# ENABLING SELF-SERVICE DATA PROVISIONING THROUGH SEMANTIC ENRICHMENT OF DATA

**Industrial thesis:** CIFRE/SAP, EURECOM and EDITE doctoral school
**Author**: Ahmad ASSAF
**Doctoral advisor**: Raphaël Troncy
**SAP supervisor**: Aline SENART

September 8, 2014

# 1. Introduction

This report aims to give a brief overview of what has been done during our third year of this PhD. We joined SAP Research in May 2012 in order to start a PhD thesis with SAP AG and EURECOM while being registered in the EDITE doctoral school in Paris.

Enterprises use a wide range of heterogeneous information systems in their business activities such as Enterprise Resource Planning (ERP), Customer Relationships Management (CRM) and Supply Chain Management (SCM) systems. An enterprise distributed IT landscape contains multiple systems using different technologies and data standards. In addition to this heterogeneity, the amount of information in enterprise databases and on-line data stores expands exponentially each year. Enterprise Big Data isn't big in volume only, but in the associated file formats and data structure (variety). The information is also often stored often in unstructured and unknown formats.

Data integration is the problem of combining data residing at different sources, and providing the user with a unified view of these data. In large enterprises, it is a time and resource costly task. Various approaches have been introduced to solve this integration challenge. These approaches were primarily based on XML as the data representation syntax, Web Services to provide the data exchange protocols and Service Oriented Architecture (SOA) as a holistic approach for distributed systems architecture and communication. However, it was found that these technologies are no sufficient to solve the integration problems in large enterprises. Recently, ontology-based data integration approaches have been suggested where ontologies are used to describe the data, queries and mappings between them. A slightly different approach is the use of the Linked Data paradigm for integrating enterprise data. Enterprises like Google and Microsoft are not only using the Linked Data integration paradigm for their information systems, but are also aiming at building enterprise knowledge bases (like the Google Knowledge Graph powered in part by Freebase) that will act as a crystallization point for their structured data.

Linked Open Data (LOD) movement has gained lots of momentum in the last years. From 12 datasets cataloged in 2007, the Linked Open Data has grown to 570 datasets and 2909 linkage relationships between the datasets containing more than 60 billion facts in 2014. Data is being published by both public and private sectors and covers a diverse set of domains from life sciences to military. This success lies in the cooperation between data publishers and consumers. Users are empowered to find, share and combine information in their applications easily.

Despite the legal issues surrounding Linked Data licenses, it is still considered a gold mine for organizations who are trying to leverage external data sources in order to produce more informed business decisions. The potential economic effect unfolding in education, transportation, consumer products, electricity, oil and gas, health care and consumer finance has an estimated potential to reach 3 trillion US Dollars annually.

Data becomes more useful when it is open, widely available and in shareable formats, and when advanced computing and analysis can yield from it. The quality and amount of structured knowledge available make it now feasible for companies to mine this huge amount of public data and integrate it in their next-generation enterprise information management systems. Analyzing this new type of data within the context of existing enterprise data should bring them new or more accurate business insights and allow better recognition of sales and market opportunities.

Business Intelligence (BI) has always been about creating new insight for business by converting data into meaning that can be shared between people to drive change in the organization. One key aspect of creating meaning is driving a common shared understanding of information also known as Semantics.

Classic BI and even the newer Agile Visualization tools focus much of their selling features on attractive and unique visualizations, but preparing data for those visualizations still remains the far more challenging task in most BI projects, large and small. **Self-service data provisioning aims at tackling this problem by providing intuitive datasets discovery, acquisition and integration techniques intuitively to the end user**.

In this thesis, we aim at creating a framework that will enable self-service data provisioning in the enterprise. Our goal is to provide a mechanism that annotates and profiles tabular data to provide better dataset description. Furthermore, we aim to aggregate the enhanced datasets descriptions so that people can search and browse through content. We also aim at providing ranking mechanism that leverages a comprehensive data quality metric and license information attached to the dataset description.

