LIDER: FP7 – 610782

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

REPORT OF THE 1ST SUMMER DATATHON ON LINGUISTIC LINKED DATA (SD-LLOD-15)

Table of Contents

1 Introduction

The 1st Summer Datathon on Linguistic Linked Data (SD-LLOD-15) took place in Cercedilla (Madrid, Spain) from 15 to 19 June 2015, organized by the LIDER project (http://www.lider-project.eu/). It was directed by Jorge Gracia from Universidad Politécnica de Madrid (Spain) and John McCrae from Bielefeld University (Germany). The main goal of the datathon was to offer persons from the industry and academia practical knowledge in the field of linked data applied to linguistics, with the final aim of allowing attendees to transform their own (or other's) linguistic data and publish it as linked data on the Web.

This datathon was the first organized on this topic worldwide. Around 65 professionals (including 44 attendees, speakers and tutors) met in the event from all around the world. The datathon was an invaluable forum not only for learning but also for the exchange of experiences and ideas related to linguistic linked data. More information can be found at http://datathon.lider-project.eu/

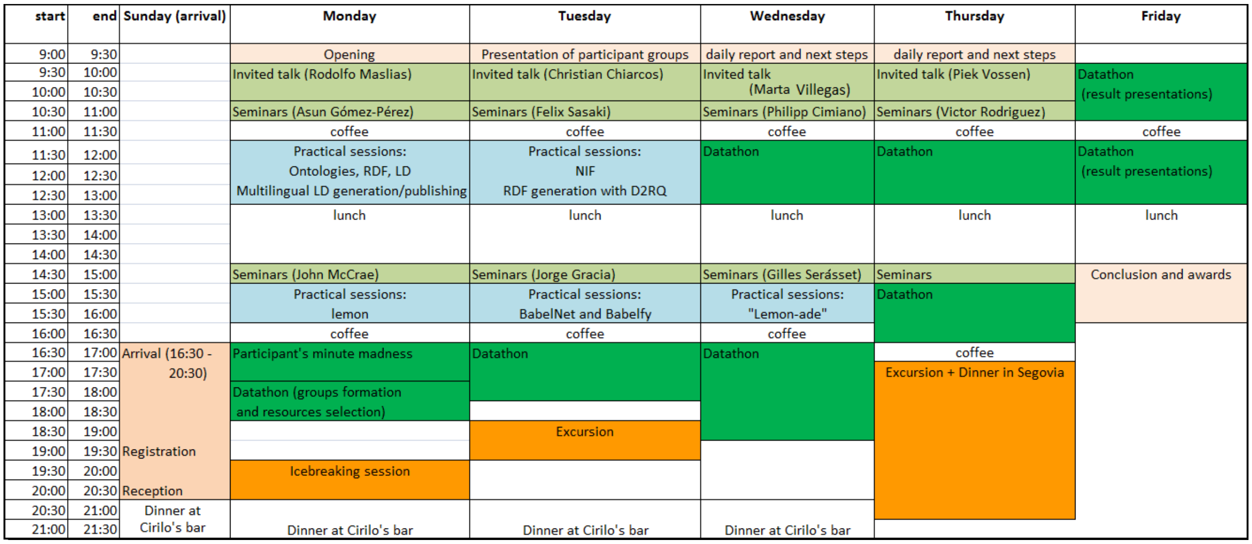
2 Contributions

2.1 Program

The detailed program of the datathon can be seen in **Figure 1**. The datathon's sessions were divided in four categories:

- **Invited talks**. Four selected invited speakers from outside the LIDER consortium were invited to give a talk (forty minutes followed by twenty minutes discussion) about a topic relevant to the datathon.
- **Seminars**, that were theoretical presentations (20 minutes + 10 for questions) given by LIDER members[1] to show novel aspects and discuss selected topics.
- **Practical sessions** to introduce the basic foundations of each topic, methods, and technologies and where participants had the opportunity to do hands-on exercises, guided by the speakers and tutors. The required materials (software and data) were pre-installed in the computers of the datathon computer rooms, and distributed also on USB sticks in case the participants preferred to use their own laptops.
- **Datathon sessions,** in which participants, organised in groups and guided by tutors, planned and performed their own project on linguistic linked data.

