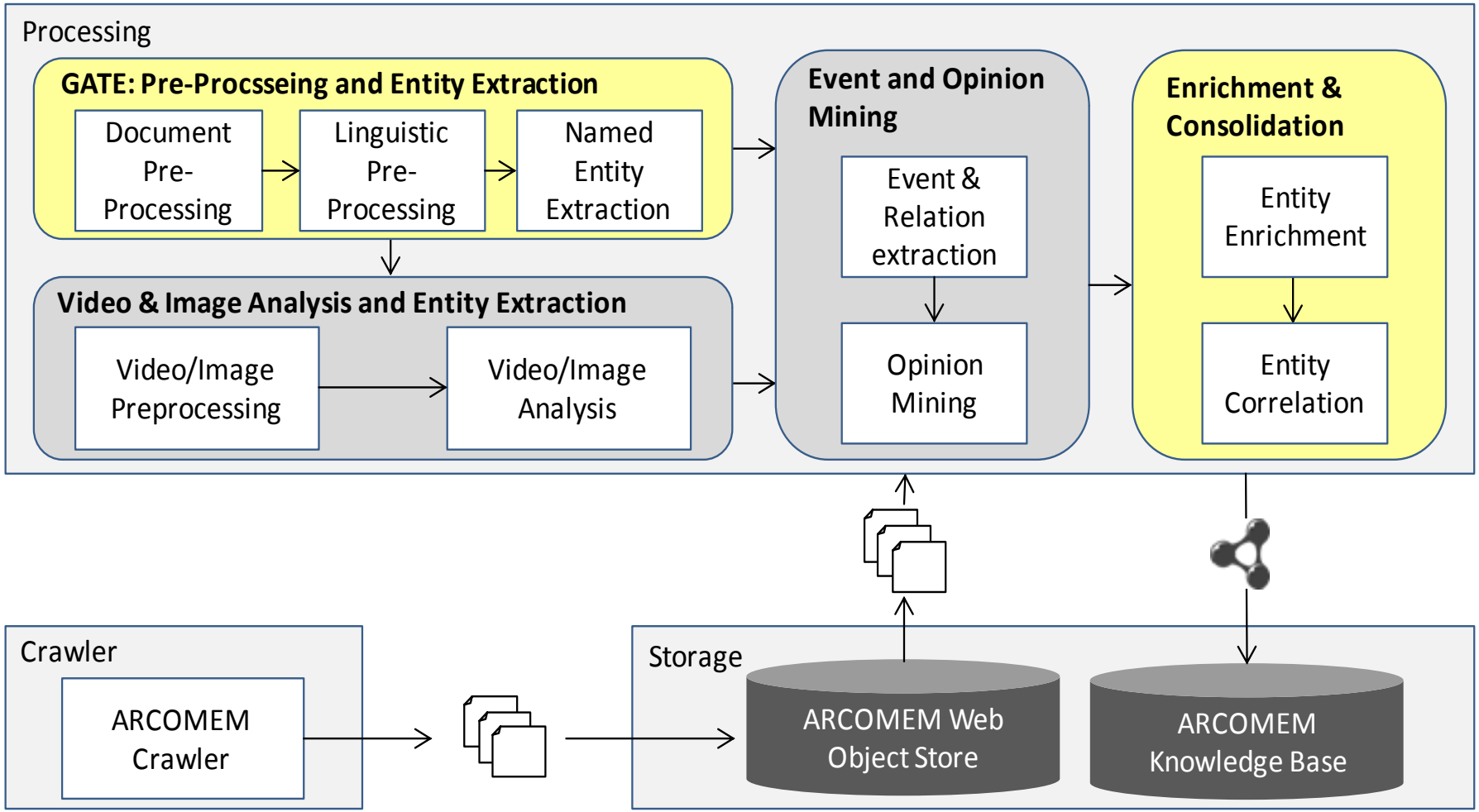
Aim

- **Extraction of Entities, Topics, Events and Opinions (ETOEs) from**
  - Web Pages
  - Social Web (Twitter, YouTube, Facebook, …)

