LIDER

LIDER

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

A. Gómez-Pérez (UPM)

asun@fi.upm.es

Project Coordinator

CSA
Budget: 1.482.000€
Starting date: 1. Nov. 2013
Duration: 2 Years

POLITÉCNICA · NUI Galway OÉ Gaillimh · DFKI · UNIVERSITÄT LEIPZIG · Universität Bielefeld · W3C

LIDER

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

Trinity College Dublin (Ireland)

DFKI (Germany)

National University of Ireland, Galway (Ireland)

Institut für Angewandte Informatik EV (INFAI, Germany)
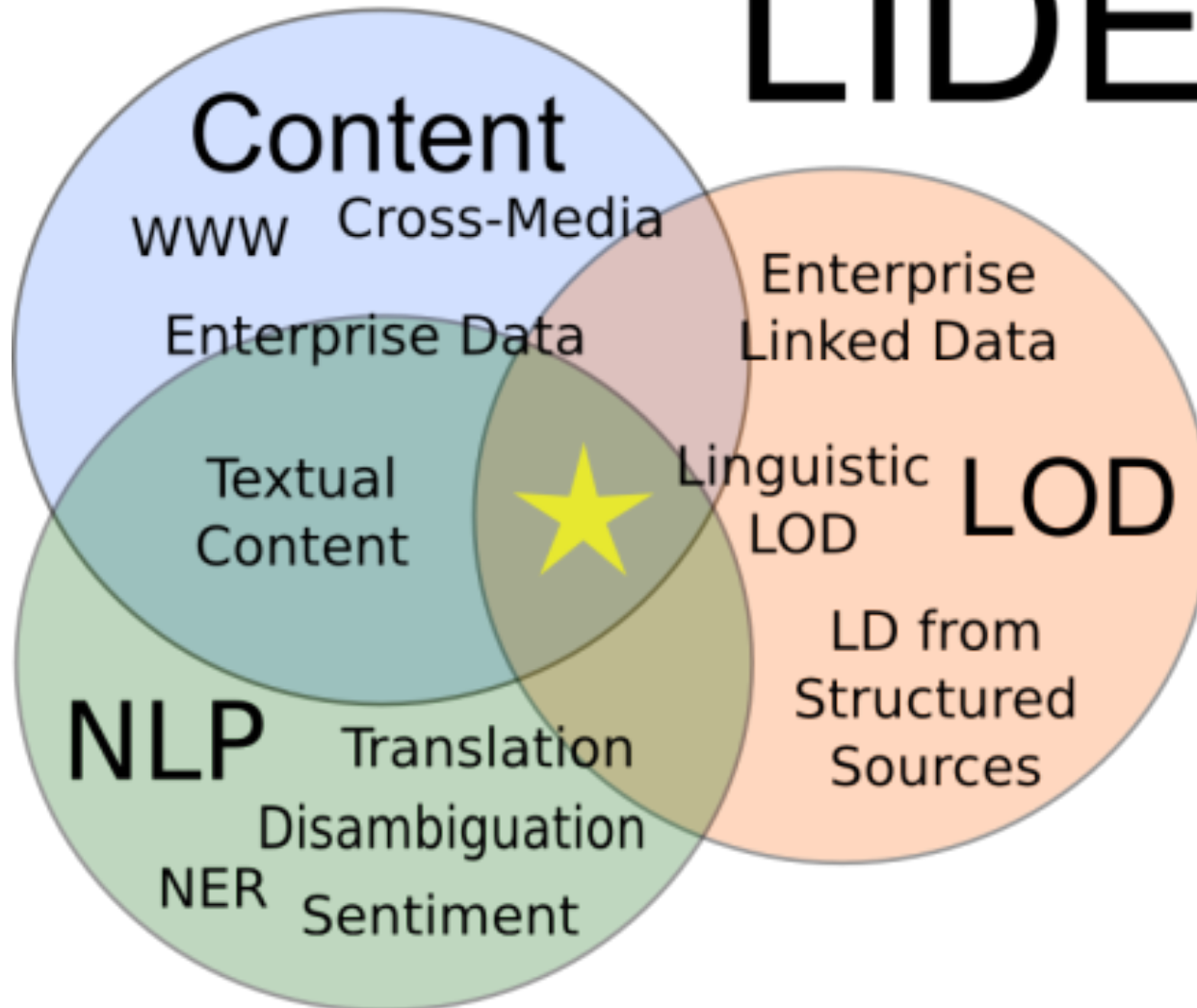
University of Bielefeld (Germany)
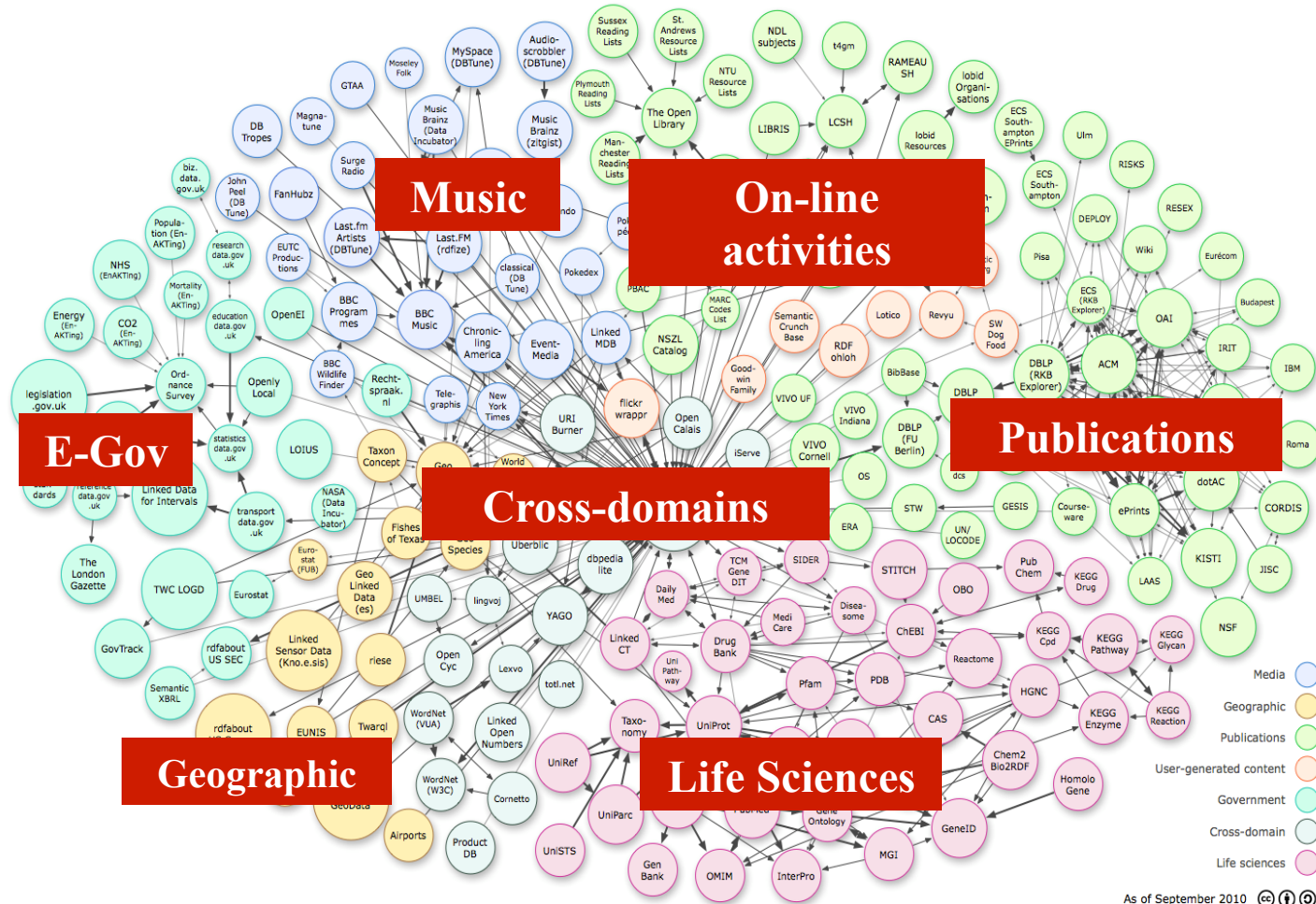
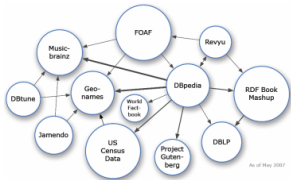**Universita degli Studi di Roma La Sapienza (Italy)**