[1] With the exception of Gilles Sérasset, attendant to the datathon from Université Joseph Fourier, Grenoble (France) who was invited by the organisers to give a seminar about DBnary.



start	end	Sunday (arrival)	Monday	Tuesday	Wednesday	Thursday	Friday
9:00	9:30		Opening	Presentation of participant groups	daily report and next steps	daily report and next steps	
9:30	10:00		Invited talk (Rodolfo Maslias)	Invited talk (Christian Chiarcos)	Invited talk (Marta Villegas)	Invited talk (Piek Vossen)	Datathon (result presentations)
10:00	10:30						
10:30	11:00		Seminars (Asun Gómez-Pérez)	Seminars (Felix Sasaki)	Seminars (Philipp Cimiano)	Seminars (Victor Rodriguez)	
11:00	11:30		coffee	coffee	coffee	coffee	coffee
11:30	12:00		Practical sessions: Ontologies, RDF, LD Multilingual LD generation/publishing	Practical sessions: NIF RDF generation with D2RQ	Datathon	Datathon	Datathon (result presentations)
12:00	12:30						
12:30	13:00						
13:00	13:30		lunch	lunch	lunch	lunch	lunch
13:30	14:00						
14:00	14:30						
14:30	15:00		Seminars (John McCrae)	Seminars (Jorge Gracia)	Seminars (Gilles Serásset)	Seminars	Conclusion and awards
15:00	15:30		Practical sessions: lemon	Practical sessions: BabelNet and Babelfy	Practical sessions: "Lemon-ade"	Datathon	
15:30	16:00						
16:00	16:30		coffee	coffee	coffee		
16:30	17:00	Arrival (16:30 - 20:30)	Participant's minute madness	Datathon	Datathon	coffee	
17:00	17:30					Excursion + Dinner in Segovia	
17:30	18:00		Datathon (groups formation and resources selection)				
18:00	18:30						
18:30	19:00			Excursion			
19:00	19:30	Registration					
19:30	20:00		Icebreaking session				
20:00	20:30	Reception					
20:30	21:00	Dinner at Cirilo's bar	Dinner at Cirilo's bar	Dinner at Cirilo's bar	Dinner at Cirilo's bar		
21:00	21:30						

Figure 1: Datathon's program.

The invited speakers and their given talks were:

- **Rodolfo Maslias (**Head of the Terminology Coordination Unit, EU Parliament), *“Institutional Terminology, Tools and Communication”*. Abstract - Slides
- **Christian Chiarcos** (Goethe University), *“Linked Open Dictionaries (LiODi) Lexical and phonological search in multilingual dictionaries”*. Abstract - Slides
- **Marta Villegas** (Universitat Pompeu Fabra, Barcelona, Spain), *“Publishing and Consuming Linked Data. (Lessons learnt when using LOD in an application)”*. Abstract - Slides
- **Piek Vossen** (VU University Amsterdam, Netherlands), *“The Global Wordnet Grid”*. Abstract - Slides

All of these presentations captured a lot attention and motivated the debate both during the discussion part and later during the coffee breaks. The four speakers were invited to stay longer, so they had the opportunity of participating in the rest of activities of the datathon and could interact more with the participants.

This is the list of imparted seminars:

- Asun Gómez-Pérez (Universidad Politécnica de Madrid), *“Maximising (Re)Usability of Linguistic Resources using Linked Data”*. Slides
- John McCrae (Bielefeld University), *“lemon: The Lexicon Model for Ontologies”*. Slides
- Felix Sasaki (DFKI) *“Roundtripping of NIF based Linguistic Linked Data with non linked data sources”*. Slides
- Jorge Gracia (Universidad Politécnica de Madrid), *“Apertium RDF: an experience in generating linguistic linked open data”*. Slides
- Philipp Cimiano (Bielefeld University), *“Linked Terminologies: applying linked data principles to terminologies”*. Slides
- Gilles Sérasset (Université Joseph Fourier), *“The DBnary eco-system, data and APIs”*. Slides

