mylider

an introduction to the LIDER project

Jorge Gracia
(on behalf of LIDER's consortium)

Ontology Engineering Group (OEG)
Artificial Intelligence Department
Universidad Politécnica de Madrid (UPM)

jgracia@fi.upm.es

Talk at OEG
16/01/2014

- The consortium

- Context

- Main goal

- Is there industrial interest for this?

- Is LOD ready for that?

- LIDER's challenge

- Expected outcomes

- LIDER's community

- Some related technologies

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

The consortium

Universidad Politécnica de Madrid (UPM, Spain) [*COORDINATOR*]

Trinity College Dublin (Ireland)

DFKI (Germany)

National University of Ireland, Galway (NUI Galway, Ireland)

Institut für Angewandte Informatik EV (INFAI, Germany)

University of Bielefeld (Germany)

Universita degli Studi di Roma La Sapienza (Italy)

GEIE ERCIM (France)

A.   Gómez-Pérez

A. Ibarrola

G.Aguado

J. Gracia

E. Montiel-Ponsoda

D. Vila

V. Rodríguez

UPM

Prof David Lewis   **Dr. Dominic Jones**

Kevin Koidl

Prof Declan O'Sullivan

Felix Sasaki   **Arle Lommel**

Thierry Declerck

NUI Galway
OÉ Gaillimh

Paul Buitelaar   **Georgeta Bordea**   **Mihael Arcan**

UNIVERSITÄT LEIPZIG

Sören Auer   **Jens Lehmann**   **Sebastian Hellmann**   **Martin Brümmer**

Prof Dr. Philipp Cimiano   Matthias Hartung   Dr. John McCrae   Dr. Christina Unger