GEIE ERCIM (France)

LIDER

LIDER

1. LOD is increasingly multilingual
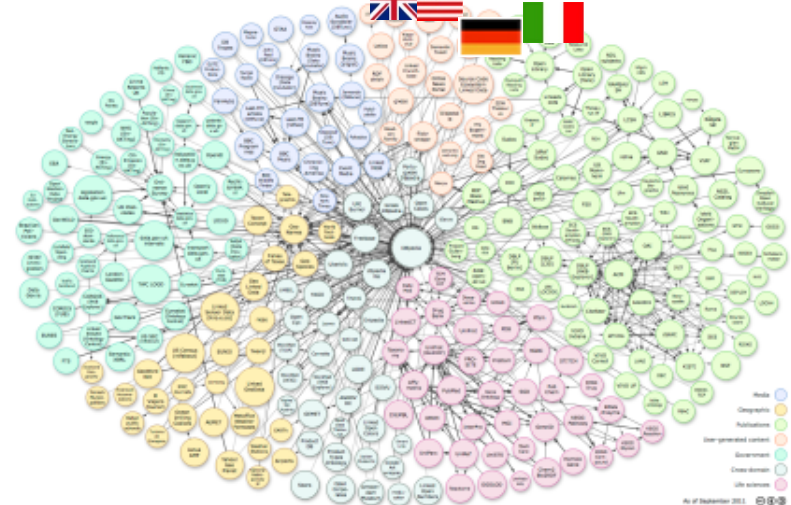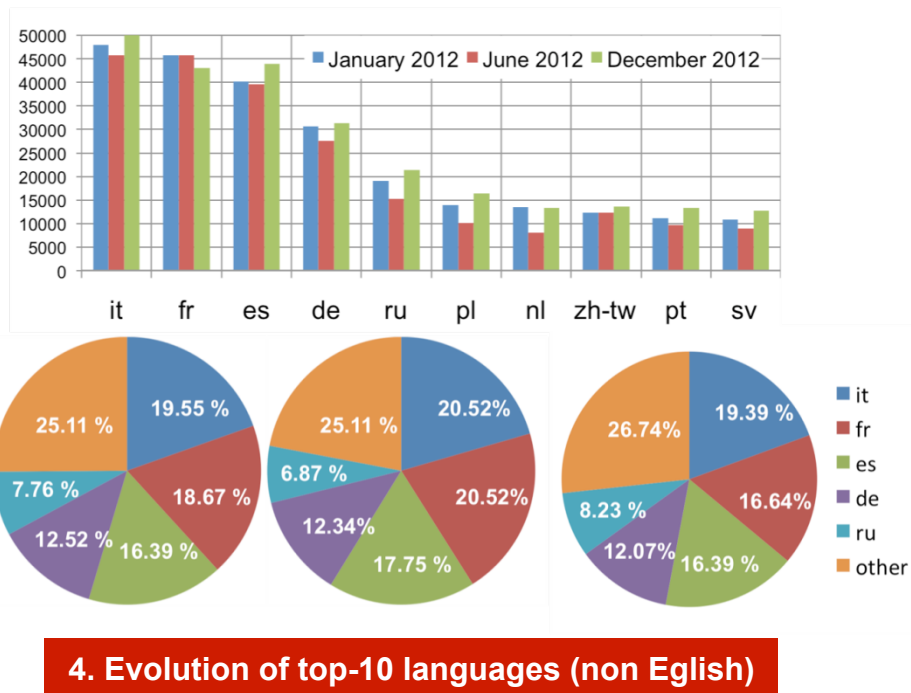2. LOD interconnects resources in many languages