- Víctor Rodriguez-Doncel (Universidad Politécnica de Madrid), *"Rights and licenses for language resources"*. Slides

And the practical sessions:

- Jorge Gracia (Universidad Politécnica de Madrid), *"Introduction to Ontologies, RDF and LD".*
- Jorge Gracia and Daniel Vila-Suero (Universidad Politécnica de Madrid), *"Multilingual LD generation and publishing".*
- John McCrae (Bielefeld University), *"lemon"*
- Ciro Baron and Bettina Klimek (University of Leipzig), *"NIF".*
- Andrejs Ābele (Insight, NUIG), *"RDF generation with D2RQ".*
- Tiziano Flati (Universitá di Roma "La Sapienza"), *"BabelNet and BabelFy".*
- Mariano Rico (Universidad Politécnica de Madrid), *"lemon-ade".*

Figure 2 shows a datathon practical session in progress.

Figure 2: Work during a practical session.

2.2 LIDER project members participation

From the LIDER consortium, six members volunteered to act as **tutors:** Gabi Vulcu (Insight, UNIG), Andrejs Ābele (Insight, UNIG), Víctor Rodríguez -Doncel (UPM), Tiziano Flati (UNIROMA1), Bettina Klimek (INFAI, University of Leipzig), Ciro Baron (INFAI, University of Leipzig). The tutors stayed in Cercedilla during the whole duration of the datathon. Every tutor had one or two datathon groups assigned to them and had the responsibility of monitoring their progress, assist them to clarify the group's goals, assure that they followed a proper methodology, and help them with any possible technical issue (asking for assistance to other LIDER members if necessary).

There were also five LIDER **speakers:** Mariano Rico (UPM), Asunción Gómez-Pérez (UPM), Daniel Vila-Suero (UPM), Philip Cimiano (University of Bielefeld), Felix Sasaki (DFKI), Ciro Baron (InfAI) as well as one non-LIDER speaker: Gilles Sérasset (Université Joseph Fourier, Grenoble). They were in charge of giving some seminar talk or leading some practical session.

Most of the tutors acted also as speakers and gave either a seminar or some practical session. Also the datathon directors acted as speakers.

In addition to tutors and speakers, some other LIDER's members with less commitment level acted as **collaborators**, to sporadically help tutors in their tasks and to assist in some logistic aspects. They were Thierry Declerck (DFKI), Guadalupe Aguado-de-Cea (UPM), and Elena Montiel (UPM).

Finally, another LIDER member, José Ángel Ramos, acted as datathon **secretary**, being in charge of all the administrative part.

2.3 Attendants

These are some basic statistics from the registered people:

- 44 participants (59 initial applicants)
- 24 different countries
- 34 cities
- 35% female 65% male
- 22% industry 78% academy

Figure 3 shows a world cloud with the different represented countries along with their relative importance (font size). A world map with the location of the different represented cities can be found in Figure 4 below.

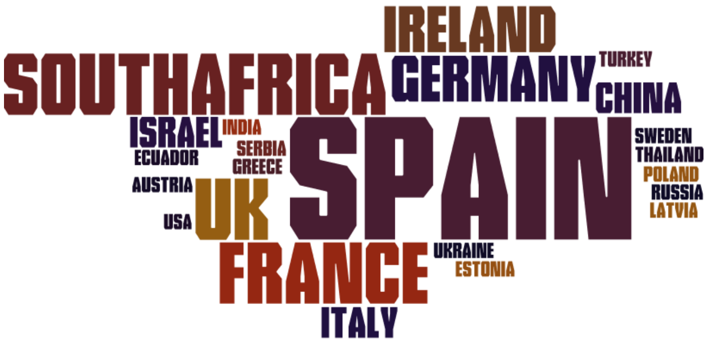


Figure 3: Word cloud with the participant countries.