Challenges

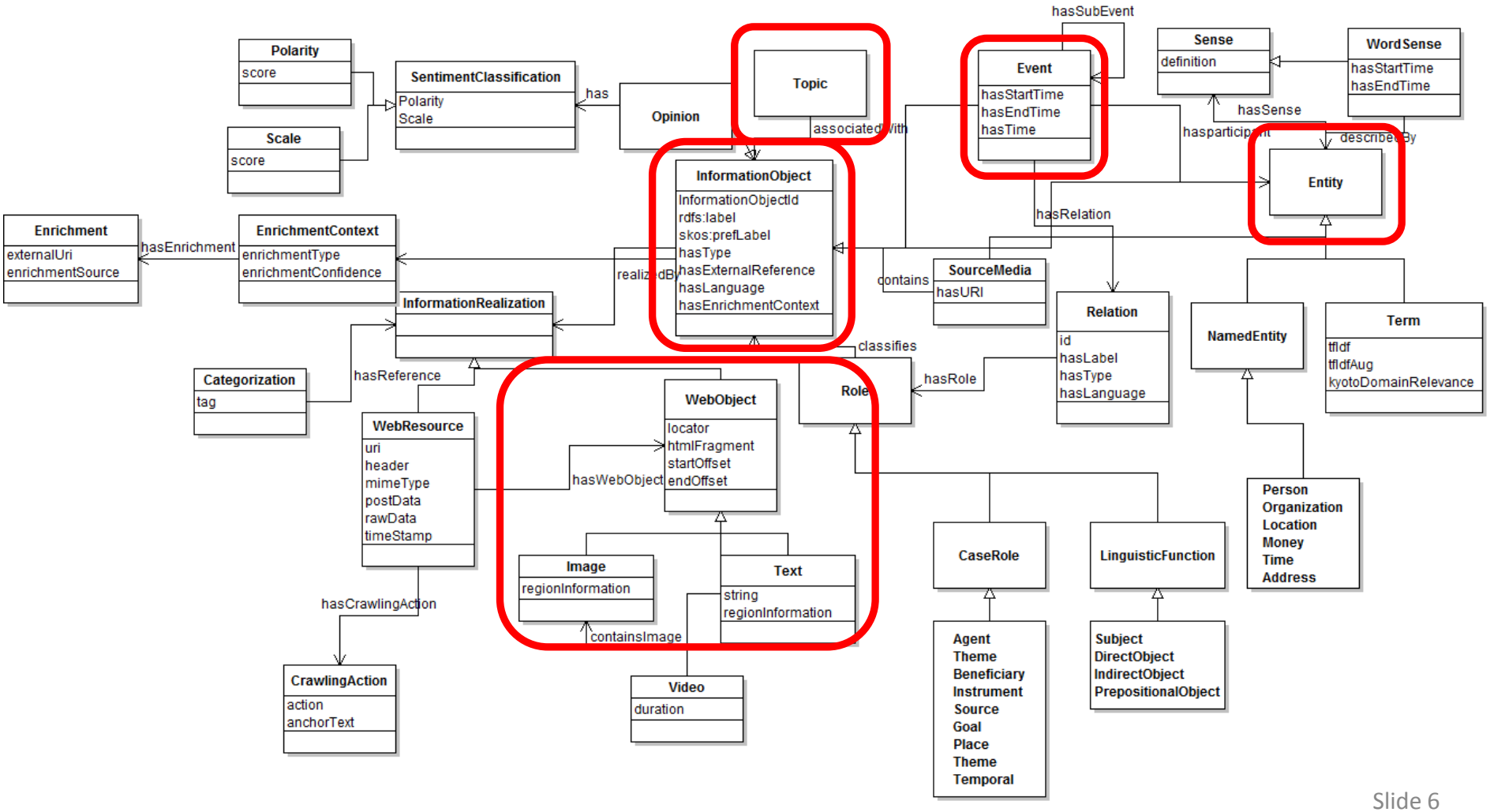
- **Entity recognition** from degraded input sources (tweets etc)
  - Advancing state of the art NLP and text mining
  - Dynamics detection: evolution of terms/entities
- **Semantic representation** of Web objects and entities
  - Appropriate RDF schemas for ETOE and Web objects
  - Exploiting (Linked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