- 2007
- 2009
- 2012

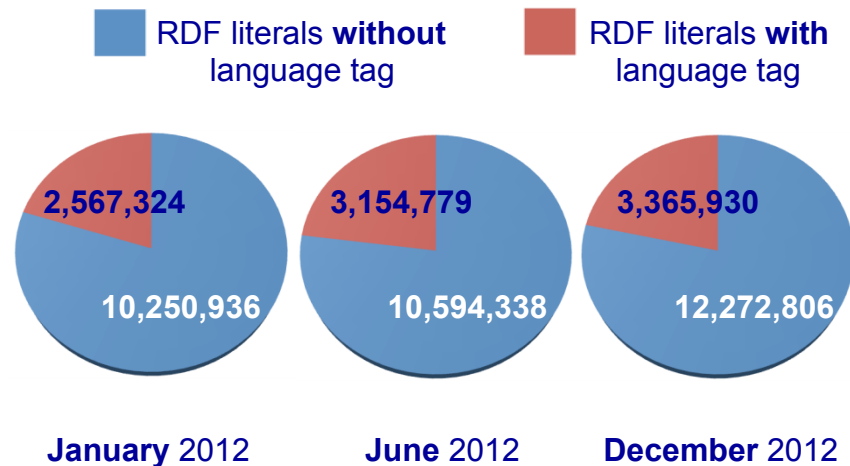# LOD is dominated by the English language

Monolingual datasets   **Multilingual** datasets   RDF literals **without** language tag   RDF literals **with** language tag

349 — 1,906   635 — 2,201   676 — 1,984

2,567,324 — 10,250,936   3,154,779 — 10,594,338   3,365,930 — 12,272,806

January 2012   **June** 2012   **December** 2012

January 2012   **June** 2012   **December** 2012

1. Number of Monolingual and multilingual datasets

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   403,714 — 2,751,065   557,785 — 2,808,145