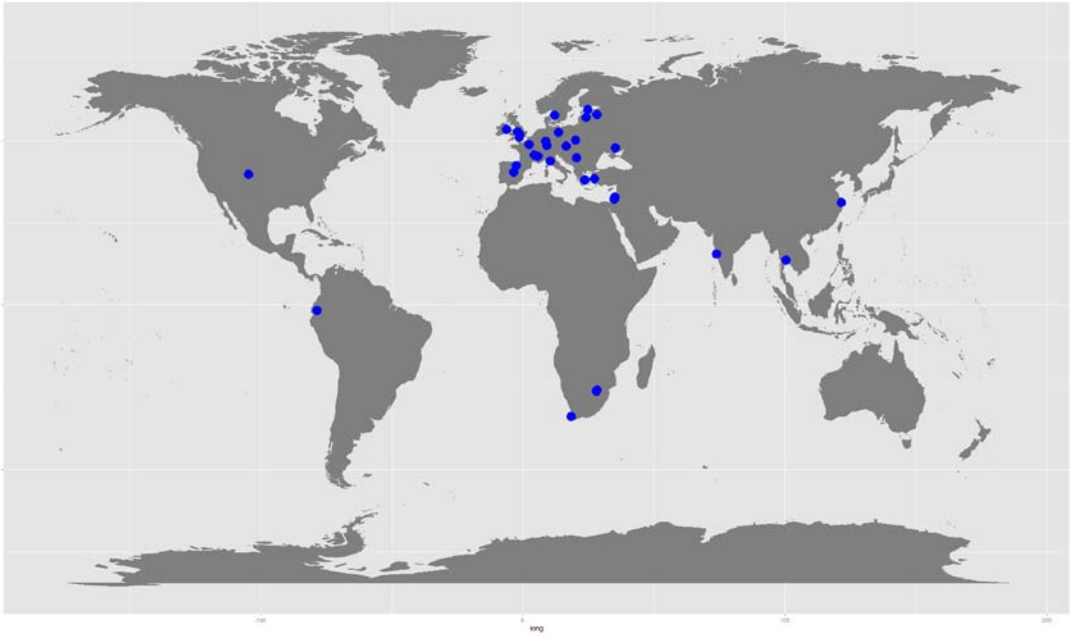


Figure 4: Location of the datathon participant's cities.

There was no unique profile among the different registered participants: there were PhD students, developers from industry, research group leaders, university teachers, etc, coming from different areas: digital humanities, computer science, linguistics, etc. Also the participant's previous experience was not uniform, ranging from little or no experience in linked data to people very experienced in semantic web technologies although willing to explore and contribute to the new LLOD paradigm.

We offered four travelling grants (up to 500€ each), intended for participants who could not cover their trip with other funds and giving preference to those coming from less-developed and/or distant countries.

The list of registered people can be found at http://datathon.lider-project.eu/#participants. In Figure 5 we show a group picture with all the datathon participants (attendants, tutors, speakers, and directors).

Figure 5: SD-LLOD'15 participants.

2.4 Developed projects

During one of the first sessions, the attendants were asked to organise themselves in working groups. Every group would have to select a leader/representative person and would have to decide on the particular topic and datasets to work with. To help in this task, the datathon directors suggested some group leaders and some possible topics, and split or merged unbalanced groups whenever it was necessary. Finally, nine working groups were organised with four to six members each one. This is a short overview of their final projects:

- Group 1: GuanXi Networks. The proposed system tries to overcome, with the use of linguistic linked data, the challenges of using linguistic resources in *language learning* and *NLP* (scattered data, lack of explicit meaning, etc.). They developed a multilingual LD network based on the integration of several resources (PDEV, Slovnyk, CEDICT, COW), using *lemon* and *translation.owl* as models. A linking method based on BabelFy was proposed and manually evaluated. Finally, a case study was described based on new Chinese words recently borrowed by English.

- Group 2: Philological Lexicons. This was about the conversion of the LIDDELL-SCOTT Greek dictionary from XML-TEI into RDF using *lemon* as model.

Additionally, they converted a corpus in Latin (chartes bourguignonnes) to NIF and linked it to BabelNet. The motivation of the latter is to facilitate the discovery of concepts and topics in Latin texts, as well as harmonizing different tagging models (e.g., POS). Linking to BabelNet was challenging owing to orthography and cultural distance issues. In parallel, they also started converting a poetry repository as LLOD, with links to VIAF. Finally, they also planned the conversion of Lewis Short's Latin dictionary and the Latin WikiQuotes.

