Introduction to the Special Issue on
Semantic Web Data Management

Roberto De Virgilio[1], Fausto Giunchiglia[2], Francesco Guerra[3], Letizia Tanca[4], Yannis Velegrakis[2]

[1] *Dipartimento Informatica e Automazione, Università Roma Tre, Rome, Italy*
[2] *Department of Information Engineering and Computer Science, University of Trento, Italy*
[3] *Dipartimento di Economia Aziendale, Università di Modena e Reggio Emilia, Modena, Italy*
[4] *Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy*
dvr@dia.uniroma3.it, fausto@dit.unitn.it, francesco.guerra@unimore.it, tanca@elet.polimi.it, velgias@disi.unitn.eu

Keywords: RDF, Data Management, Conceptual Modeling, Semantic Web.

1. Introduction

During the last decade we have witnessed a tremendous increase in the amount of data that is available on the Web in almost every field of human activity. Financial information, weather reports, news feeds, product information, and geographical maps are only few examples of such data, all intended to be consumed by the millions of users surfing the Web. The advent of Web 2.0 applications, such as Wikis, social networking sites and mashups have brought new forms of data and have radically changed the nature of modern Web. They have transformed the Web from a publishing-only environment into a vibrant place for information exchange. Web users are no longer plain data consumers but have become active data producers and data dissemination agents, contributing further to the increase of the information plethora on the Web.

For the successful discovery, sharing, distribution and organization of this information, the ability to understand and manage the semantics of the data is of paramount importance. This need gave birth to the Semantic Web vision that aims at the creation of a well-defined, reusable and machine-understandable form of the semantics of the Web data. The exponential growth of these data, however, made clear the need for contributions from two more disciplines: Conceptual Modeling and Data Management, research fields that have been around for more than three decades, are mature enough and can offer high-quality research results alongside efficient implementations for real-world application scenarios. Conceptual modeling deals with the development of advanced techniques and tools that allow accurate representation of, and reasoning on, artifacts that model real-world objects and information. Data management, on the other hand, deals with the development of techniques for the efficient and effective storage, querying, retrieval and management of large amounts of information of different nature, i.e., relational data, XML documents, blogs, social networking data, multimedia items, etc. While developments in Semantic Web, Conceptual Modeling and Data Management have traditionally followed independent paths, we claim that only through interdisciplinary synergies can advanced and efficient data management techniques for the Web be achieved. This special issue collects contributions that span the three areas and promote the implementation of the Semantic Web vision.

One of the fundamental challenges of Web information processing is the automated identification of related entities. Given the variety of knowledge available on the Web for the users' business or personal needs, spanning from fast, short, ready-to-consume news or posts to well-structured, formal ontology instances in the Semantic Web, users require fast access focussed to all important information about given entities, including, besides entity properties, also related events, people, situations and similar. This problem has at least two major facets: the first is *entity matching*, that is, identifying the concepts on the Web that, though having different names or abstractions, actually refer to the same real-world entity; the second is *entity-centric resource gathering*, i.e. collecting, around an entity, all the Web resources that are related to it.

The problem of named entity matching resembles the entity resolution problem of relational databases, but it is more difficult since, in the Web case, the information is often partial or incomplete. In the paper *Quality-aware similarity assessment for entity matching in Web data* Surender Yerva, Zoltan Miklos and Karl Aberer propose some strategies to design combined similarity functions that capture the degree of belief about the equivalence of two entities. The method relies on the combination of multiple evidences, with the help of estimated quality of the individual similarity values and with particular attention to missing information, which is common in the Web context. The technique is satisfactorily tested on the two cases of person-name disambiguation and Twitter-message classification.

The paper *Structured Data Clouding across Multiple Webs*, by Silvana Castano, Alfio Ferrara and Stefano Montanelli, introduces the notion of inCloud (information Cloud), a collection of related web resources built for a target entity of interest by distinguishing, also in a visual way, how prominent each retrieved web resource is with respect to the given entity, and by organizing web resources according to their mutual levels of closeness. In this paper the authors also discuss and evaluate an application of the inCloud approach to real web resources about movies.

The next fundamental problem of Semantic Web data is effective and efficient data access. In the database realm this problem is commonly addressed by research on query-answering, while the web is traditional domain of information retrieval. Within the Semantic Web, techniques from the two disciplines are blended with semantic techniques enabling surprisingly effective approaches.

In *Ontological Query Answering under Expressive Entity-Relationship Schemata*, Andrea Cali, Georg Gottlob and Andreas Pieris take a step forward in bridging the gap between conceptual modeling and semantic reasoning, by addressing the problem of answering conjunctive queries under constraints representing schemata expressed in ER+, an extension of the Entity-Relationship model. ER+ grants the expression of a variety of constraints like functional and mandatory pareticipation constraints on relationshiops, and is-a hierarchies among mentities or relationships. After defining the notion of ER+ schema separability, the paper gives a syntactic characterization of separable ER+ schemata, analyzes the complexity of the conjunctive query answering problem under separable ER+ schemata, and show that the addition of negative constraints does not increase the complexity of query answering.

The paper *PEST: Fast Approximate Keyword Search in Semantic Data using Eigenvector-based Term Propagation* by Klara Weiand, Fabian Kneissl, Wojciech Lobacz, Tim Furche and Franois Bry presents PEST, a novel approach to the approximate querying of RDF graph-structured data that propagates term weights between data items which are connected in the data structure. Besides the graph structure, also the knlowledge of the application semantics – like wiki page closeness or links between persons in social networks– is exploited to build the PEST matrix, a generalization of the Google Matrix used in PageRank. The paper concludes with extensive experiments including a user study on a real life wiki, showing the improvement over several other ranking approaches, while not loosing in performance.