## 2. Third Year Activity

We continued the literature survey spanning different fields ranging from Business Intelligence, Data Analysis and Data Integration in the Semantic Web. Doing so helps in enriching our overall knowledge about these fields which helped us in defining more our research scope. In this year,

we have refined the architecture and the core contributions of this Thesis (Figure 1). The black boxes represent the contributions of this thesis.

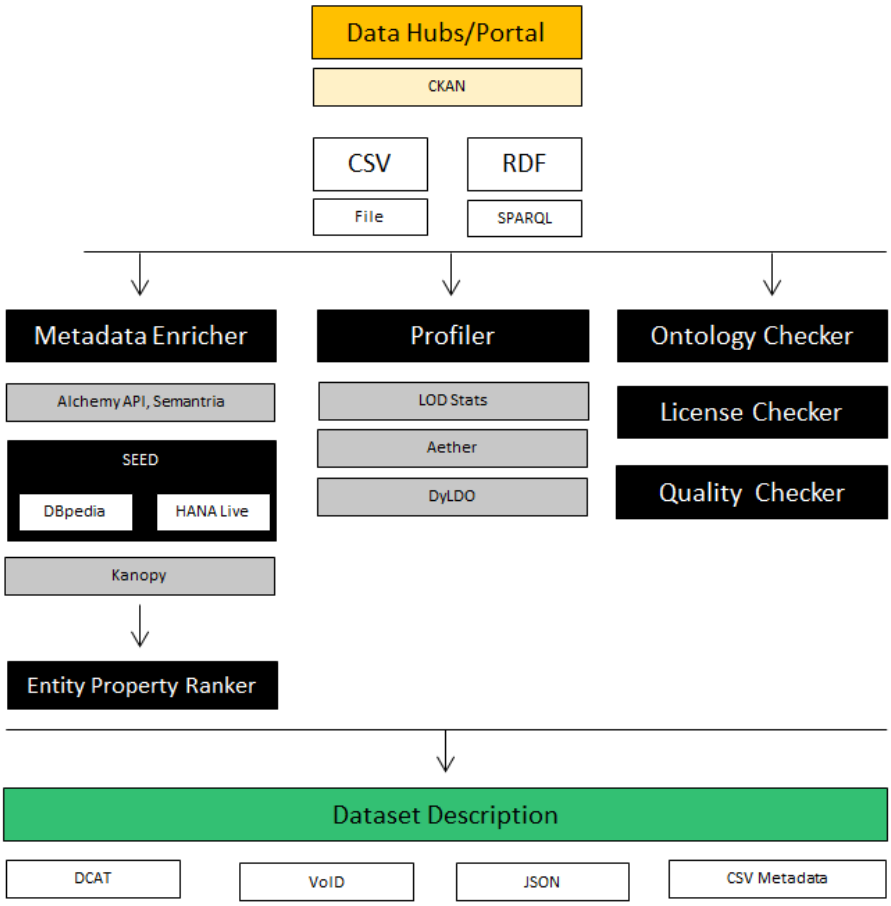


**Figure 1: Architecture of our proposed self-service data provisioning framework**

## 2.1 Metadata Enricher and SEED

We have participated in an internal project at SAP called ***Business Intelligence Graph (BIG).*** The Business Intelligence Graph is a set of foundation services for BI applications such as SAP Lumira to simplify the experience of Decision Makers and Analysts. With the BI Knowledge Graph (BIG), we harvest BI artefacts, BI usage and user profiles and store them in the HANA Graph Engine. We provide query, recommendation and ranking services as external services. BIG offers the following key features:

- A graph repository that summarizes and links reference data, cards, BI artifacts, user profile, usage and context data

- Services for feeding and maintaining the data into the repository
- Services for executing cards
- Services for querying, usage analytics, and recommendation

The technical development of BIG followed the Scrum methodology which is an iterative and incremental agile software development method for managing software projects and product or application development. I was the UI/UX lead as well as a contributor to the core functionalities of designing the data model, implementing the services in the back-end and data aggregation and feeding.