- Group 3: META-SHARE & LRE-Map. The motivation of this work was LR metadata harmonization and reconciliation, in particular the Meta-Share and the LREMap datasets. ODRL was used to represent license data. They mapped the upper nodes of the ontologies into general categories and created a new ontology using Protégé. Then, the metadata was converted into RDF according to the new model. As future work, such metadata will be published as LLOD and integrated in LingHub.

- Group 4: Terminology on Demand. The motivating use case is a Spanish speaker who does not speak English and wants to extract some knowledge from twitter via SPARQL queries. The system performs term extraction and candidate translations from tweets based on IATE. Then, the original TBX is converted into RDF, and combined with the annotations generated by group 9. Both the IATE and twitter terminologies and annotations were in a triple store and could be queried.

- Group 5: South African Languages. This work was about converting multilingual agricultural data in South African languages as LLOD. The original data came from searchable PDFs, and were converted into CSV to allow their later processing. The extracted entities were searched in BabelNet and Falcon to establish external links. Eleven lexicons were created in RDF with added translations in English and some external links.

- Group 6: Lemonification of two language resources. Two resources were converted into RDF using *lemon*: a Swedish lexicon (saldo) and a dictionary from the Oxford University Press (OUP), each one having different license schemes. Links to DBnary and WordNet were explored. Some OUP dictionary examples were represented in NIF. An evaluation was carried out, based on a sense-sense DBnary-saldo based gold standard. The generated RDF was uploaded into the DBnary service and exposed in a SPARQL endpoint.

- Group 7: K Dictionaries. The project was about converting into RDF a multilingual dictionary (with Spanish as main node), initially in XML and with a privative license. A subset of the elements was left out for the conversion into RDF (some complex structures such as collocations, idioms, etc. that would need a more careful analysis). The models for the RDF representation were *lemon*, lexinfo and SKOS. The produced RDF was loaded in Fuseki and the results queried via SPARQL.

- Group 8: Getty LOD Ontology localisation. The goal was to convert a subset of the Getty LOD Ontology (Agent types: Artists) into RDF (using the translation.owl ontology) and localize it in as many languages as possible. They got translations from several translation systems (including Google Translate, Baidu, Yandex, Bing) and combined them in order to rank the candidates. All the possible translations were included in order to allow for future user-based feedback. As next step, a disambiguation step should be included in order to avoid including noisy translations.

- Group 9: Semantic enrichment of Twitter. The goal was to provide enriched Twitter open data and to publish it as LOD. They used the Open American National Corpus of tweets, and used DBpedia and BabelFy for their semantic enrichment. They were converted into NIF and stored the resulting RDF in Fuseki. They worked in synergy with group 4.

After a voting among the participants, the *GuanXi Networks* project, by Group 1, was selected as the best datathon project and therefore declared as winner of the **"best datathon result" award** (600€, evenly split). Figure 6 shows a picture of the group members after receiving the prize.

Figure 6: Winners of the "best datathon result" award with the datathon organizers after receiving the prize.

2.5 Social aspects

One of the objectives of LIDER (and therefore of the datathon) has been community creation. To favour this, we promoted interaction among participants both through the social activities and the work in groups.

To that end we organised a "**lightning talk session**" on the first day, in which each attendant had to present their background and motivation in one slide in one minute, as a way to know each other better and to help with the group formation. Also in the first day we had an informal **"icebreaking session"** with a variety of social games.

In the second day we had an **excursion** in the surrounding area with a professional guide, in which we visited the surrounding woods, a Roman road, and a water reservoir. On Thursday, we visited the historical city of Segovia, having dinner in a prestigious local restaurant.

