

# Enabling Entity-Based Aggregators for Web 2.0 Data

Ekaterini Ioannou<sup>1</sup>, Claudia Niederée<sup>1</sup>, Yannis Velegarakis<sup>2</sup> – <sup>1</sup>L3S Research Center, <sup>2</sup>University of Trento

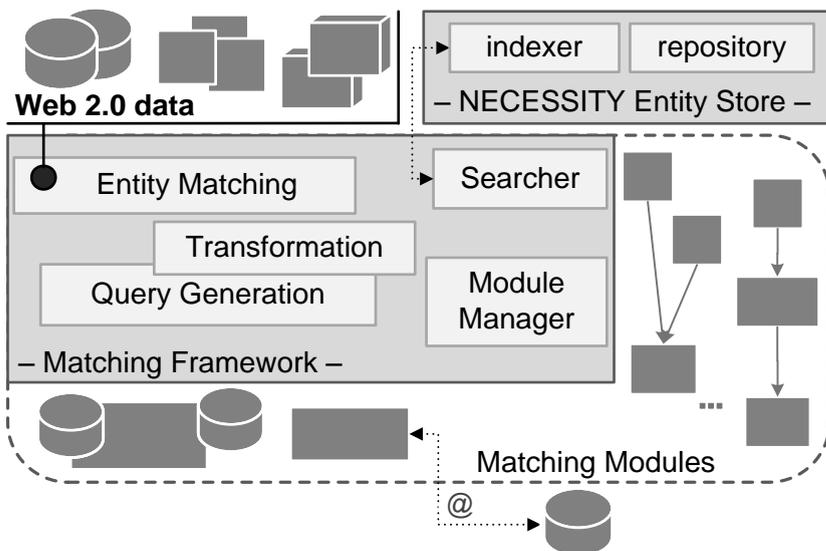
## CHALLENGES

With the growing amount of Web 2.0 data, aggregators that combine structured, semi-structured, and unstructured information from heterogeneous sources, become a promising direction for Social and Semantic Web applications. However, such Web 2.0 data with its user-generated content, its various quality levels, and especially its heterogeneity, make the required integration an extremely difficult task.

## SOLUTION

- We adopt the suggestions of the Linked Data movement and propose entity-based aggregation.
- We perform integration by identifying and linking data that refer to the same real world object (e.g., an event, a location, an organization, a person).

## ENTITY-BASED AGGREGATORS



New query requests are processed from the entity matching component:

- Request is first sent to query generation that analyzes it and generates an initial query for the entity store.
- This query is then analyzed for selecting the module that is more appropriate for evaluating the specific request. This is based on the module profiles that contain information about module description, classification, and matching capabilities.
- The selected matching module revises the generated query, and sends the revised query to the entity store.
- The store processes the query and returns a small set of candidate entities.
- These candidate entities are then returned to the module for performing matching and identifying the entity that corresponds to the given request.
- The matching results are returned as an answer to the request.

## EVALUATION

Entity store with 6.5 million entities describing people, organizations, countries, cities, roads, mountains, etc.

- (A) Queries extracted from news events with OpenCalais, e.g., *name='Carla Boni' position='singer' country='italian' medicalCondition='long illness'*  
 Entity as first answer: 81% of the queries  
 Execution time: 0.02 seconds
- (B) Queries from Web page text, e.g., *Jacques-Yves Cousteau French explorer*  
 Entity as first answer: 74% of the queries  
 Execution time: 0.03 seconds