Datos abiertos de Interés Lingüístico

Prof. Dr. Asunción Gómez-Pérez

Artificial Intelligence Department

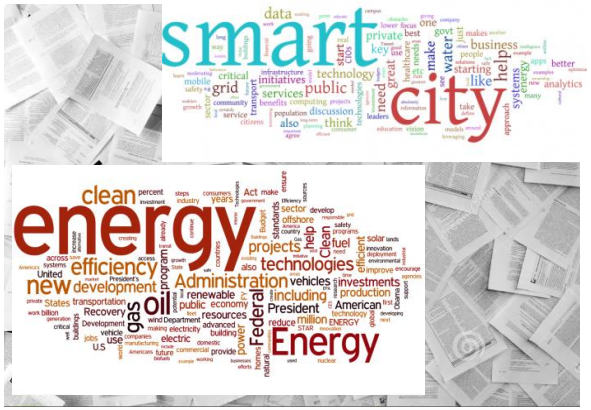
Universidad Politécnica de Madrid
Campus de Montegancedo sn
28660 Boadilla del Monte, Madrid

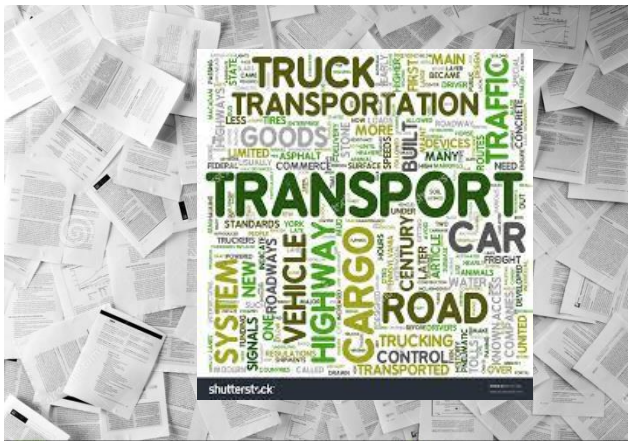
http://www.oeg-upm.net

asun@fi.upm.es
Phone: 34.91.3367439
Fax: 34.91.3524819

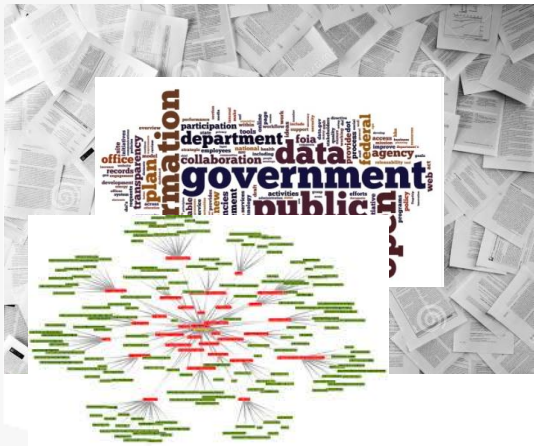
Domain Terminologies

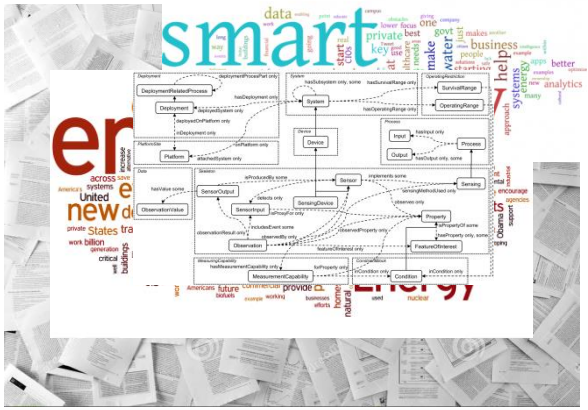
Domain Terminologies



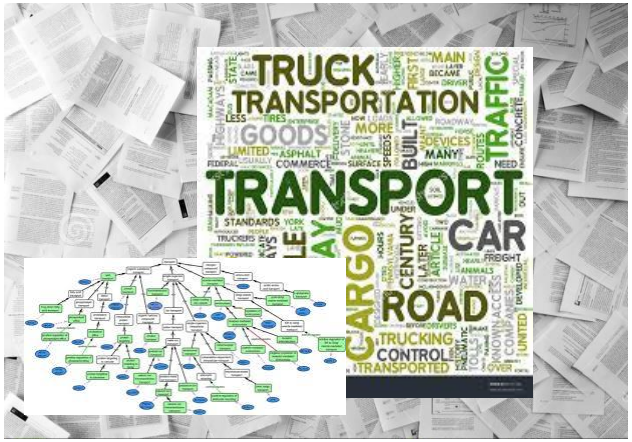


Domain Terminologies





MedDRA

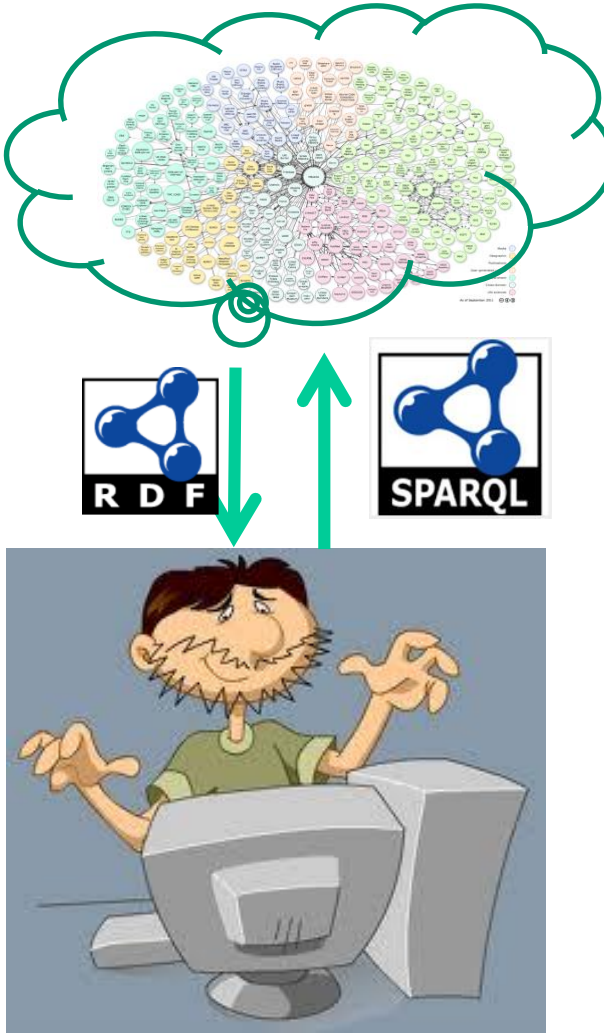


The Open Data stars

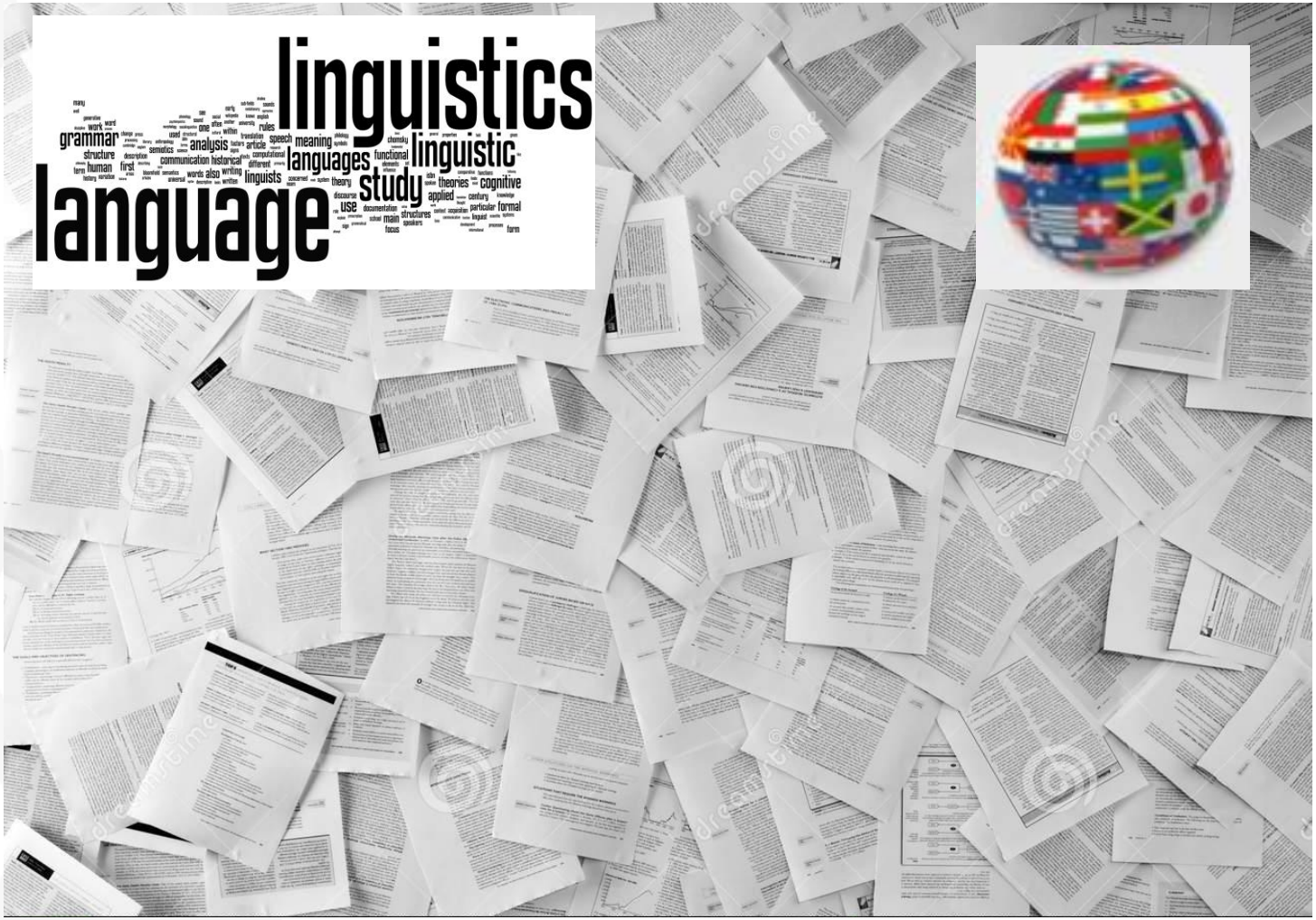
- ★ Make your stuff available on the Web (whatever format) under an open license
