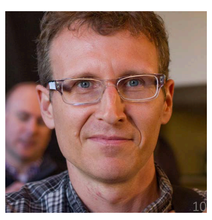
Self-Service Data Provisioning Through Semantic Enrichment of Data

Ahmad Assaf† , Aline Senart† and Raphaël Troncy‡

†SAP Research, SAP Labs France SAS
805 avenue Dr. Maurice Donat,06254 Mougins Cedex, France
first.last@sap.com

‡EURECOM
2229 route des crêtes, 06560 Sophia Antipolis, France
raphael.troncy@eurecom.fr



Fabien Martins
Business Analyst
Department of Transport - France

Fabien wants to compare the number of accidents that happened last year in France which involve consumption of illegal chemicals with those in the United States

Driving accidents chemicals

Dataset Integration and Enrichment

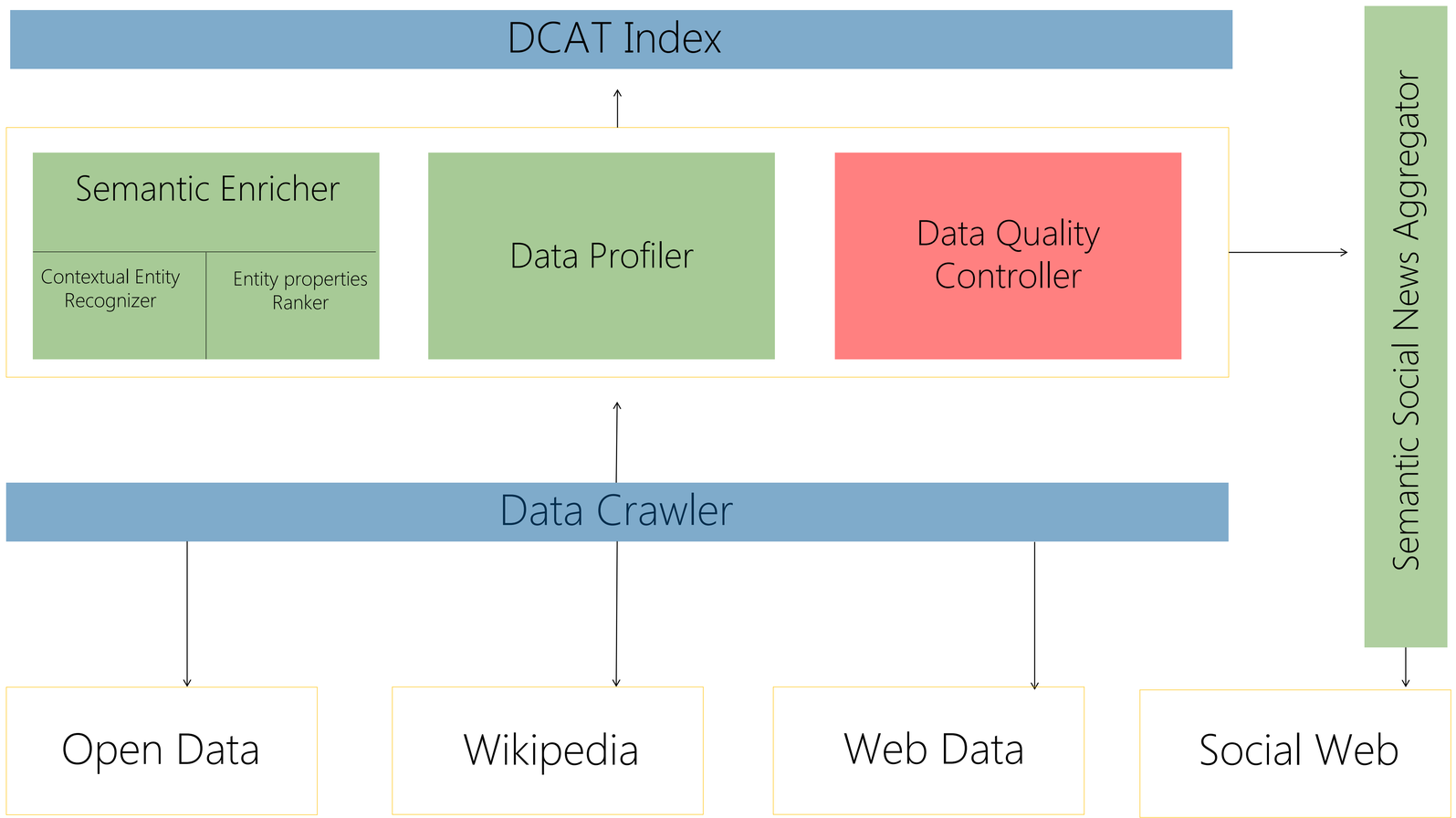
- The **Semantic Enricher** annotates datasets based on the semantics of the data at the instance level
- The **Contextual Entity Recognizer** is able to identify the most relevant type of an entity taking into account contextual information
- The **Entity Properties Ranker** is able to assess which properties of an entity are more "important" than others
- The **Data Profiler** is used to examine the data in order to understand its content and structure. It applies a set of profiling tasks that includes statistical, dependency, redundancy and uniqueness.
- The **Semantic Social News Aggregator (SNARC)** is a service that federates queries to various Social Media sources like Twitter, Google+, YouTube, etc. and reconciles the results against a semantic context

Road traffic accidents involving alcohol (% of all traffic crashes)
World Health Organization
Alcohol – Drink – Car – Transportation Method , Accident

Prescription drugs and alcohol top drugged driving accident survey, marijuana comes in last
Oct 14 2013

Did you know almost half of driving accidents occur at night? How can new technology make rain and snow disappear from your car? How well do you deal with driving in precipitation?
Dec 6 2013

Drivers who used #drugs & #alcohol were 23 times more likely to be in a fatal car accident: new study. http://bit.ly/1beYfre
Sept 26 2013



Data Quality Controller

Building on the main four principles for publishing Linked Data, we grouped the quality attributes into four main categories:

- **Quality of the entities** : quality indicators that focus on the data at the instance level (i.e. syntactic checkers)
- **Quality of the dataset**: quality indicators at the dataset level
- **Quality of the semantic model**: quality indicators that focus on the semantic models, vocabularies and ontologies
- **Quality of the linking process**: quality indicators that focus on the inbound and outbound links between datasets

Drunk driving in the United States
Wikipedia
Alcohol – United States – Drunk - Driving

Transportation Accidents by Mode
Bureau of Transportation Statistics
Transportation – Accident – Car – Bus - Motorcycle

Alcohol Beverage Sampling Program
Department of the Treasury
Alcohol – Beverage – Wine - Spirits

- The datasets descriptions are exposed using the DCAT vocabulary. This choice came from the fact that the Open Data Support is promoting the DCAT-AP (and consequently DCAT) as the standard for describing datasets and catalogs in Europe
- We need to implement mechanisms that are able to periodically check for updates and apply live edits when needed