Moreover, we have created a new project inside of SAP called ***Semantic Enrichment of Enterprise Data (SEED).*** It aims at enriching data at the instance level with Semantic meta-information extracted from open knowledge bases like DBpedia. SEED implements entity ranking algorithms that take into account string similarity using HANA text search and the popularity in the DBpedia dataset via the number of incoming associations. Furthermore the algorithms are used to determine the most appropriate types for an entity. The developed entity disambiguation is further used to enhance schema matching, improve data integration by providing data cleansing functionalities and automatic domain detection for Linked Open datasets.

## 2.2 Entity Property Ranker

We have worked on a so-called Entity Property Ranker module that aims at suggesting what are the important properties for a particular entity. More precisely, we have reverse-engineered the Google Knowledge Graph in order to represent explicitly what choices Google is making when displaying facts in its Search Engine Results Panel (SERP), abstracting this to types. This work has been published as a poster at the Extended Semantic Web Conference (ESWC) 2014.

**Abstract:**
Entities play a key role in knowledge bases in general and in the Web of Data in particular. Entities are generally described with a lot of properties; this is the case for DBpedia. It is, however, difficult to assess which ones are more "important" than others for particular tasks such as visualizing the key facts of an entity or filtering out the ones which will yield better instance matching. In this paper, we perform a reverse engineering of the Google Knowledge graph panel to find out what are the most "important" properties for an entity according to Google. We compare these results with a survey we conducted on 152 users. We finally show how we can represent and explicit this knowledge using the Fresnel vocabulary.

**Reference:**

Assaf, Ahmad; Atemezing, Ghislain Auguste; Troncy, Raphaël; Cabrio, Elena. What are the important properties of an entity? Comparing users and knowledge graph point of view. ESWC 2014, 11th Extended Semantic Web Conference, May 25-29, 2014, Anissaras, Crete, Greece

## 2.3 Data Quality Checker

We have proposed "An Objective Assessment Framework for Linked Data Quality" that we have described in a paper submitted to a special issue on Data Quality in the International Journal on Semantic Web and Information Systems.

**Abstract:**

The standardization of Semantic Web technologies and specifications has resulted in a staggering volume of data being published. However, data should be of good quality to be integrated properly. In this paper, we propose an objective assessment framework for data quality that issues a certificate for a given Linked Open Data repository. This framework helps on one hand data owners to rate the quality of their datasets and get some hints on possible improvements, and on the other hand data consumers to choose their data sources from a ranked set. In a previous work, we identified potential quality issues of Linked Data and listed quality principles for all stages of data management. We refine this work here with a framework composed of objective quality indicators and associated metrics. For each indicator, we selected a set of tools and systems that can be used to rate datasets according to key quality principles. This allowed us to discover that most of the tools cover only a small subset of indicators and to identify those that are not covered at all.

# 3. Future Work

For this final year, we plan to finalize the development of one component in this architecture (Profiler), to conduct an extensive evaluation and to write the thesis manuscript.

## Profiler Component:

- Develop a CKAN/DKAN/DCAT compliant crawler, that, providing a datahub URI, enables to crawl parts (e.g. a specific group, a specific dataset) or an entire dataset catalog. The metadata descriptions will be stored in a database.
- Develop methods for selecting and sampling datasets. We limit ourselves to datasets with CSV downloadable data and for which the CSV files can be validated against a CSV lint.

- Develop methods for creating topic models of this sampled data, after an enrichment step performed using the **SEED** module.
- Expose this enriched CSV metadata following the best practices proposed by the W3C CSV on the Web Working Group so that others can re-use this annotation and enrichment steps.

## Evaluation:

- Set up the evaluation protocol by specifying the datasets used in the experiments
- Evaluate **SEED** against Freebase and DBpedia lookup for usage in annotating datasets

Dr. Raphaël Troncy