ETOE Processing Chain



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

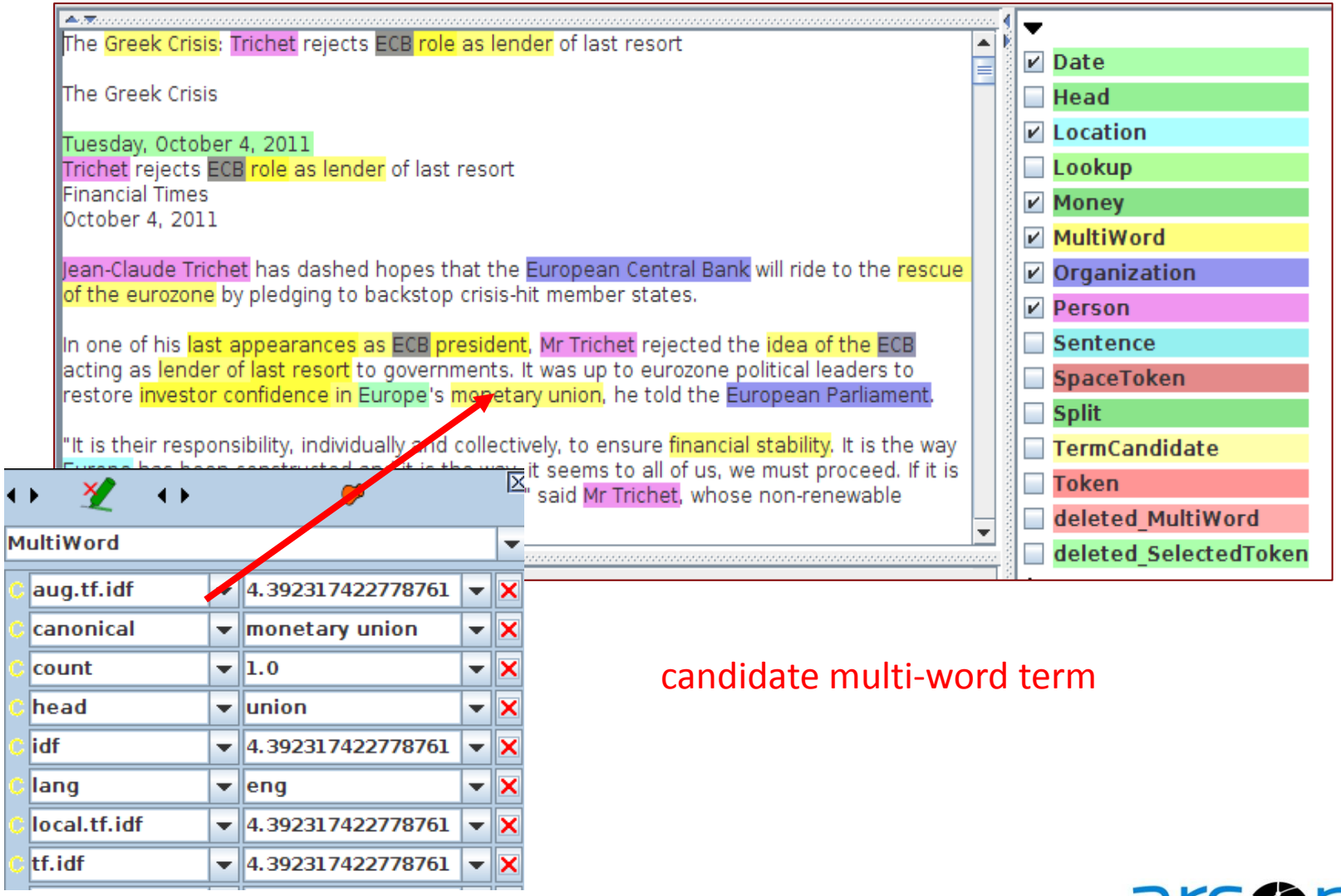
ARCOMEM research challenges:

- 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



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)

