Scientific Report

First name / Family name	Riccardo Albertoni
Nationality	Italian
Name of the *Host Organisation*	Ontology Engineering Group (OEG), Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid (UPM) 28660 Boadilla del Monte, Madrid, España
First Name / family name of the *Scientific Coordinator*	Asunción Gómez Pérez
Period of the fellowship	01/03/2012 to 28/02/2013

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

The research activity carried out during this fellowship has enlarged the applicability and usability of my asymmetric context-dependent instance similarity[1](hereafter referred as instance similarity for short) when analysing linked data entities. It results in two main outcomes: (i) **SSONDE**, which is an open source framework deploying the instance similarity on linked data entities[2]; (ii) a definition of **linkset quality,** which aims at predicting data missing generated by complementing a dataset via its linksets, and eases in assessing if two interlinked datasets can be analysed by SSONDE. The SSONDE framework has been developed in collaboration with IMATI-CNR, and it is expected to attract potential users for my instance similarity, so that, future extensions of SSONDE can be driven and funded through a direct involvement in third party projects. The linkset quality is meant to estimate gains and losses when complementing a linked data dataset with its interlinked resources. It contributes to the SSONDE applicability as well as to the overall linked data conditioning practices. Below, I describe my work in accordance with the ABCDE Research Training Programme originally submitted to ERCIM.

Activity 1 - Identification of application scenarios where to exploit SSONDE: In the context of the activities and projects running at OEG, SSONDE seems particularly promising when applied on linked data provenance descriptions. At OEG, the notion of provenance is employed in different application domains. For example, it is deployed to aggregate travelers' blogs and materialize their travels (*touristic domain*) and to characterize research objects in terms of their underpinning scientific experiments (*"open research/science" domain*). In these domains, SSONDE can be deployed to compare travels, travellers, and research objects. It can be employed to complement suggestions obtained from recommendation systems, or to support end-users in sifting resources represented as linked data entities. An example of application of SSONDE in the open science domain is provided in [A].

Activity 2 – Identification of current limitations, bad practices and technological gaps in the analysis of linked data resources: two complementary approaches has been deployed to identify limitations, bad practices and technological gaps: (i) the analysis of the state of art pertaining to linked data applications, crawling tools, provenance and data quality practices; (ii) the development of a test application in a scenario related to "open research and science" domain.

The complementary approaches have pointed out that *not all the linked data architectural patterns are equally applicable when analysing data with SSONDE*. At the moment, among the existing architectural patterns, i.e., On-The-Fly Dereferencing Pattern, Query Federation Pattern and Crawling Architectural Pattern, the last is the most suitable when applying SSONDE-alike analysis tools on third parties linked datasets. This conclusion has been drawn considering that

a) On the fly dereferencing of large sets of entities is a slow process, which is even quite inefficient in term of bandwidth. Unless we are analysing a modest set of resources, such architectural pattern is not recommendable;
b) Query federation architectural pattern strongly relies on SPARQL endpoints federation which at the moment is not completely supported: (i) SPARQL syntactical constructs for federation and "follow up" queries have been only recently proposed for W3C Recommendation, thus SPARQL technology is still evolving to become fully compliant with this recommendation proposal; (ii) SPARQL endpoints available in the

---

[1] R. Albertoni, M. De Martino: Asymmetric and Context-Dependent Semantic Similarity among Ontology Instances. J. Data Semantics 10: 1-30 (2008)

[2] http://code.google.com/p/ssonde/

LOD cloud extensively differ in their stability, efficiency, and compliancy to SPARQL 1.1, thus querying SPARQL endpoints efficiently in articulate consumption scenario is still quite challenging.

c) No matter which architectural pattern you select, the harmonization of independently served linked datasets will be a critical issue. However, the crawling architectural pattern addresses the harmonization issues with off-line dedicated frameworks (e.g., LDIF). Whereas, On-The-Fly Dereferencing and Query Federation Patterns might require to deal "on the fly", which is intrinsically more difficult. Considering that the reliability of SSONDE's similarity results is strongly affected by inconsistencies in vocabulary mapping and entity consolidations, and that the resources published as linked data are generally far from being satisfyingly mapped and consolidated, applying SSONDE on "on the fly" harmonized resources is not really recommendable.

d) In general there is no guarantee that datasets integrated for a group of consumers will suit target applications of others, and current linked data quality practices have difficulties in assessing the completeness and reliability of harmonization efforts undertaken by third parties. Often the impossibility to assess at what extent the harmonization has been already reached forces in re-harmonizing the datasets almost from the scratch, so that, harmonized datasets can be finally trusted.