Universität Bielefeld

**Roberto Navigli**   **Paola Velardi**   **Maud Ehrmann**   **Andrea Moro**

Context

Content analytics methods allow discovering relevant knowledge and making sense of the continuously growing amount of unstructured data on the Web or within organizations, for the purpose of improved decision making.
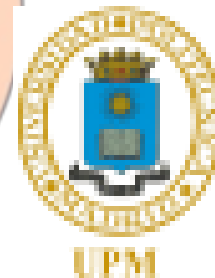
Content analytics methods use natural language processing (NLP), trend analysis, contextual discovery and predictive analytics to uncover patterns and trends across such unstructured content
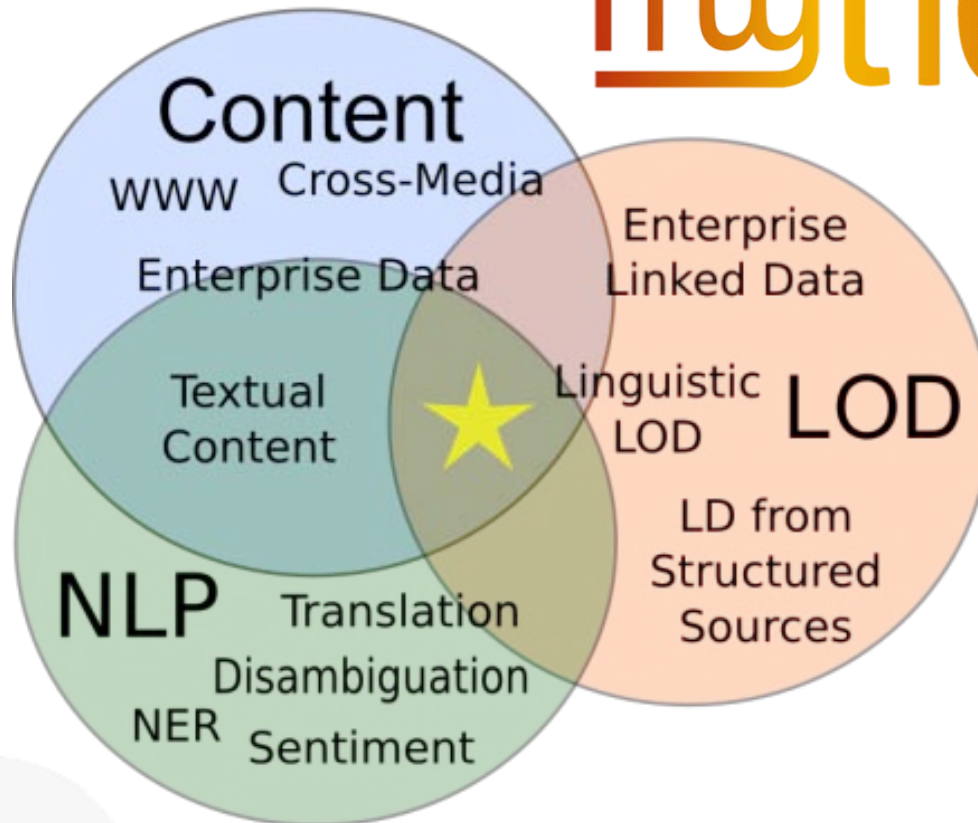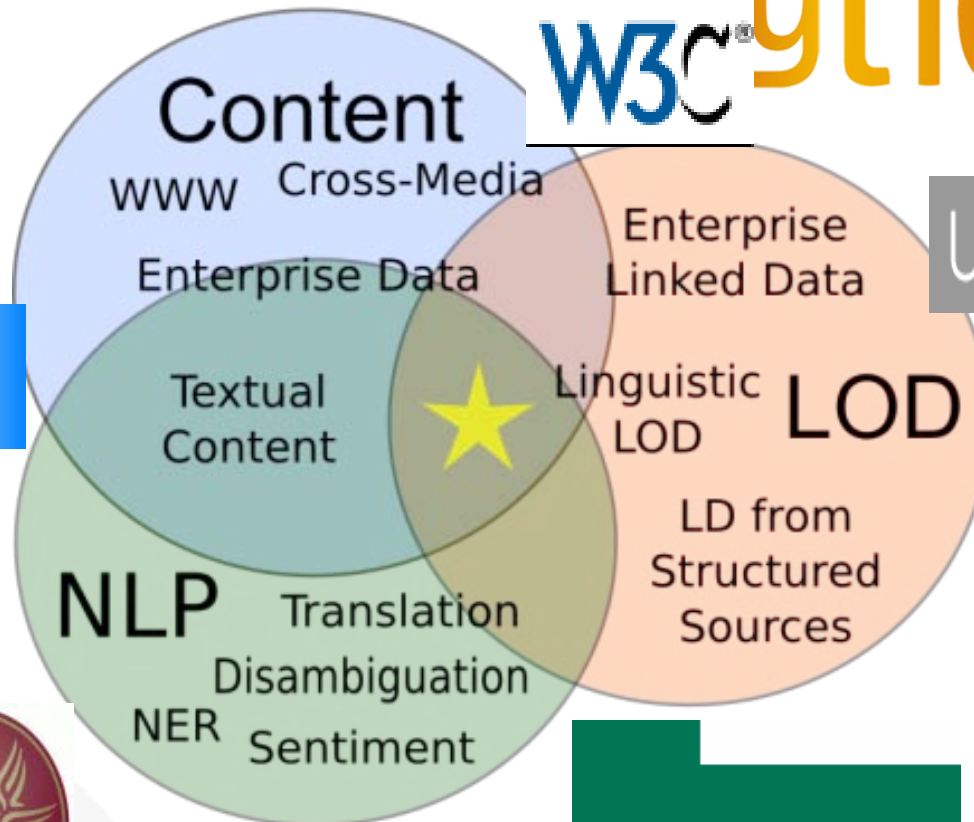
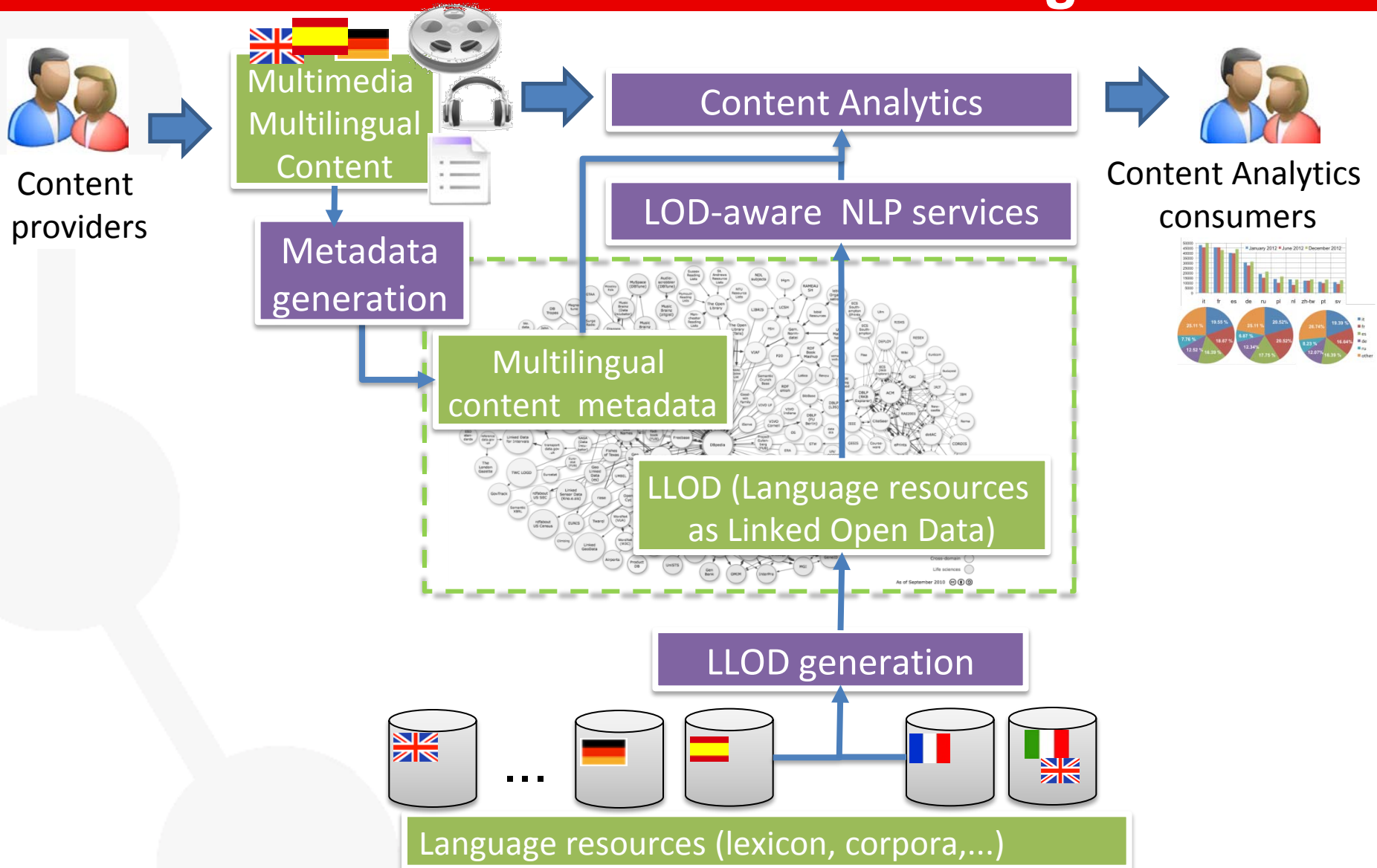# Main goal