The Internet social networks also helped us to disseminate our event. For instance, the datathon activities and social aspects were disseminated in Twitter by using the #sdllod15 (and optionally #LiderEU) hashtags. See https://twitter.com/hashtag/sdllod15?src=hash

The event motivated also some blog posts by non-LIDER members, such as

- http://www.maslias.eu/2015/06/one-big-cloud-all-terminology-all.html (by Rodolfo Maslias, EU Parliament),
- http://inmyownterms.com/linked-data-connecting-the-terminology-dots/ (by Patricia Brenes, Inter-American Development Bank),
- http://rgcl.wlv.ac.uk/2015/06/23/summer-datathon-2015-winners-madrid/ (by the Research Group in Computational Linguistics at the University of Wolverhampton),
- http://kaiko.getalp.org/about-dbnary/21-languages-are-now-available/ (by Gilles Sérasset, Université Joseph Fourier),
- http://news.ecust.edu.cn/news/35158?important=1 [in Chinese] (by East China University Of Science and Technology) .

2.6 Participant's opinions

According to the post-event survey[2], most of the participants evaluated the event very positively, acknowledging the opportunities they had for learning and doing networking in an inspiring environment. In a three-degree scale, the *organisation* of the datathon was considered "very good" by the 95% of participants and "reasonable" by 5%. Nobody rated it negatively ("poor"). The feedback and assistance got by *tutors* was considered "very good" by the majority (84%) and "reasonable" by the rest (16%). Again, nobody rated it negatively ("poor").

The majority of participants (90%) considered that the *focus* of the datathon was neither too academic nor too industry oriented, but "just right". An 84% of participants considered that the *atmosphere* of the datathon was conductive to *learning* and a 100% considered it conductive to *networking*.

Regarding the type of sessions, most of the participants considered *all the sessions* enjoyable and beneficial to them. However we detected that they missed more time for purely practical activities. For instance, despite the duration of *invited talks* and *seminars* was considered "just right" (79% and 89% respectively), the duration of the *practical sessions* and *datathon sessions* were considered "too short" (78% and 79% respectively). This is an improvement point to be considered in future editions of the datathon.

In terms of *learning per topi*c, this is a summary table with the participants' opinions on how much they did learn on each topic:

[2] The satisfaction questionnaire was distributed online during the last day of the datathon and it was answered by 19 participants (43% out of the total).

	A lot	Something	Little	Not at all
Semantic Web	50%	39%	11%	0%
Linked Data	50%	39%	11%	0%
Language resources	28%	50%	17%	5%
Linguistic Linked Data	78%	17%	5%	0%
Linguistic Linked Data applications	44%	39%	11%	6%
Lemon	42%	42%	11%	5%
NIF	37%	53%	5%	5%

Finally, the *social activities* were also very well rated. All the participants (100%) who answered the questionnaire considered that they benefited “a lot” from the social activities.

2.7 Key Points of the Datathon

In the following, we summarise the main outcomes and benefits of the datathon:

- Increase of awareness of Linguistic Linked Data.
- Community creation: the datathon attracted many people interested in linguistic linked data, putting them in connection both among them and with other experts in the field (invited speakers, tutors, etc.). The participants were encouraged to:
    - Join the relevant W3C and OKF community groups, specially W3C Linked Data for Language Technologies (LD4LT) community group and OKF Open Linguistics working group,
    - Keep the contact with the other participants (by email, LinkedIn, etc.),
    - Finish the work started in groups during the datathon and submit it to relevant workshops/conferences.
- Dissemination of guidelines and best practices for linguistic linked data[3]. All the guidelines, reference cards, etc. generated in the context of the LIDER project[4] and W3C community groups were disseminated among the participants. Also the Linghub aggregator of linguistic metadata[5] was introduced to the participants.
- Through the projects developed by the participants, several benefits were obtained:
    - Increase of the (future) amount of linguistic linked data on the Web[6],
    - Identification of a number of potential applications and use cases for linguistic linked data technologies,
    - (Partial) development of such linked data based applications.

[3] http://www.lider-project.eu/guidelines

[4] Deliverables D2.1.1, D2.1.2

[5] http://linghub.lider-project.eu/

[6] The RDF generated in the projects will be eventually published as LLOD on the Web, but most of it needed some more elaboration at the time the datathon was finished. This should be done by the own participants, with the assistance of LIDER members if required.