UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

eXascale Infolab

Prof. Dr. **Philippe Cudré-Mauroux**
Université de Friboug
Département d'informatique
eXascale Infolab
Bd Pérolles 90
1700 Fribourg

T +41 26 300 8332
F +41 26 300 9726
pcm@unifr.ch
www.exascale.info

Mountain View, December 1, 2015

Rapport de thèse

Enabling Self-Service Data Provisioning Through Semantic Enrichment Data (M. Ahmad ASSAF)

In his thesis, Mr. Assaf tackles the overall problem of designing an effective framework to enable data provisioning in the enterprise. The author makes a number of important contributions in this context, including contributions on data integration and enrichment, on data maintenance and discovery, and on data quality. The research questions explored by Mr. Assaf are highly relevant nowadays given the rapid extension of open datasets that are contributed online as well as the increasing importance of data integration inside companies. The contributions of Mr. Assaf provide a solid foundation for solving those problems and are relevant both from a scientific and from an application perspective.

The first chapter of Mr. Assaf's thesis gives an introduction to his work, presents the two use case scenarios used throughout the rest the thesis (a data analyst use case and a data portal administrator use case), introduces the three main research challenges tackled (on data integration and enrichment, data maintenance and discovery, and data quality) and gives an outline of the overall thesis. The rest of the thesis is organize into two distinct parts: Part I, spanning from Chapter 2 to Chapter 5, focuses on problems and applications related to dataset profiling. Part II, ranging from Chapter 6 to Chapter 8, is dedicated to the description of two prototypes in enterprise data enrichment.

Part I starts with Chapter 2, which introduces a number of important concepts used throughout the rest of the thesis, including Semantic Web standards, the notion of Linked Open Data, as well as important concepts in data profiling and data quality.

Chapter 3 is in my opinion one of the central chapters of the thesis. In this chapter, the author conducts a comprehensive survey of seven metadata models including the well-known CKAN, VoID, DCAT, and Schema.org models. The author proposes a compelling classification of the dataset models along several dimensions. He also describes in detail a series of dataset model mappings that can be used to compare the various models as well as to promote interoperable or translational services across models. Following this compelling survey, the author introduces the Harmonized Dataset modeL (HDL), which basically is a superset model built from the existing models. Mr. Assaf convincingly motivates the need for a harmonized model, and successfully introduces a rich model capable of expressing enough meta information to let any potential data consumer fully capture the given dataset. Finally, Mr. Assaf describes the benefits of leveraging such an expressive model for a number of important use cases, including for dataset discovery, search, and spam detection.

In Chapter 4, Mr. Assaf motivates the need for automated tools that can identify various issues in the metadata of published datasets as well as correct them automatically. The notion of information entropy introduced by the author in this context unclear (it is different from the usual notion of information entropy, and its correlation to data lifespan is in my opinion debatable). Subsequently, the author introduces the main software artifact he created for his thesis, namely Roomba, a scalable and automated tool for extracting, validating, correcting and generating descriptive profiles of linked datasets. Roomba takes the form of a processing pipeline that combines techniques for data identification and crawling with a set of pluggable modules for various profiling tasks. The descriptions of Roomba, of its extraction components and of its profile processing capabilities are from my perspective quite compelling. Roomba is a convincing data profiling tool that could be helpful in countless data manipulation scenarios. The author also presents the results of an extensive experimental evaluation designed by running his framework on a series of prominent data portals validating the correctness of his approach and framework.

Chapter 5 concludes the first part of this thesis by tackling the important problem of automated data quality assessment for linked datasets. The chapter starts by a critical review of the related work in this domain, highlighting a number of gaps in previous work (such as the lack of clear distinction between what can be automatically measured and what cannot). Subsequently, the author introduces a comprehensive evaluation and assessment framework for measuring the quality of linked data at the dataset level (leveraging more than 60 quality indicators), and extends his Roomba software in order to support most of the objective quality indicators introduced in his framework. The resulting approach is, to the best of my knowledge, the most comprehensive and automated approach to measure the quality of linked datasets. By running Roomba on the LOD cloud, the author makes a number of convincing arguments on the low quality of currently available linked data along several axes.

Part II, dedicated to the enrichment of enterprise data, starts in Chapter 6 with a summary of Data Integration, semantic enrichment, and Business Intelligence techniques. The author also gives a succinct description of relevant tools from the SAP ecosystem. The description of those tools is however rather superficial and lacks technical depth (for instance, HANA is introduced as a “revolutionary platform” without backing such a claim from a technical perspective). Finally, the author gives a brief overview of several social media platforms exposing data that could be relevant in the context of enterprise data enrichment.

Chapter 7 examines the need for knowledge bases in large enterprises. While the motivation behind this need is succinctly discussed (i.e., to facilitate the provision of data integration services), I would have welcome some additional explanation about this as it is rather untypical today to resort to such generic knowledge bases in companies. The author then details the challenges he faced when importing the DBPedia knowledge base into SAP HANA. The author decides to leverage a column store to import DBPedia, which is interesting, though he does not discuss the important technical tradeoffs in this context (see for example the work by Daniel Abadi or Lefteris Sidirourgos on this topic). In addition, he presents a number of interesting tools that he designed for entity disambiguation, property ranking and semantic enrichment. While these tools are interesting, I would have welcome additional technical detail in this context also, for instance on the choices made when implementing those tools (as there are many known approaches in this domain), as well as additional experimental results, for example on standard datasets or comparing the author’s approach to well known baselines. Finally, Mr. Assaf introduces a compelling schema matching approach leveraging Linked Data to map cell values onto instances that substantially improves the results of the considered baseline.

Chapter 8 is dedicated to the issue of aggregating social contents. The author succinctly presents a prototype, called SANR, that uses semantic web technologies to aggregate social news. The description of the prototype is interesting, though the author does not discuss the technical choices behind the prototype in detail (e.g., on how to create the semantic models). Also, neither the efficiency nor the effectiveness of the tool is validated empirically.

Chapter 9 concludes the manuscript by offering a qualitative analysis on how the various advances described throughout this thesis can be leveraged to solve the use cases presented in the introduction. Finally, a number of extensions and future work are discussed.

Overall, I found Mr. Assaf's dissertation compelling both from a technical and from a scientific perspective. The thesis is clear, well-written, and nicely structured, though it is composed of two fairly heterogeneous parts (probably due to the fact that the student worked in collaboration with two institutions). The state-of-the-art sections give a comprehensive yet to-the-point overview of the various domains touched by this work. The contributions of this thesis are from my perspective scientifically sound—leveraging state of the art methodologies and techniques from the Semantic Web, Database, and Information Retrieval fields—modulo the few remarks that I raised above. The main piece of software resulting from this thesis, Roomba, is a very interesting artifact that could be used in many different contexts relating to linked data exploration or processing. The author also showed the validity of his results through a successful series of publications in international workshops and conferences.

To summarize, I consider this thesis as being of interest to the research field considered, and support that this work be defended in order to grant Mr. Assaf the title of Doctor of Philosophy at Telecom ParisTech.

Prof. Dr. Philippe Cudré-Mauroux
Professeur Associé, Uni. Fribourg
Directeur, eXascale Infolab