Enable a new generation of LOD-aware NLP techniques to support content analytics of multilingual multimedia content

It is NOT about solving the technical aspects but about establishing the basis and the means to that end (guidelines, best practises, community, ...)

Content providers

Multimedia Multilingual Content

Content Analytics

Content Analytics consumers

Metadata generation

LOD-aware NLP services

Multilingual content metadata

LLOD (Language resources as Linked Open Data)

LLOD generation

...

Language resources (lexicon, corpora,...)
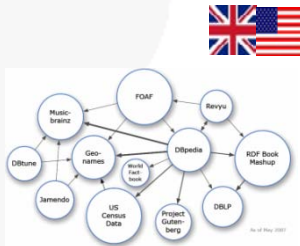
Is there industrial interest for this?

- Multilingual multimedia content annotation

    Increase demand for NLP services that combine text processing with Multimedia meta-data and media processing components.

- LOD generation from linguistic resources

    Data is already being published by companies, but not linguistic resources as LLOD

- LOD-based NLP services for Content Analytics

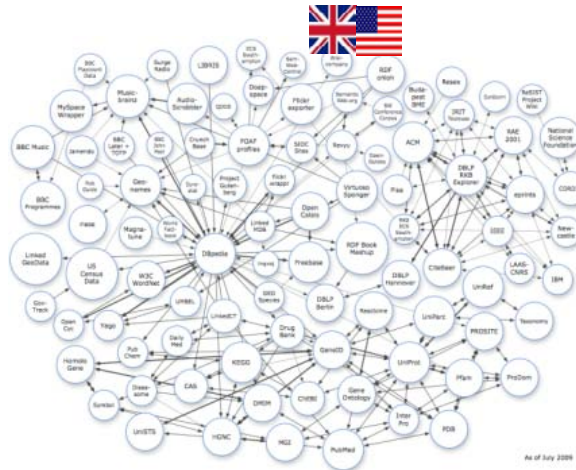   CA related companies that actively use Dbpedia: OpenCalais, Zemanta, Ontos, Yahoo!, Nerd, etc.

Is LOD ready for that?

- 2007

- 2009

- 2012
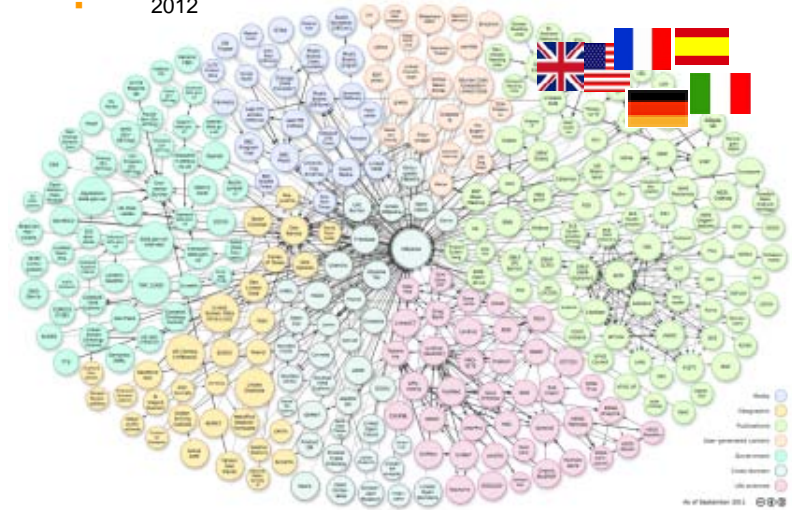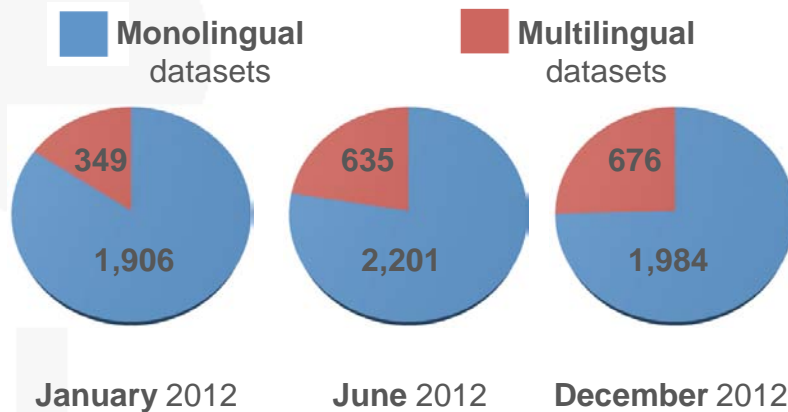
Monolingual datasets **Multilingual** datasets

349 | 1,906 — January 2012
635 | 2,201 — June 2012
676 | 1,984 — December 2012

1. Number of Monolingual and multilingual datasets

RDF literals **without** language tag RDF literals **with** language tag

2,567,324 | 10,250,936 — January 2012
3,154,779 | 10,594,338 — June 2012
3,365,930 | 12,272,806 — December 2012
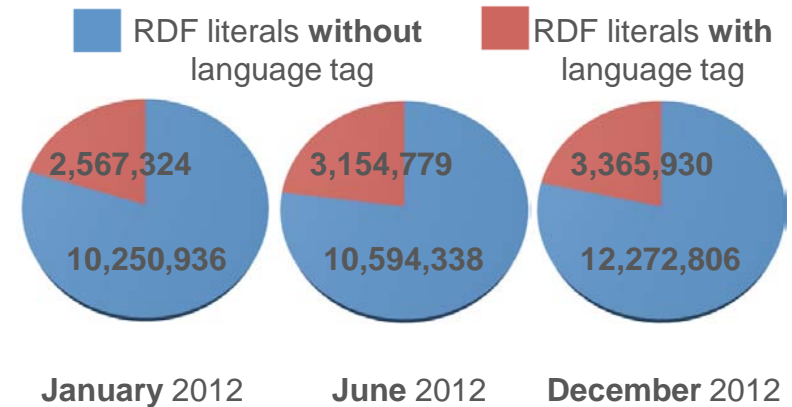
2. Current usage of language tagging capabilities in RDF

