This document is part of the Coordination and Support Action CRACKER. This project has received funding from the European Union's Horizon 2020 program for ICT through grant agreement no.: 645357.

Deliverable D3.5

Coordination with and support of LIDER

Authors: Penny Labropoulou, Stelios Piperidis, Katerina Gkirtzou

Dissemination Level: Public

Date: 30 June 2016

Status: Final

Grant agreement no.	645357
Project acronym	CRACKER
Project full title	Cracking the Language Barrier
Type of action	Coordination and Support Action
Coordinator	Dr. Georg Rehm (DFKI)
Start date, duration	1 January 2015, 36 months
Dissemination level	Public
Contractual date of delivery	30/06/2016
Actual date of delivery	30/06/2016
Deliverable number	D3.5
Deliverable title	Coordination with and support of LIDER
Type	Report
Status and version	Final
Number of pages	7
Contributing partners	ELDA
WP leader	ATHENA RC
Task leader	ATHENA RC
Authors	Penny Labropoulou (ATH), Stelios Piperidis (ATH), Katerina Gkirtzou (ATH)
Internal reviewer	Georg Rehm
EC project officer	Pierre-Paul Sondag
The partners in CRACKER are:	• Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany • Charles University in Prague (CUNI), Czech Republic • Evaluations and Language Resources Distribution Agency (ELDA), France • Fondazione Bruno Kessler (FBK), Italy • Athena Research and Innovation Center in Information, Communication and Knowledge Technologies (ATHENA RC), Greece • University of Edinburgh (UEDIN), UK • University of Sheffield (USFD), UK

For copies of reports, updates on project activities, and other CRACKER-related information, contact:

DFKI GmbH
CRACKER
Dr. Georg Rehm
Alt-Moabit 91c
D-10559 Berlin, Germany

georg.rehm@dfki.de
Phone: +49 (0)30 23895-1833
Fax: +49 (0)30 23895-1810

Copies of reports and other material can also be accessed via http://cracker-project.eu.

Contents

1 Introduction

One of the objectives of the CRACKER project is to coordinate and support resource sharing activities underpinning high-quality multilingual technology research and development, and to build bridges with concurrently running activities preparing multilingual digital service infrastructure(s).

As part of this action line, task 3.3 focuses on LIDER. The LIDER project (http://www.lider-project.eu), a 2-year project that ended on 31/12/2015, aimed to provide the basis for the creation of a Linguistic Linked Data (LLD) cloud that can support content analytics tasks of unstructured multilingual cross-media content. More specifically, the purpose of this task is to coordinate with, provide feedback to and support LIDER in rendering the META-SHARE metadata schema into RDF and developing the underlying OWL ontology, as a metamodel to be used for the documentation of Language Resources that can populate the LLD cloud.

This deliverable reports on the activities that have taken place and their outcomes.

2 Activities

The core object of the collaboration with the LIDER project has been the conversion of the META-SHARE metadata model into an RDF ontology and the subsequent conversion of the metadata records contained in the META-SHARE inventory (www.meta-share.org) into a compatible RDF format.

The collaboration has been established in the framework of the W3C (World Wide Web Consortium) Community Group “Linguistic Data for Language Technology” (ld4lt, https://www.w3.org/community/ld4lt/). Under the umbrella of the W3C, the ld4lt group aims to consult with current and potential users of linguistic data to assemble use cases and requirements for Language Technology applications that use Linked Data. A wiki dedicated to the META-SHARE/OWL metamodel (https://www.w3.org/community/ld4lt/wiki/Meta-Share_OWL_metamodel) was created for this purpose and regular conference calls took place between experts of both communities.

In addition, the *1st Summer Datathon on Linguistic Linked Open Data* (SD-LLOD-15, http://datathon.lider-project.eu/) has served as an opportunity to build upon and improve the META-SHARE/OWL ontology and combine it with the LRE-Map that includes user-provided descriptions of Language Resources presented and discussed in LRE conference articles (http://www.resourcebook.eu).

3 META-SHARE/OWL metamodel

The META-SHARE/OWL ontology has been the object of two consecutive efforts:

- initially, an ontology[1] was built for a subset of the META-SHARE schema, focusing on text corpora [4]
- this ontology has been extended to cover the whole schema; in this process, changes have also been made on the basis of theoretical and technical considerations, in order to better accommodate the RDF format.

The outcome of these efforts, the OWL version of META-SHARE ("MS-OWL") is published at: http://purl.org/net/def/metashare.

In this process, we have considered the following issues:

- W3C recommendations for RDF vocabularies,
- linking to or adopting relevant widely used RDF vocabularies for similar concepts, including
    - the Data Catalogue Vocabulary (DCAT) for representing catalogues of datasets,
    - the PROV ontology for modelling provenance information,
    - the Rights Expression Languages, such as CC-REL and ODRL, for describing copyright and licensing information in RDF,
    - the bibo ontology for describing publications
    - the lexvo ontology for the representation of languages;
- adhering to a well-designed and principled conversion of the XSD elements and values into the appropriate RDF entities, classes and properties;
- modelling considerations arising from the transformation of a hierarchical structure as used by META-SHARE into the flat structure of RDF, which resulted in changes in the naming conventions of the metadata elements, in the attachment of the properties, etc.