The paper *An Ontology-Based Retrieval System Using Semantic Indexing* by Soner Kara, Ozgur Alan, Orkunt Sabuncu, Samet Akpinar, Nihan Cicekli and Ferda Alpaslan introduces a system for ontology-based information extraction and its application to the soccer domain. The authors propose keyword-based semantic retrieval, where performance is improved considerably using domain-specific information extraction, inferencing and rules, and semantic indexing is used to solve simple structural ambiguities and to improve scalability. The accent in this paper is on the three key issues of usability, scalability and performance of the retrieval system, and a detailed evaluation shows the performance gain due to domain-specific information extraction and inferencing.

Finally, since the amount of available Semantic Web data has grown significantly over the last few years, the need for testing scalability and performance of Semantic Web systems is growing as well. However current benchmarking either address schema-less RDF datasets or rely on fixed RDFS schemas. A step forward is takeniIn *PoweRGen: A Power-law Based Generator of RDFS Schemas*, where Yannis Theoharis, George Georgakopoulos and Vassilis Christophides present the first RDFS schema generator, which takes into account the features exhibited by real SW schemas. PoweRGen generates synthetic schemas which adhere to morphological characteristics like the subsumption hierarchy depth or structure and respect the power-law functions given as input with an accuracy around 95%. The power-law functions concern the combined in- and out-degree distribution of the property graph and the out-degree distribution of the transitive closure of the subsumption graph.

We consider with particular pride this special issue, where the three main issues of the Semantic Web data challenge, namely conceptual modeling, information access and performance, are addressed with accuracy and competence, and hope that our readers will enjoy, and take advantage of, their reading like we did while preparing it.

2. Editors' Information

Roberto De Virgilio is a Research Assistant at the University of "ROMA TRE" under the supervision of Prof. Riccardo Torlone. His PhD referred to Adaptation of Web Information Systems respect to the context of a client. The last years his research focuses on Semantic Web Information Management at different levels of abstraction. He is the author of several papers on Web Information Systems management and Semantic Web published in international journals and conferences, a chapter of the book "*Mobile Information Systems: Infrastructure and Design for Adaptivity, and Flexibility*" edited by Springer and editor of a Springer-Verlag book on Semantic Data Management which will be published by December 2009.

His research interests are: Data integration, adaptive information systems, Web based information systems, XML data management, Semantic Web, data modeling and database design, model management, Web Information Systems, Personalization.

Fausto Giunchiglia, ECCAI Fellow, is a Professor of Computer Science. He has done research in various related areas including knowledge management, data and knowledge representation, reasoning with context and formal methods. He has been program or conference chair various events, including: IJCAI 2005, Context 2003, AOSE 2002, Coopis 2001, KR&R 2000. He has been editor or editorial board member of around ten journals, including: Journal of Autonomous Agents and Multi-agent Systems, Journal of applied non Classical Logics, Journal of Software Tools for Technology Transfer, Journal of Artificial Intelligence Research. He has been Member of the ECCAI Fellows Selection Committee, of the IJCAI Board of Trustees member (01-11), President of IJCAI (05-07), President of KR, Inc. (02-04), Advisory Board member of KR, Inc., Steering Committee member of the CONTEXT conference. Relevant to the topic of this proposal, he has developed influential work on context,, ontology (semantic) matching, and data and knowledge management, and lately also on semantic Web and Web 2.0 related issues (e.g., lightweight ontologies).

Francesco Guerra received his PhD in information engineering at the University of Modena and Reggio Emilia in 2003. Since November 2005, he has been an assistant professor of Computer Engineering at the Faculty of Economics at the University of Modena and Reggio Emilia, where he teaches computer science and enterprise information systems. His main research interests include integration of heterogeneous information sources, ontologies, and the Semantic Web. He is part of the program committee of international conferences and workshops, he was co-chair of the SWAE workshop (Semantic Web Architectures For Enterprises); he was reviewer for the Information Sciences Journal (ELSEVIER), for the Electronic Commerce Research and Applications journal (ELSEVIER), for the IEEE/ACM Transactions on Computational Biology and Bioinformatics, for the International Journal of Metadata, Semantics and Ontologies (IJMSO), and he is guest editor of the special issue on Search using Metadata, Semantic, and Ontologies of the IJMSO journal. He will be guest editor of the special issue on Information Overload of the IEEE Internet Computing Review (2010).

Letizia Tanca is a Full Professor of Technologies for Information Systems and of Database Systems and an author of e than 100 papers published in international journal and conference proceedings, on databases and database theory, author of the book "Logic Programming and Databases", coauthored with S. Ceri and G. Gottlob, and editor of a Springer-Verlag book on Semantic Data Management published in December 2009. She has been involved in, and has coordinated, European and Italian research projects. Her research interests range over all database theory, especially on deductive, active and object oriented databases, graph-based languages for semistructured data. Her most recent research interests include context-aware information personalization, pervasive and mobile information systems. Letizia Tanca has been a referee of several international journals, and a member of the program committee of a large number of international conferences, and is a member of the board of the Informatics Europe association.

Yannis Velegrakis is a faculty member of the Department of Information Engineering and Computer Science of the University of Trento. His research area of expertise includes tool building for schema and ontology mapping, interoperability, data translation, information integration, data exchange, view updates, view maintenance and meta-data management. Prior to joining the University of Trento, he held a researcher position at AT&T Research Labs in the US. He has also spent time as a visitor at the University of California, Santa-Cruz, the IBM Almaden Research Center, and the Center of Advanced Studies of the IBM Toronto Lab. He was a member of the committee for the CIMI cultural profile of the ANSI/NISO Z39.50 standard. He has served in program committees of many national and international conferences, has been a reviewer for numerous international journals and he is general chair for VLDB 2013. He holds 3 US patents and has been a Marie Currie fellow.