RDF literals **with English** tag RDF literals **with other** language tag

431,660 | 2,135,664 — January 2012
403,714 | 2,751,065 — June 2012
557,785 | 2,808,145 — December 2012
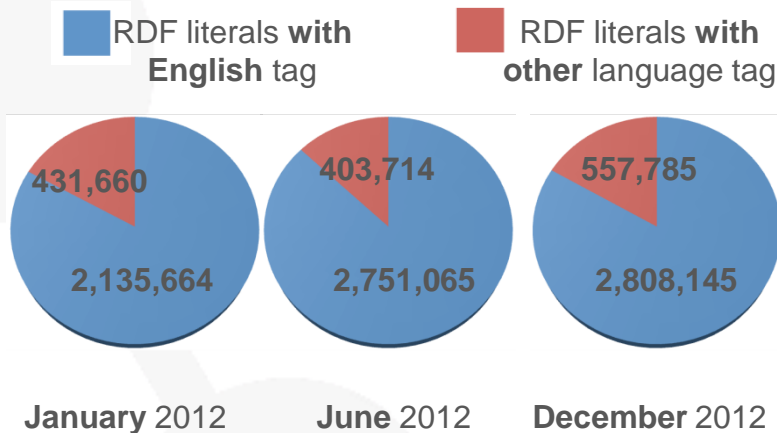
3. English tags versus other languages' tags

January 2012 — June 2012 — December 2012

it, fr, es, de, ru, pl, nl, zh-tw, pt, sv

Pie chart 1: 19.55%, 18.67%, 16.39%, 12.52%, 7.76%, 25.11%
Pie chart 2: 20.52%, 20.52%, 17.75%, 12.34%, 6.87%, 25.11%
Pie chart 3: 19.39%, 16.64%, 16.39%, 12.07%, 8.23%, 26.74%

Legend: it, fr, es, de, ru, other
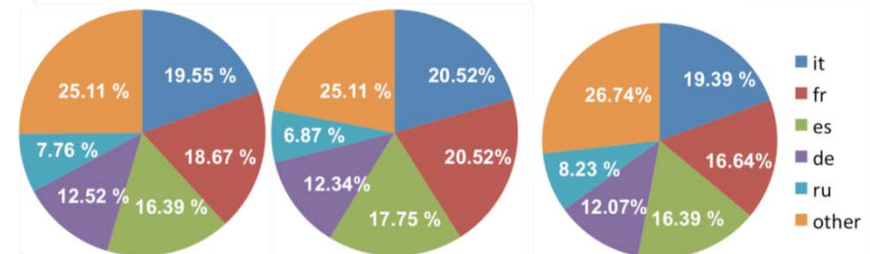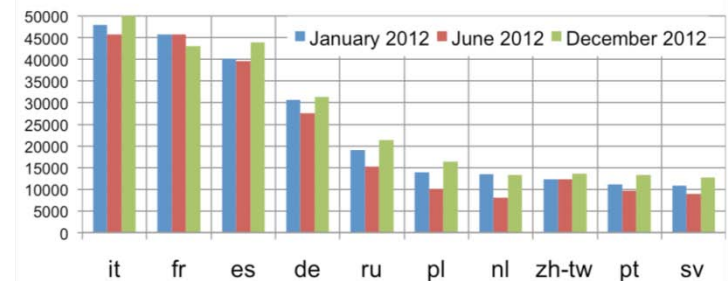
4. Evolution of top-10 languages
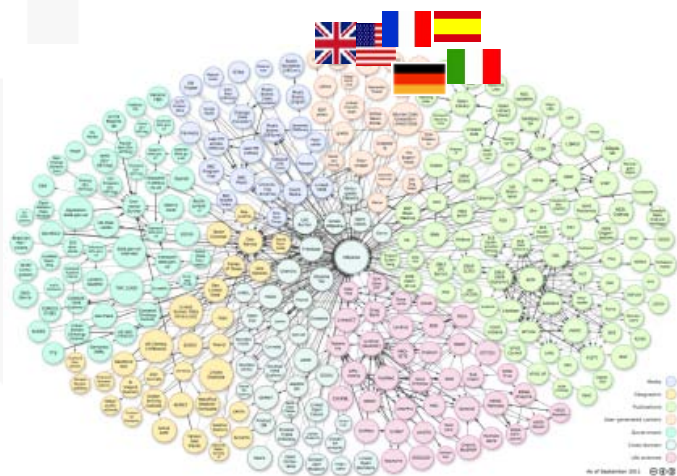
Linguistic LOD (LLOD) cloud