The ontology and the theoretical discussion about the RDFization process have been presented at two international workshops [2, 3].

Moreover, the licensing component of the model has received special attention and has been re-structured in order to take full advantage of Rights Expression Languages that are used for representing licences in machine-readable format. As a result, the licensing component has also been published as a separate module at: http://purl.org/NET/ms-rights ("MS-Rights vocabulary"). The MS-Rights vocabulary extends the CC-REL and ODRL vocabularies addressing the requirements set for the distribution of Language Resources.
A set of licenses, including the META-SHARE ones, has been encoded in RDF according to this vocabulary and can be accessed at: http://rdflicense.appspot.com/.

4 Conversion of META-SHARE metadata records into RDF

As a proof of concept, a subset of metadata records exported into XML from META-SHARE has been converted by the LIDER consortium into RDF triples and fed into linghub (http://linghub.lider-project.eu/). Linghub has been designed as a portal

[1] The first version of this ontology is available at: https://raw.githubusercontent.com/martavillegas/metadata/master/MetaShare.ttl

facilitating the discovery of language resources; it collects metadata from other sources as well (namely CLARIN VLO, lre-map and datahub) and makes them available under a common RDF schema [1].

5 MS/OWL and LRE-Map

The harmonisation of the META-SHARE and the LRE-Map metadata schemas has been initiated as a small-scale research project during the *1st Datathon on Linguistic Linked Data*.

The aim was to reconcile the two different schemas, import both into a common ontology, convert the metadata records of the two sources into RDF and try to deduplicate and merge common records. The scope of the project was intriguing given the differences of the two schemas in design principles (a strict vs. a more relaxed metadata schema) and implementation (XSD format vs. relational database).

The work that was accomplished consisted in mapping the upper nodes of the two ontologies into general theoretically motivated categories of language resources, creating a new ontology for these using Protégé, importing a subset of metadata records from both sources in the new ontology and making SPARQL queries on them[2]. The outcome of this work has not been published yet as it has not been finalised.

6 Current status

Although the LIDER project has officially ended, we intend to pursue the work that has already started towards providing META-SHARE records as linked data. Currently, we are working on improving the MS-OWL ontology (e.g., improvement of definitions, consistency checking, clearing up of the ontology in depth etc.). The ontology also needs to be updated to the newest version of the META-SHARE schema, as delivered for the CRACKER project, and complemented with supplementary material, such as an online tutorial web page, with extensive examples and rich definitions for human consumption, when the ontology is requested by a typical Web browser [5]. Once the improvements are finalised, we intend to move the ontology to a META-SHARE-owned namespace, so that it can be better curated.

We are also investigating ways of exporting META-SHARE metadata records in RDF format through a dedicated SPARQL endpoint.

[2] https://drive.google.com/file/d/0BwvuzIAhamr9cFdVTWJkUTVnbEk/view

7 References

[1] McCrae, John P., P. Cimiano , V. Rodriguez-Doncel, D. Vila-Suero, J. Gracia, L. Matteis, R. Navigli, A. Abele, G. Vulcu & P. Buitelaar (2015), "Reconciling Heterogeneous Descriptions of Language Resources". *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications, ACL-IJCNLP*, July 2015, Beijing, China. http://www.aclweb.org/anthology/W/W15/W15-4205.pdf

[2] McCrae, John P., P. Labropoulou, J. Gracia, M. Villegas, V. Rodriguez-Doncel & P. Cimiano (2015) " One Ontology to Bind Them All: The META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web". In Gandon, Fabien, C. Gueret, S. Villata, J. Breslin, C. Faron-Zucker, A. Zimmermann (eds.) *The Semantic Web: ESWC 2015 Satellite Events*, Portoroz, Slovenia, May 31 -- June 4, 2015, Revised Selected Papers http://dx.doi.org/10.1007/978-3-319-25639-9_42

[3] Rodriguez-Doncel, V. and P. Labropoulou (2015) "Digital Representation of Rights for Language Resources". *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications, ACL-IJCNLP*, July 2015, Beijing, China. http://www.aclweb.org/anthology/W15-4206

[4] Villegas, M., M. Melero, N. Bel (2015) "Metadata as Linked Open Data: mapping disparate XML metadata registries into one RDF/OWL registry". *Proceedings of the Language Resources Evaluation Conference, LREC 2014*: 393-400

[5] W3C (2008), "Best Practice Recipes for Publishing RDF Vocabularies", 28 August 2008. https://www.w3.org/TR/swbp-vocab-pub/#negotiation