January 2012   **June** 2012   **December** 2012

January 2012   June 2012   December 2012

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

it   fr   es   de   ru   other

19.55 %   18.67 %   16.39 %   12.52 %   7.76 %   25.11 %

20.52%   20.52%   17.75 %   12.34%   6.87 %   25.11 %

19.39 %   16.64%   16.39%   12.07%   8.23 %   26.74%

3. English tags versus other languages' tags
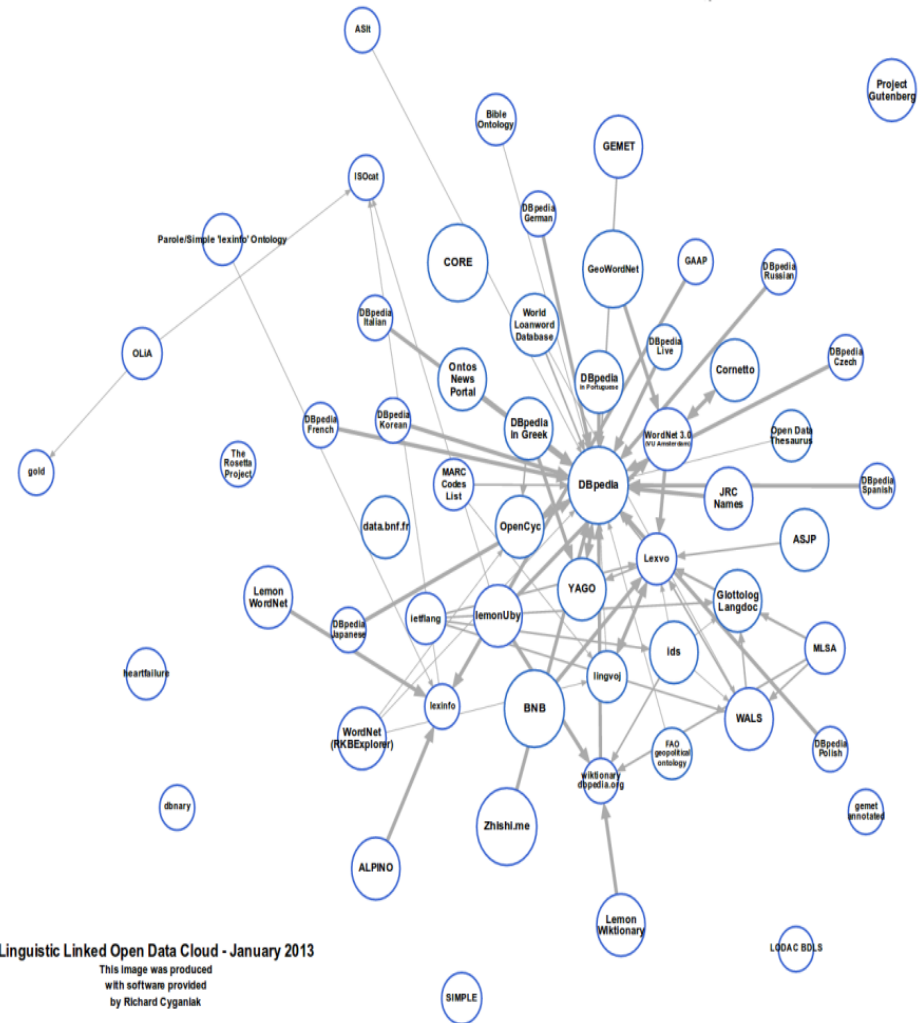
4. Evolution of top-10 languages (non Eglish)

Linguistic LOD (LLOD)

- Subset of LOD
- Linguistic and Open resources in RDF interconnected with other Linguistic and Open resources
- Not too many linguistic resources as LOD

Linguistic LD (LLD)

- Licensed linguistic linked data

LOD, LLOD and LLD as a source of large background knowledge for NLP

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

- State-of-the-art methods in content analytics are based on statistical and machine learning approaches

- The need
  - availability of large quantities of qualitative background knowledge and of interconnected multilingual language resources.