By the way, in a longer perspective, we would like to consider also scenarios where crawling architectural pattern can't be applied. For example, scenarios in which data is frequently updated or maintained distributed because of its size or licence. In these cases, the query federation pattern becomes an option. A remarkable amount of efforts are currently on-going to overcome technological and scientific issues in query federation, and hopefully, SPARQL endpoint federation will be soon feasible. So at the end, the most critical issue when applying SSONDE-alike analysis tool on third party interlinked datasets is the completeness and reliability of third party served harmonization. For this reason, I've decided to spend the whole Activity 4 developing the linkset quality.

Activity 3 – Consolidation of the instance similarity framework: the instance similarity has been released as SSONDE, an open source JAVA framework working in a linked data setting and released under the GNU GPL v3 Licence[2]. Besides, new modules have been developed taking into account limitations and technological gaps identified developing a test-application in the "open research/science" domain. In particular, (i) *performance optimization* has been deployed considering caching of intermediate similarity results. Cashing techniques have significantly speeded up the assessment of similarity in the "open research and science" scenario described in [A] moving the comparison of researchers from 1318 seconds to 33 seconds; (ii) the SSONDE prototype has been opened with respect to further *domain driven extensions* re-encoding the context parameterization in JSON and making external data similarity pluggable in the system; (iii) A very *preliminary interface* to visualize the similarity matrix has been arranged exploiting third parties frameworks[3]. However, the deployment of a visualization framework to sift entities according to SSONDE's similarity results requires a longer-term and dedicated effort.

Activity 4 - Development of practices and tools easing data conditioning process: contributions to the practices for linked data conditioning have been done developing the notion of linkset quality [B]. Linkset quality aims at predicting data missing generated by complementing a dataset via its linksets. In particular, linked data publishers can take advantage of the proposed quality to check if a linkset they have provided is good enough

[3] http://code.google.com/p/magic-table/

or must be improved. Linked data consumers are expected to consider this measure (a) to better understand whether they should or not rely on the harmonization offered by a linkset; (b) to have a first guess of what is the next action to complete the linkset; (c) to rank possible linkset alternatives.
A proof of concept prototype has been developed to ease the data conditioning process. At the moment, the linkset quality mainly focuses on linkset completeness estimating possible losses in completeness when fusing two interlinked datasets. This notion is going to be extended considering further "quality dimensions". For example, a definition of linkset LOD accessibility and availability is currently on-going. The inclusion of quality indicators in the existing RDF schemas describing linksets and datasets is planned as future work.
Activity 5: Dissemination of results has been carried out through short visits to ERCIM institutes and the submissions to journals, conferences and workshops listed in the following sections. No direct contributions to standardization processes have been done being the involvement of OEG in the W3C incubators no strictly related to the core activities of this fellowship.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

