LIDER: Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe

Guadaplupe Aguado de Cea[1], Paul Buitelaar[2], Philipp Cimiano[3], Maud Ehrmann[6], Asunción Gómez Pérez[1], Jorge Gracia[1], Sebastian Hellmann[4], David Lewis[5], John McCrae[3], Roberto Navigli[6], Felix Sasaki[7], Christina Unger[3]

[1] Ontology Engineering Group, Universidad Politécnica de Madrid
[2] INSIGHT, National University of Ireland, Galway,
[3] Semantic Computing Group, University of Bielefeld
[4] Institut für Angewandte Informatik, University of Leipzig
[5] Trinity College Dublin
[6] Sapienza University of Rome
[7] German Research Center for Artificial Intelligence (DFKI) & W3C/ERCIM

Abstract. LIDER is a support action (SA) funded by the European Commission under FP7. Its goal is to define best practices and a reference architecture, develop a roadmap as well to create a community supporting the release of language resources as Linguistic Linked Open Data, fostering the discoverability of such datasets as well as their exploitation by linked-data-aware natural language processing and content analytics services. LIDER is seeking to network with projects in the area of Natural Language Processing, Linked Data and Big Data Analytics.

1 Motivation

In the Big Data era, content analytics methods allow to discover relevant knowledge in and make sense of the continuously growing amount of unstructured data that exists on the Web as well as within organizations, institutions and companies for the purpose of improved decision making. Unstructured content can be textual or in some other form including voice, images or video, and may comprise of content in different natural languages. While there is an explosive growth in the amount of unstructured content available on the Web, the effective management, discovery and exploitation of this content still remains a challenge for companies and researchers alike. The exploitation of content at a Web scale in content analytics processes is hampered by a number of barriers including i) organizational barriers, ii) language barriers, iii) data format and modality barriers as well as iv) economic barriers caused by the fact that many datasets are not freely available, making their exploitation difficult for SMEs. In particular language resources still remain in data silos with the result that their exploitation in content analytics processes is far from straightforward.

Contributing to overcoming these barriers, LIDER promotes the availability of free, open and interoperable (FOI) language resources and technologies to lower the entry cost for exploitation of available data in content analytics processes. In order to make language resources including annotated corpora but also bilingual lexica or terminologies better discoverable and exploitable by NLP and content analytics services, clear guidelines on how to publish, discover and exploit such resources are urgently needed. This will support the development of NLP services leveraging on Linguistic Linked Open Data (LLOD) (see [1] for the benefits of representing language resources as Linked Data).

2 Objectives and Outcomes

The objectives of LIDER are to:

- **build an open and sustainable community** of stakeholders in industry, research and standards, interested in the use of free, interlinked, and semantically interoperable linguistic ressoruces for content analytics; these participants will guide the project with use cases and requirements.
- **develop a set of guidelines and best practices** supporting the publication, easier discovery and exploitation of linguistic linked data. Special attention will be paid to the availability of licensing and provenance information [2]. As proof-of-concept, LIDER will transform a substantial number of language resources into Linked Data following these guidelines.
- **provide a reference architecture** for LLOD-based content analytics services for common tasks, developing a set of services as proof-of-concept.
- **define a roadmap for LLOD-based content analytics in enterprises:** unifying requirements across stakeholders from industry (with a special focus on SMEs), research and standardization bodies such as W3C.

3 Key Project Data

The LIDER Project is a support action (SA) funded by the European Commission under FP7. The project started in November 2013 and will run for two years. The partners of the project are: Universidad Politécnica de Madrid (coordinator), Trinity College Dublin (vice coordinator), DFKI, National University of Ireland, Galway, University of Leipzig, University of Bielefeld, Sapienza University of Rome, W3C/ERCIM. You will find more information about the LIDER project at http://www.lider-project.eu/ as well as on Twitter via hashtag `#lider-project`.

References

1. C. Chiarcos, J. McCrae, P. Cimiano, C. Fellbaum, "Towards Open Data in Linguistics, Lexical Linked Data, In: A. Oltramari, P. Vossen, P. Quin and E. Hovy (eds.) New Trends of Research in Ontologies and Lexical Resources, pp 725, Springer, 2013

2. D. Vila Suero, V. Rodríguez Doncel, A. Gómez Pérez, P. Cimiano, John P. McCrae, G. Aguado de Cea, 3LD: Towards high quality, industry-read linguistic Linked Licensed Data, Proceedings of the European Data Forum (EDF), 2014.