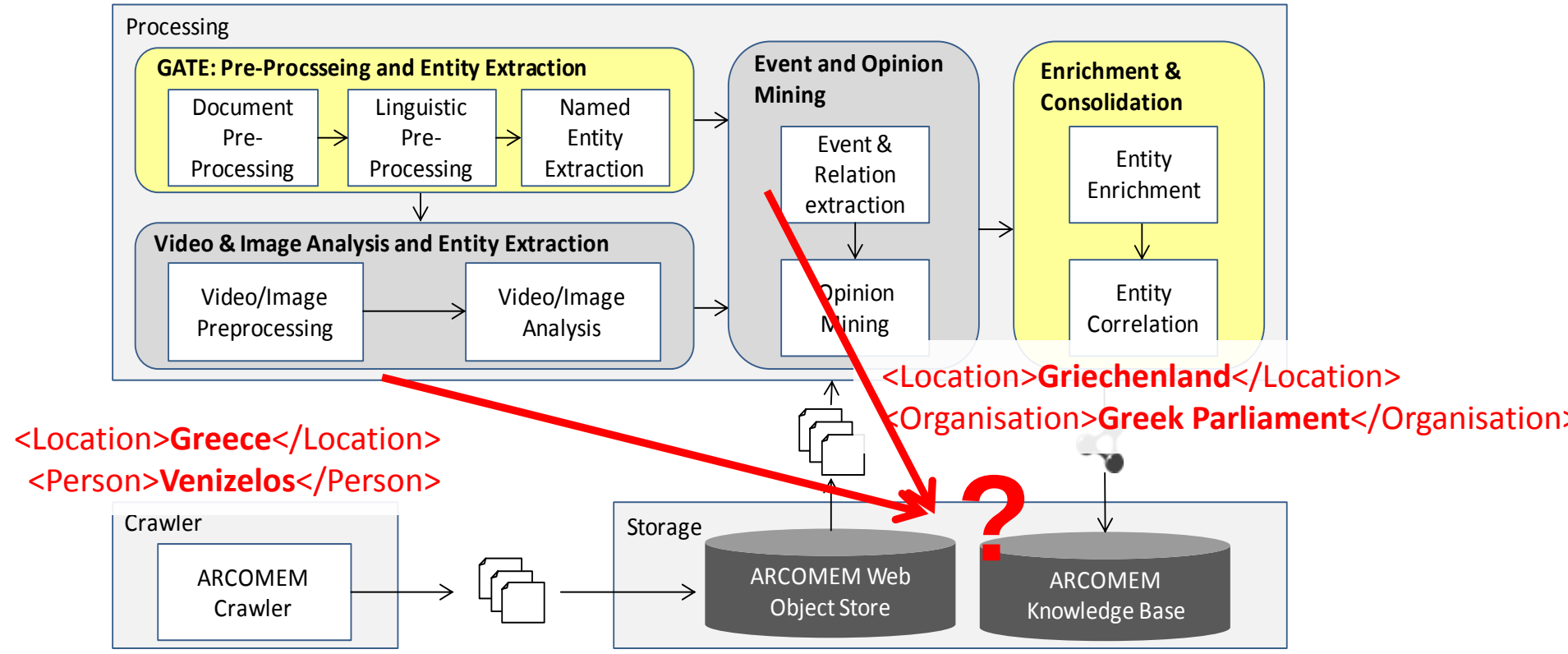
ETOE extraction results: evaluation

- **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

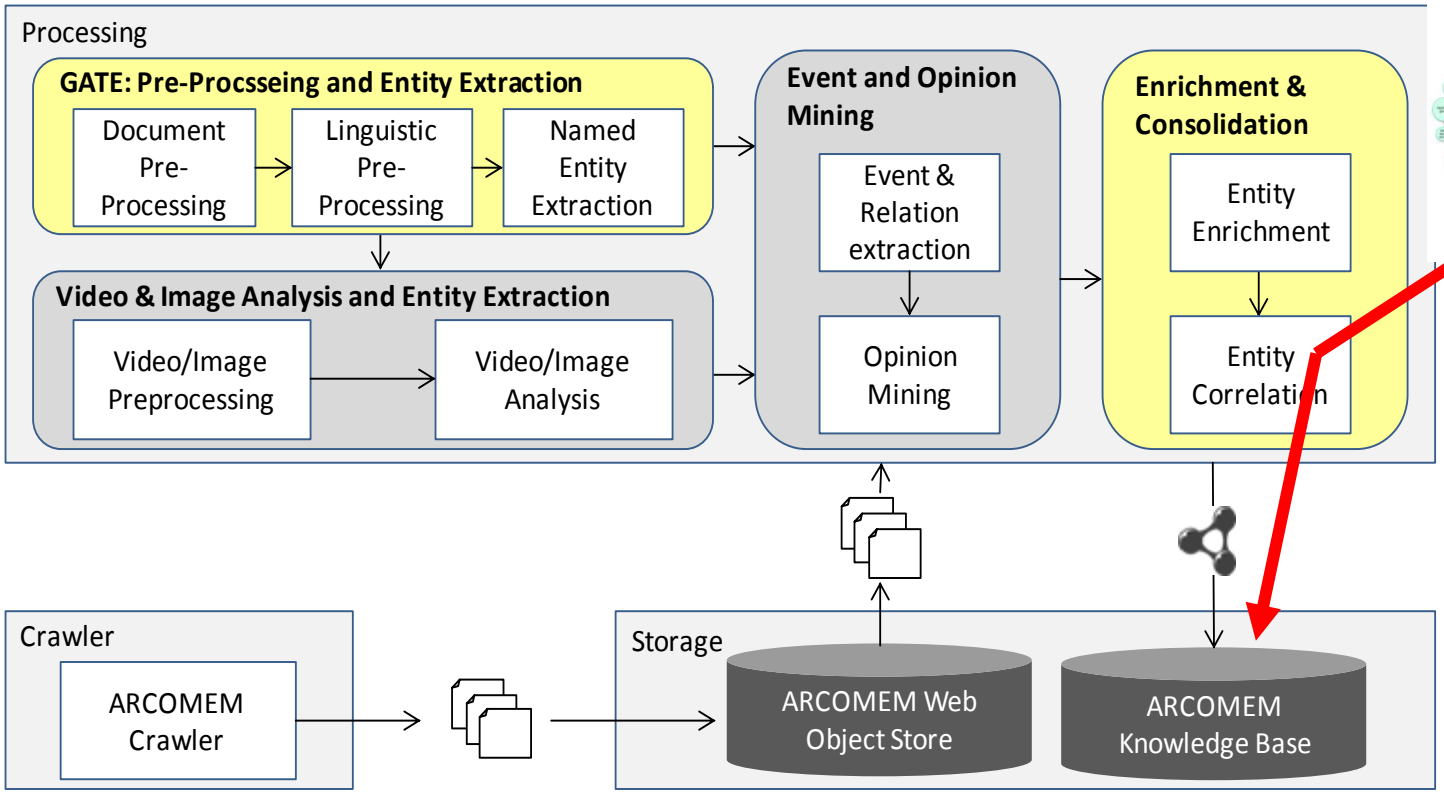
Data extracted from **different components** or during **different processing cycles** not aligned
=> consolidation, disambiguation & correlation required.