- Which extensions to the LOD are needed to support a new generation of large-scale content analytics applications that will overcome language barriers.
  - identification of key NLP tasks that require background knowledge
  - Specification of a new generation of NLP services that are LOD-aware and can exploit LOD
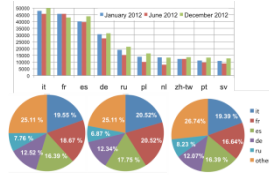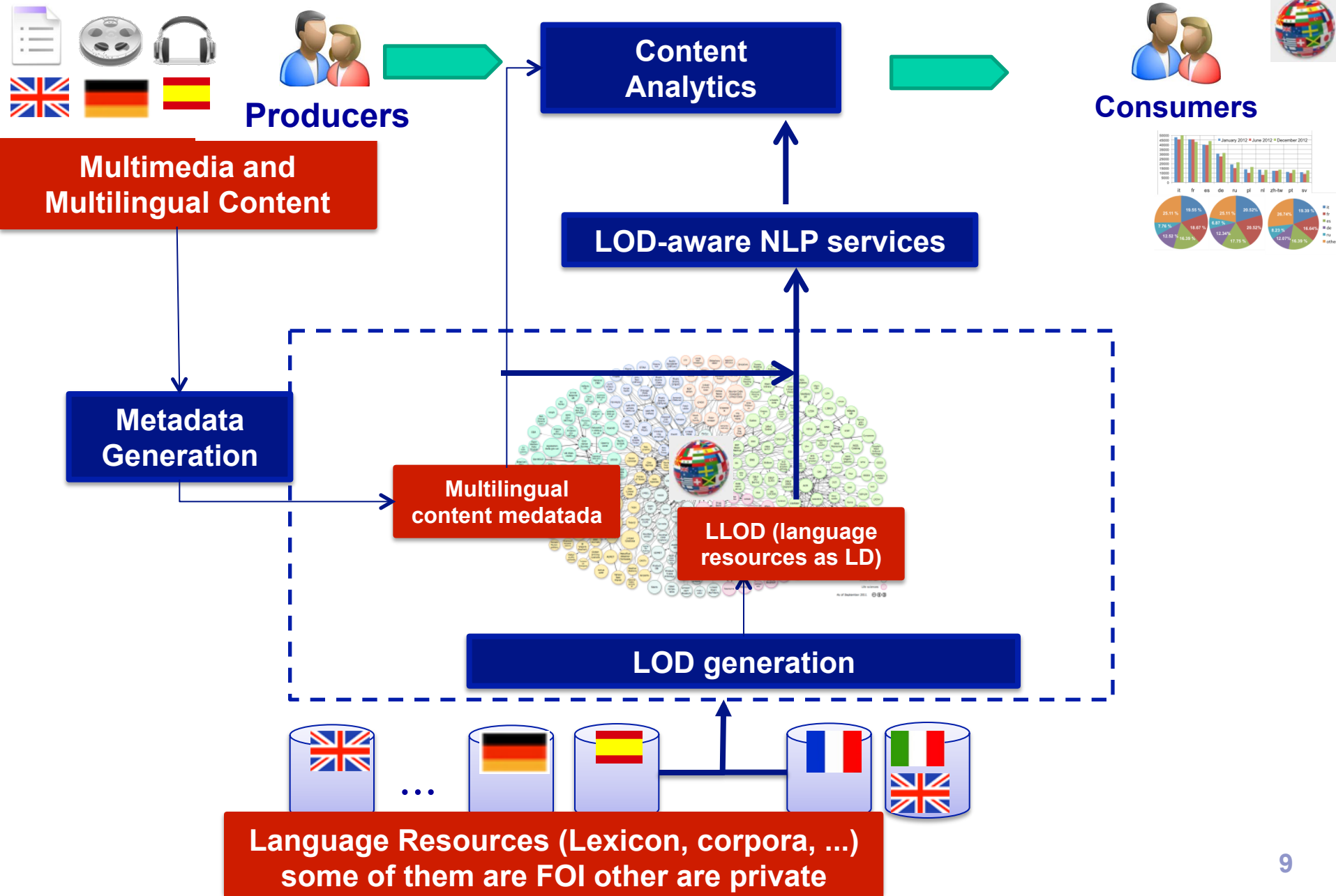
- *Multilingual multimedia content annotation.*
  - o *Increase demand for NLP services that combine text processing with Multimedia meta-data and media processing components.*

- *LOD generation from linguistic resources*
  - o data is already being published by companies, but not linguistic resources as LLOD