[A]Title: SSONDE: Semantic Similarity On liNked Data Entities.
Authors: Riccardo Albertoni, Monica De Martino.
Status: accepted/published.
Published in Metadata and Semantics Research Communications in Computer and Information Science 2012 (MTSR), Communications in Computer and Information Science. ISSN: 1865-0929 (Print) 1865-0937 (Online), pp 25-36.
Abstract: The paper illustrates SSONDE, a framework to assess semantic similarity on linked data entities. It describes the framework architecture, its design assumptions and its configuration functionalities. SSONDE relies on an instance similarity in which asymmetricity and context dependence are specifically conceived to compare linked data resources according to their metadata. Two different applications to consume linked datasets are illustrated showing SSONDE as a building block technology to sift linked data resources.
[B]Title: Assessing Linkset Quality for Complementing Third-Party Datasets.
Authors: Riccardo Albertoni, Asunción Gómez Pérez.
Status: accepted/to appear.
Accepted in the Third International Workshop On Linked Web Data Management (LWDM 2013), it will appear in the Joint EDBT/ICDT '13 Workshop Proceedings, March 18 - 22 2013, Genoa, Italy, ACM Press.
Abstract Linked data best practices are getting extremely popular: various companies and public institutions have started taking advantage of linked data principles for exposing their datasets, and for relating their datasets to those served by third parties.
Such enthusiasm is due to the linked data promise of evolving into a Global Data Space.
Linksets are sets of links relating datasets and they surely play a fundamental role in this promise. However, a stable and well-accepted notion of linkset quality has not been yet defined. This paper contributes to overcome this lack by proposing a linkset quality measure. Among the different quality dimensions that can be addressed, the proposed measure focuses on completeness. The paper formally defines novel scoring functions and proposes an interpretation of these functions when maintaining and complementing third party datasets.
[C]Title: EARTh: an Environmental Application Reference Thesaurus in the Linked

Open Data Cloud.
Authors: Riccardo Albertoni, Monica De Martino, Sabina Di Franco, Valentina De Santis, Paolo Plini.
Status: pending
Submitted to Semantic Web Journal – Interoperability, Usability, Applicability
Abstract: The paper aims at providing a description of EARTh, the Environmental Application Reference Thesaurus. EARTh represents a common general terminology for the environment, which has been published as a SKOS dataset in the Linked Open Data cloud. It promises to become a core tool for indexing and discovery environmental resources by refining and extending GEMET, which is considered the de facto standard when speaking of general-purpose thesaurus for the environmental domain in Europe. The paper illustrates the main key characteristics of EARTh as a guide to its usage. It clarifies (i) the methodology adopted to define the EARTh content; (ii) the design and technological choices made publishing EARTh as Linked Data; (iii) the information pertaining to its access and maintenance. Descriptions of EARTh applications and future relevance are also highlighted.

III – ATTENDED SEMINARS, WORKSHOPS, CONFERENCES

- Adaptive Semantic Data Management Techniques for Federations of Endpoints, a course tough by Maria Esther Vidal (Universidad Simón Bolívar), 4-8 June 2012, Polytechnic University of Madrid, Spain;
- 6th Metadata and Semantics Research Conference (MTSR 2012), 28- 30 November 2012, Cadiz, Spain;
- Big Data Spain 2012, 16 November 2012, Madrid, Spain;
- Ninth Semantic Summer School on Ontological Engineering and Semantic Web school, 10 July 2012, Cercedilla, Spain;
- 1st OEG RDFa collaborative tripleton (ORCO), 8th February, Polytechnic University of Madrid, Spain.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

From 14th October to 23rd October 2012: Short visit at Exmo Group led by Prof. Jérôme Euzenat, INRIA Grenoble Rhône-Alpes, Montbonnot, Grenoble, France. During my visit I presented the seminar "SSONDE Semantic Similarity On liNked Data Entities" and I discussed about possible extensions of SSONDE to include some of the data similarity functions provided by OntoSim. Besides, I discussed with Jérôme Euzenat about the linkset quality I have designed during my ERCIM fellowship and possible future collaborations.
From 28th January to 1st February 2013: Short visit at the Information System Laboratory of the Institute of Computer Science at Foundation for Research and Technology-Hellas (FORTH), Heraklion, Creete, Grece. During my visit I presented the seminar "SSONDE Semantic Similarity On liNked Data Entities", and I discussed my research activities with some of the researchers working at FORTH. Among the others, I met Irini Fundulaki, Dimitris Plexousakis, Kostas Stefanidis, Pavlos Fafalios, Martin Doerr, Giorgos Flouris, Carlo Alocca, Evangelia Daskalaki. With Yannis Tzitzikas, who was my host at FORTH, I had the occasion to talk about my instance similarity, the linkset quality and possible collaborations in running and future the activities at FORTH.