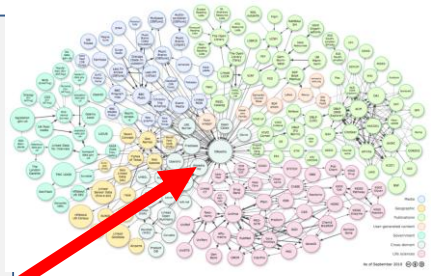


Data clustering & enrichment

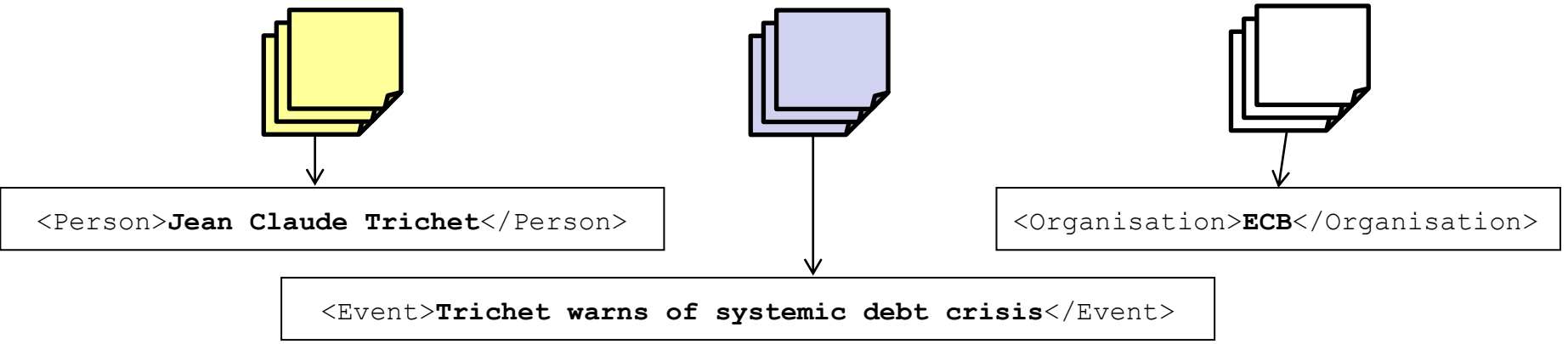
Enrichment of entities with related references to Linked Data, particularly reference datasets (DBpedia, Freebase, …)