- *LOD-based NLP services for Content Analytics*
  - o CA related companies that actively use Dbpedia: OpenCalais, Zemanta, Ontos, Yahoo!, Nerd, etc.
  - o multilingual LLOD would be vital for reaching EU-wide and global markets
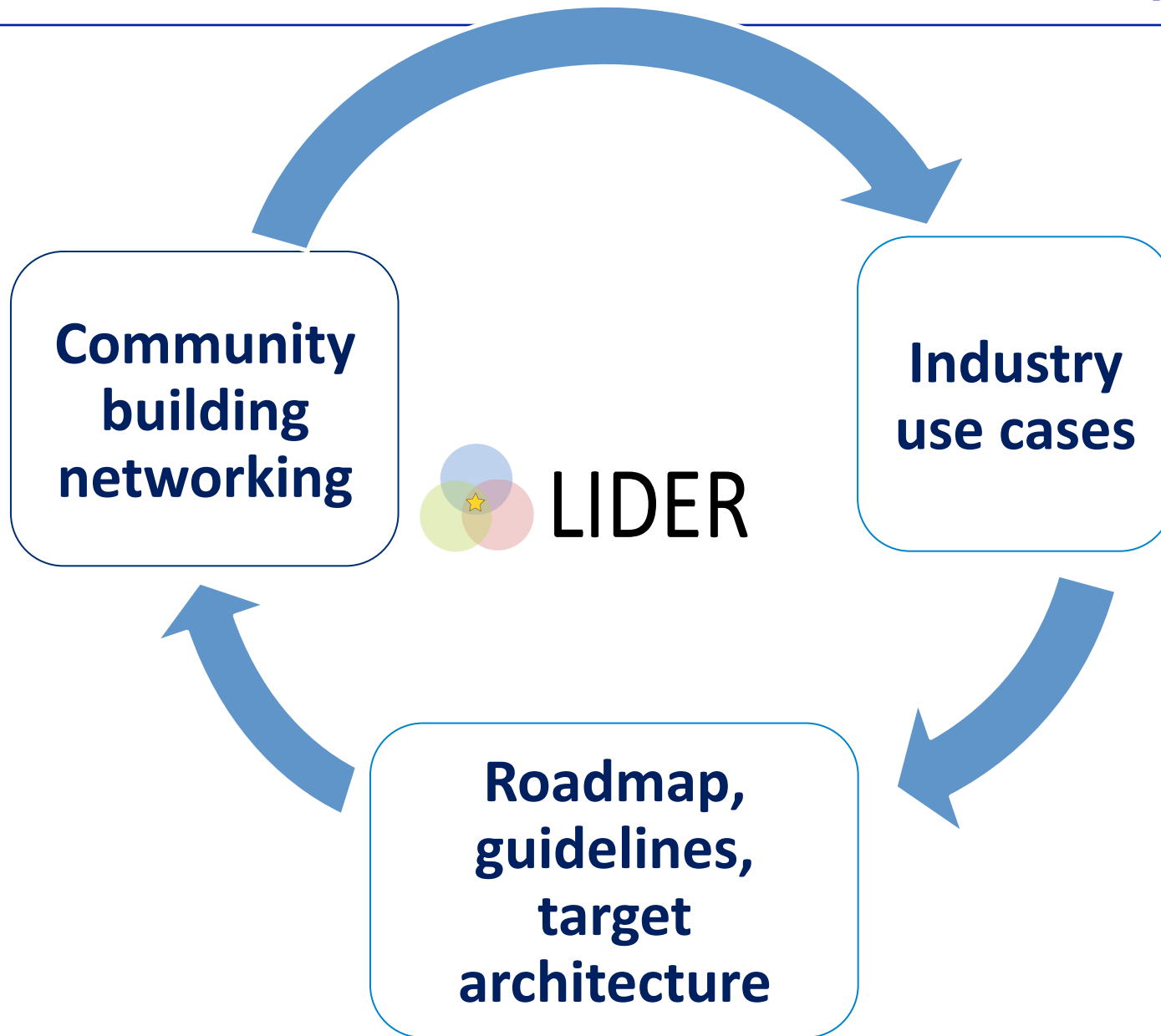
1. Definition of business use cases

- Extract requirements needed to exploit LLD in content analytics processes
- Extract common and frequent NLP-based tasks that are needed for content analytics.

- Business use cases: LLD in CA

2. Definition of Guidelines and best practices for:

- Multimedia and multilingual content metadata generation and use
- LLD generation
- NLP services built on top of LLD

- Guidelines and best practices: LLD for CA
- Linguistic LOD

3. Reference Architecture and Roadmap for content analytics

- Reference architecture: reference model + architectural patterns
- Roadmap involving the academic community and industry

- LLD Reference Architecture
- Roadmap: LLD for CA

4. Community Building and Dissemination

- o Industrial Board
- o Open community Events tailored to the different audiences
  - Roadmapping Workshops
  - Surveys to localization industry and general Web companies
  - Sessions at W3C Multilingual Web Workshop and European Data Forum
  - Publication of best practices material via W3C community groups
  - Hackathons
- o Community portal
  - Relying on http://www.multilingualweb.eu portal and the related social channels
- o Dissemination activities

- 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.

lider-community@delicias.dia.fi.upm.es

Lider will help with travelling grants to participants in Roadmapping WS

LIDER

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

A. Gómez-Pérez (UPM)

asun@fi.upm.es

Project Coordinator