LOD cloud

Linguistic Linked Open Data Cloud - January 2013
This image was produced
with software provided
by Richard Cyganiak

Is LOD ready for that?

Not really:

- Multilingualism causes many issues in (still unsolved)
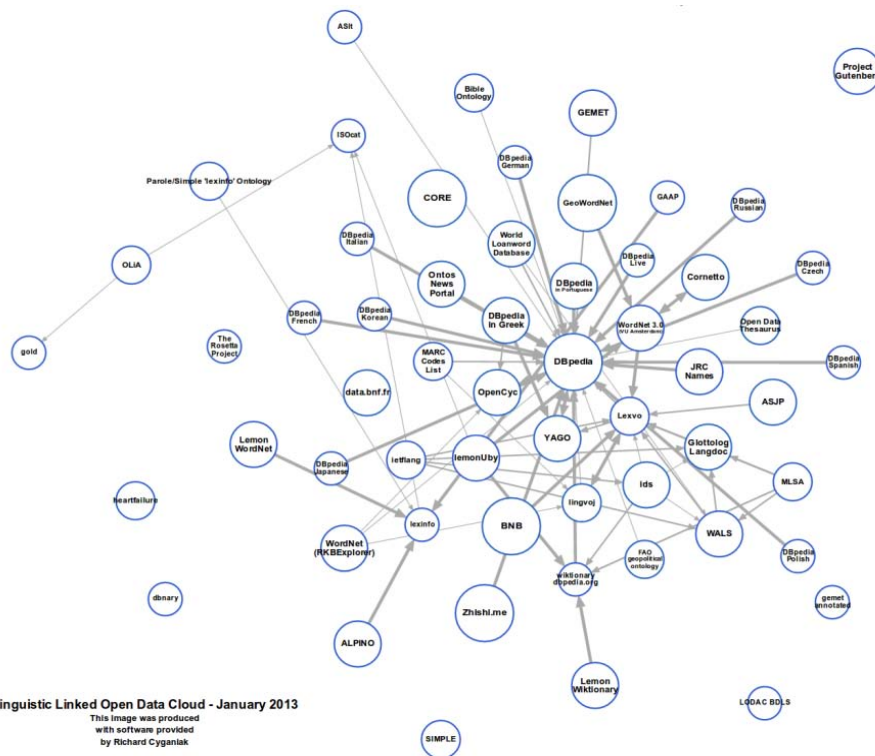- LLOD is sparse, small, and not really "linguistic"

LIDER's challenge

Which extensions to the current LOD are needed to support a new generation of large-scale content analytics applications that will overcome language barriers?

- Identification of new linguistic resources, metadata, guidelines, tools, etc. for the LOD

- Linguistic Licensed Linked Data (3LD)

- Specification of a new generation of NLP services that are LOD-aware and can exploit LOD