=> use enrichments for correlation/clustering/consolidation

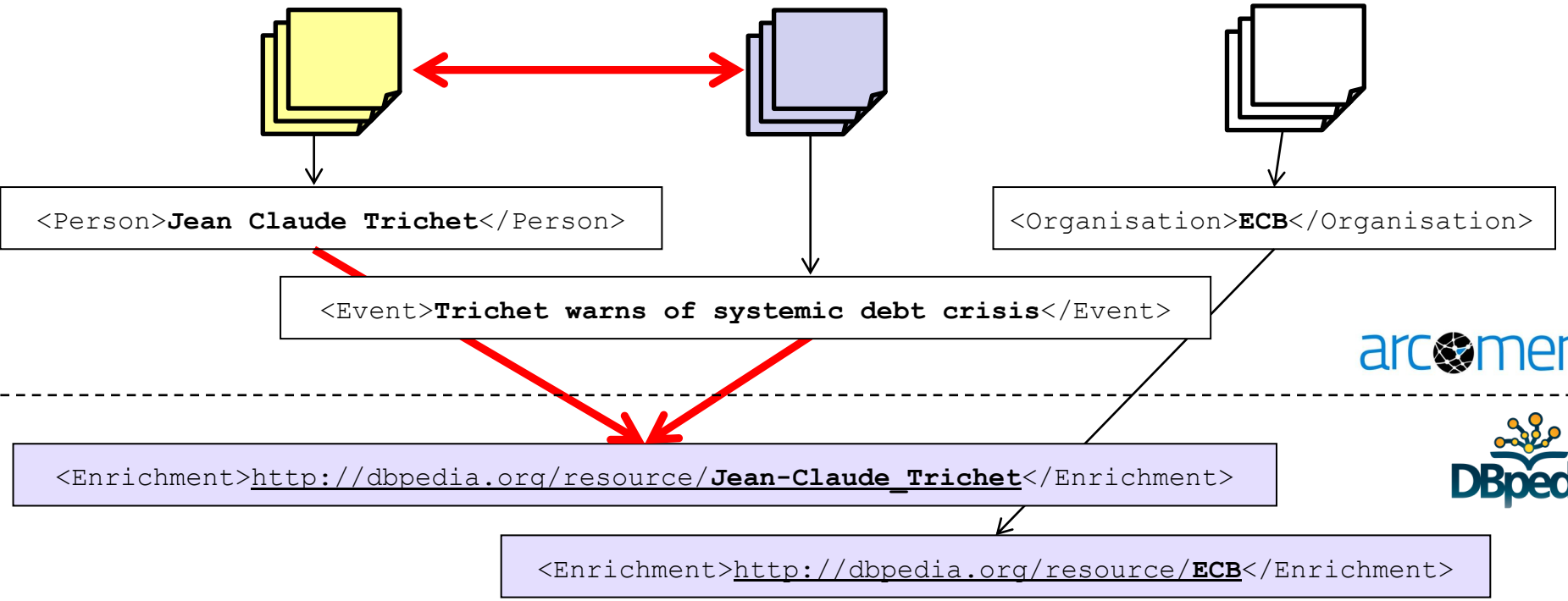




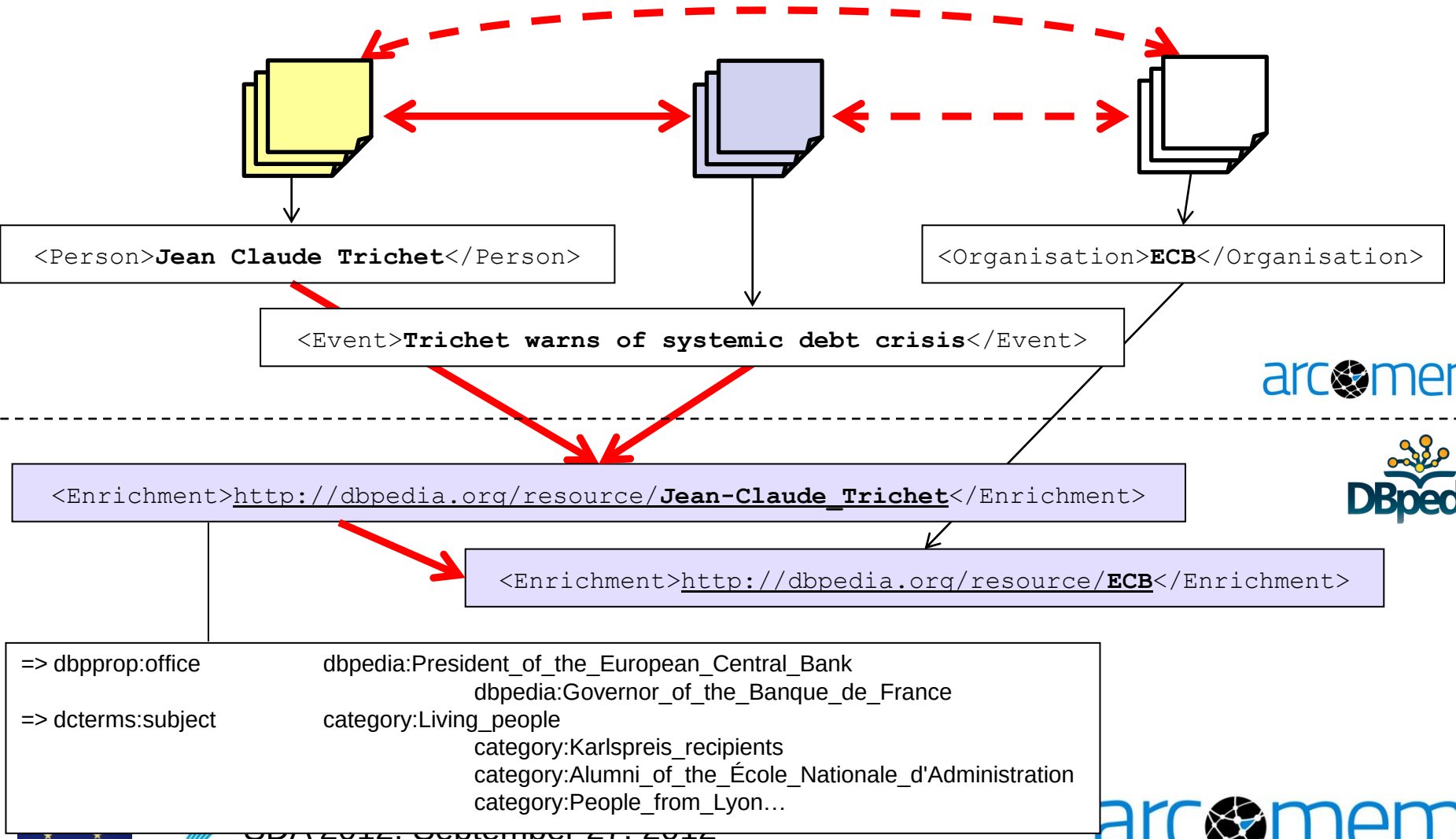
Enrichment for clustering and correlation: example



