

KEYSTONE COST Action IC1302
Short Term Scientific Mission Report

STSM Topic: Keyword Search over the Web of Things

STSM Applicant: Prof. Omar Boucelma, Aix-Marseille University (AMU), France,
omar.boucelma@univ-amu.fr

STSM Host: Prof. Francesco Guerra, University of Modena e Reggio Emilia (UNIMORE), Italy,
francesco.guerra@unimore.it

STSM Reference: ECOST-STSM-IC1302-050417-084469

STSM Dates: 07.04.2017 – 22.04.2017

Purpose of the STSM

One of the objectives of the STSM was to explore a useful combination of keyword-based search mechanisms within a Semantic WoT (SWoT) framework in order to enable the development of intelligent applications such as recommendation of things of interest or advanced data analytics.

The motivation behind this objective are twofold: first, to leverage of both KEYSTONE results and partners background for SWoT applications; second to pave the road for a midterm collaboration that could involve master students on both sides in this new exciting research topic.

Description of the work carried out during the STSM

During the first week, we worked on a specific key issue for Big Data processing, that is cleaning Map Reduce workflows. More specifically, we illustrated one of the specific topics that has been addressed during the **1st KEYSTONE training school: Keyword Search on Big Data**, that is *Provenance and Data quality*.

The goal of this work was to explore and represent through a formal model how data cleaning can improve the quality of parallel workflow systems such as MapReduce. We came up with a design of a provenance-aware quality assessment model and illustrated its applications for debugging distributed computing. A **paper resulting** from this work has been **submitted** for publication.

The second week was dedicated to the Semantic World of Things topic (SWoT). The main idea behind SWoT is to consider *Things* as first-class citizens of the semantic web. Until recently, the *Internet of Things* consisted of a set of applications built upon physical devices such as sensors, cyber-physical systems and so forth. Such devices use proprietary protocols and formats, which makes it difficult to capture the application logics at a reasonable semantic level, and even harder to ensure interoperability. After reviewing the SWoT literature, we

proposed a research agenda. The first phase of this research agenda -- which consists in building a semantic model for the WoT, is underway, since a master student at AMU is already involved and we expect to involve a UNIMORE student as well.

Description of the main results obtained

We introduced PAQA, a Provenance-Aware Quality model for quality assessment in MapReduce programs. The model extends W3C PROV-DM Provenance model with quality assessment rules. The model allows the representation of both cleaning tasks and MapReduce processes in a graph that can be used along different directions: providing a reporting tool to the final users, ensuring reproducibility, or repairing the wrong data by means of lineage.

Regarding the Semantic World of Things, our findings led us to set up a research agenda that consists of:

- Providing a semantic model for the WoT;
- Exploring a discovery/mashup approach to promote a service-oriented approach for the development of WoT applications;
- Enhancing the discovery process by means of keyword search.

Finally, apart from the technical results obtained during this STSM, the efficient networking dimension of KEYSTONE has been demonstrated: indeed, besides the STSM host and applicant, the work carried out during this STSM has also led to the involvement of students. Two post-docs (Julien Lacroix, from AMU, and Matteo Interlandi, formerly from UNIMORE and currently at UCLA) participated to the PAQA work, while another master student has been involved in the SWoT program.

Future collaboration with the host institution

We are planning to pursue the research work developed during this STSM: (1) Extend the PAQA model to allow a more generic quality assessment framework, and (2) contribute to the SWoT.

We would like also to involve more students and staff in setting un an exchange programme through ERASMUS - Higher Education Student and Staff Mobility.

Foreseen publications/articles resulting from the STSM

A paper entitled "Cleaning MapReduce Workflows" has been submitted to the 4th *International Symposium on Big Data Principles, Architectures & Applications (BDAA 2017)*, co-located with the 15th *annual International Conference on High Performance Computing & Simulation (HPCS 2017)*.