Expected outcomes

- Open and sustainable community
- Guidelines and best practices for multimedia and multilingual metadata generation
- Reference architecture
- Research and Innovation Roadmap
- Linguistic LOD as a Free Open and Interoperable (FOI) infrastructure
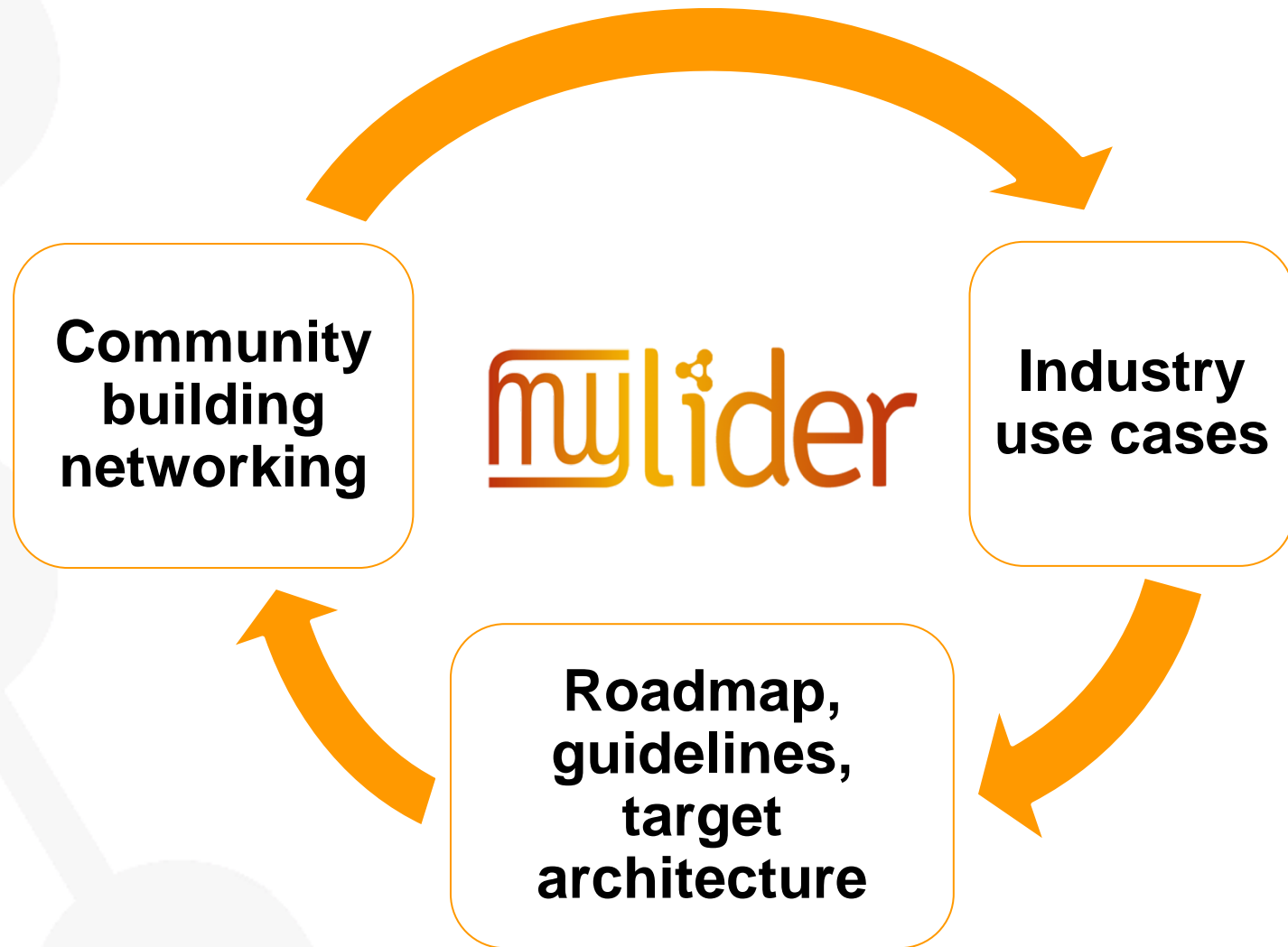
LIDER's community

- Use case definition from industry will be input to the roadmap

- Linguistic resources →LLOD

- Validation of guidelines and reference architecture

- Participation in surveys

- Participation in events (roadmapping WS, hackatons, etc.

#mlw

http://www.multilingualweb.eu/

#Lider-eu

Project-Portal (UPM)

www.lider-project.eu

Public Mailing List (UPM)

Linked Data for Language Technology (Industrial Board)

BP-MLOD W3C-CG
OntoLex W3C-CG

Some Related Technologies

An RDF-based lexicon model for ontologies

Linked Data Stack

Creating Knowledge out of Interlinked Data

http://stack.linkeddata.org/
http://stack.lod2.eu/blog/

A very large multilingual encyclopedic dictionary and semantic network

Thanks for your attention !