Enrichment for clustering and correlation: example

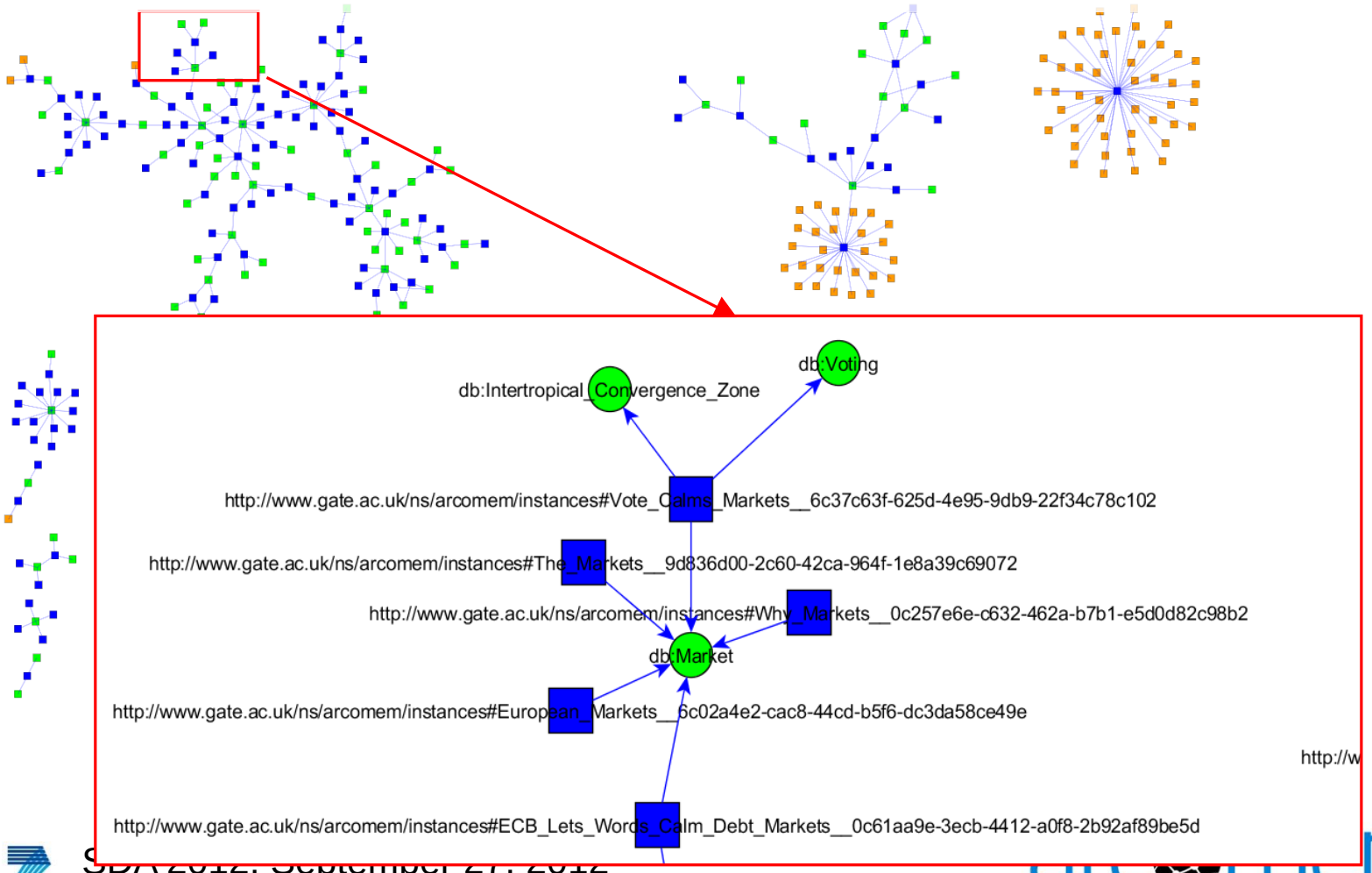


Enrichment for clustering and correlation: example



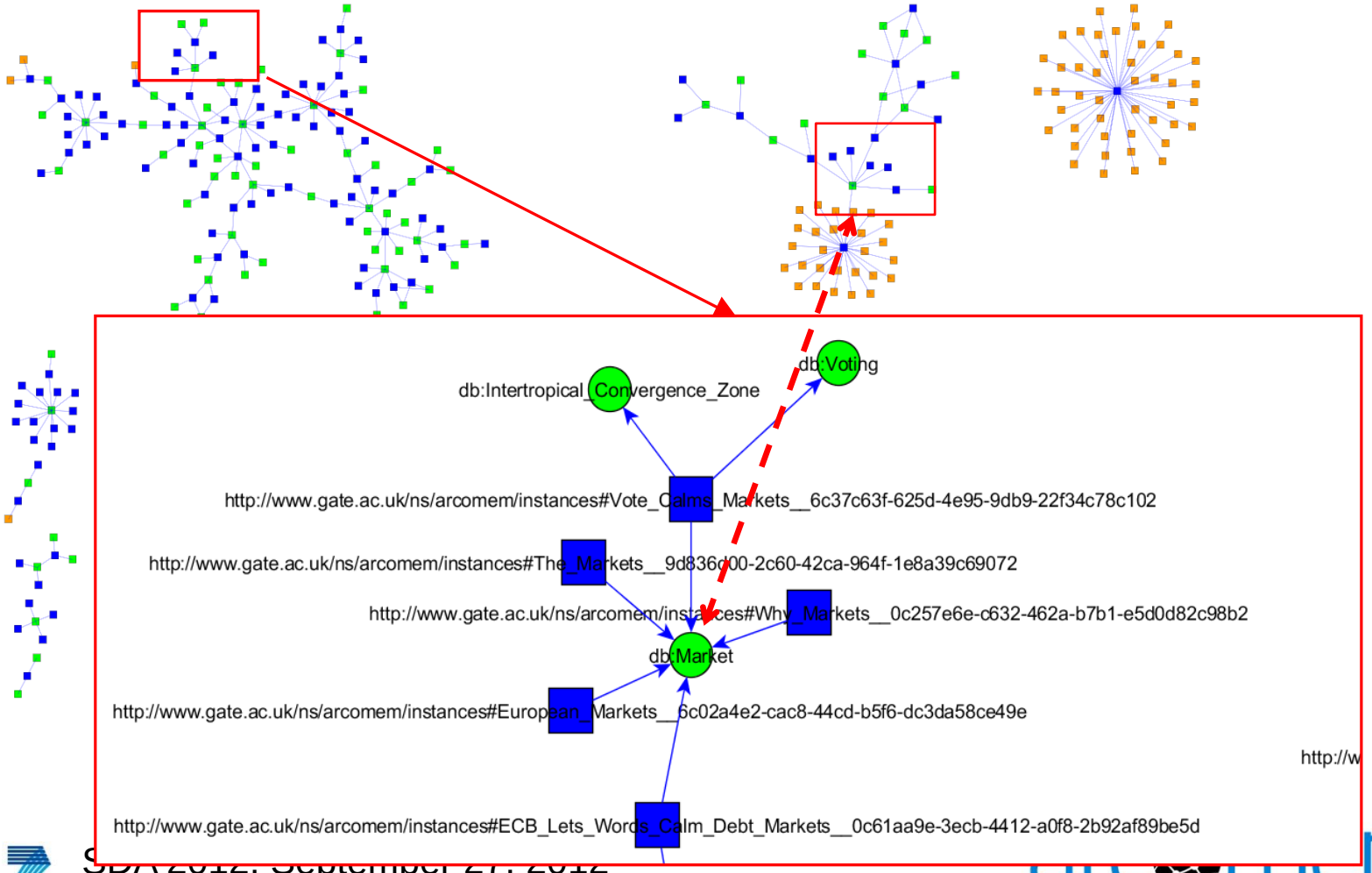
ARCOMEM entities and enrichments - graph

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



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**



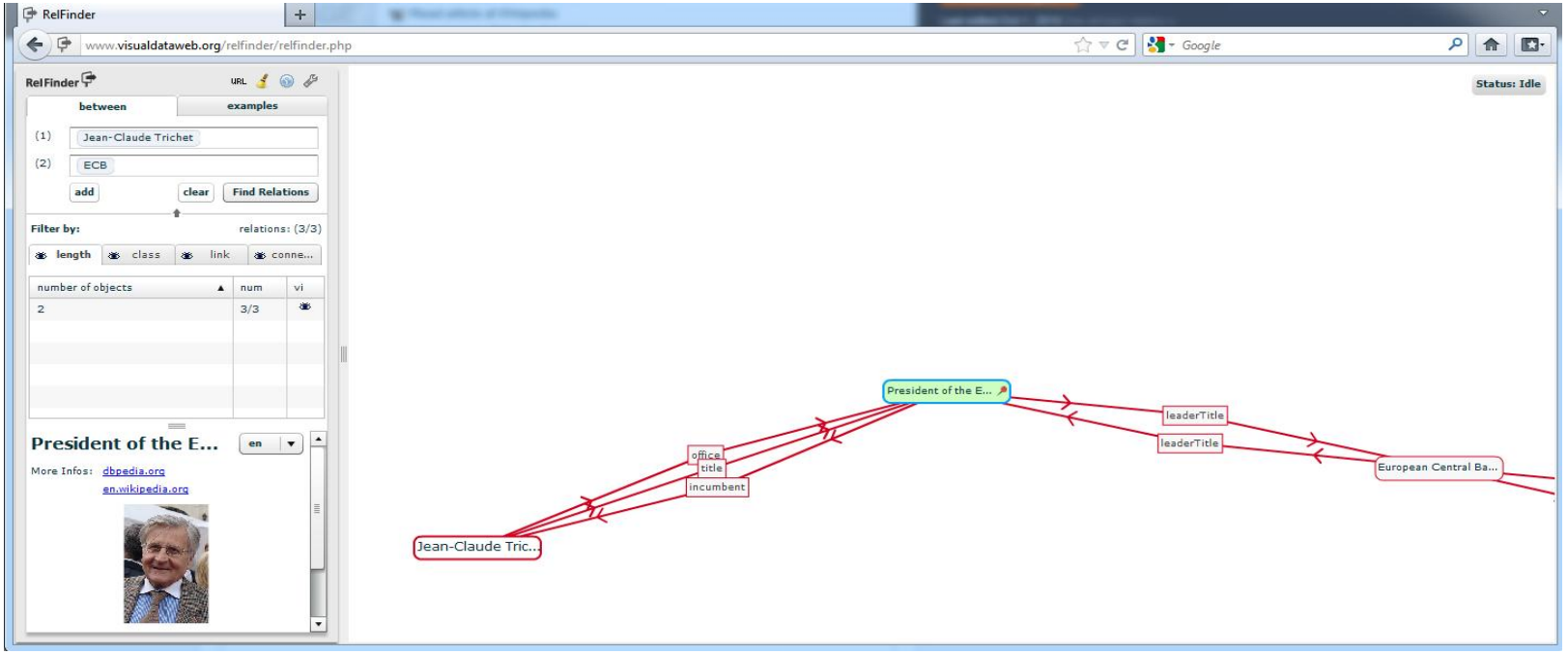
Clustering of entities via enrichment relatedness

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



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/

THANK YOU

CONTACT DETAILS

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