- ★★ Make it available as structured data (e.g., Excel instead of image scan or a table
- ★★★ Use non-proprietary formats (e.g., CSV instead of Excel)
- ★★★★ **Use URIs to identify things, so that people can point at your stuff**
- ★★★★★ **Link your data to other data to provide context**

Linked Data allows uniform access

1. **Agree on ontologies** for describing metadata and domain data
2. Unified and standardized language for describing resources (**RDF(S)**)
3. Unified and standardized **query language** (SPARQL)
4. Standardized non-proprietary APIs
5. Links to other resources



Open data in linguistics



Open data in linguistics

In Boo

As files on the Web following standards

★ ★ ★

REAL ACADEMIA ESPAÑOLA

La institución | Obras académicas | Biblioteca y Archivo | Consultas lingüísticas

Inicio » Recursos » Diccionarios » Diccionario de la lengua española

Diccionario de la lengua española

El *Diccionario de la lengua española* (*DRAE*) es la obra de referencia de la Academia. La edición actual —la 22.ª, publicada en 2001— incluye más de 88 000 entradas.

red

á é í ó ú ü ñ

Ayuda

red.

Artículo enmendado

(Del lat. *rete*).

1. f. Aparejo hecho con hilos, cuerdas o alambres trabados en forma de mallas, y convenientemente dispuesto para pescar, cazar, cercar, sujetar, etc.

2. f. Labor o tejido de mallas.

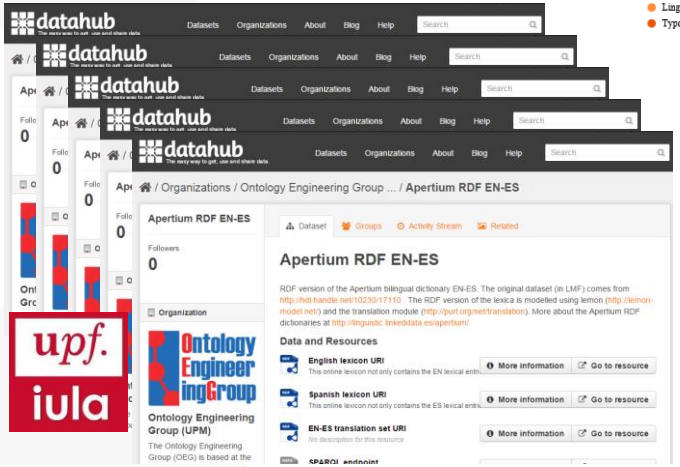
3. f. redecilla (II prenda de malla para el pelo).

4. f. Lugar donde se vende pan u otras cosas que se dan por entre verjas.

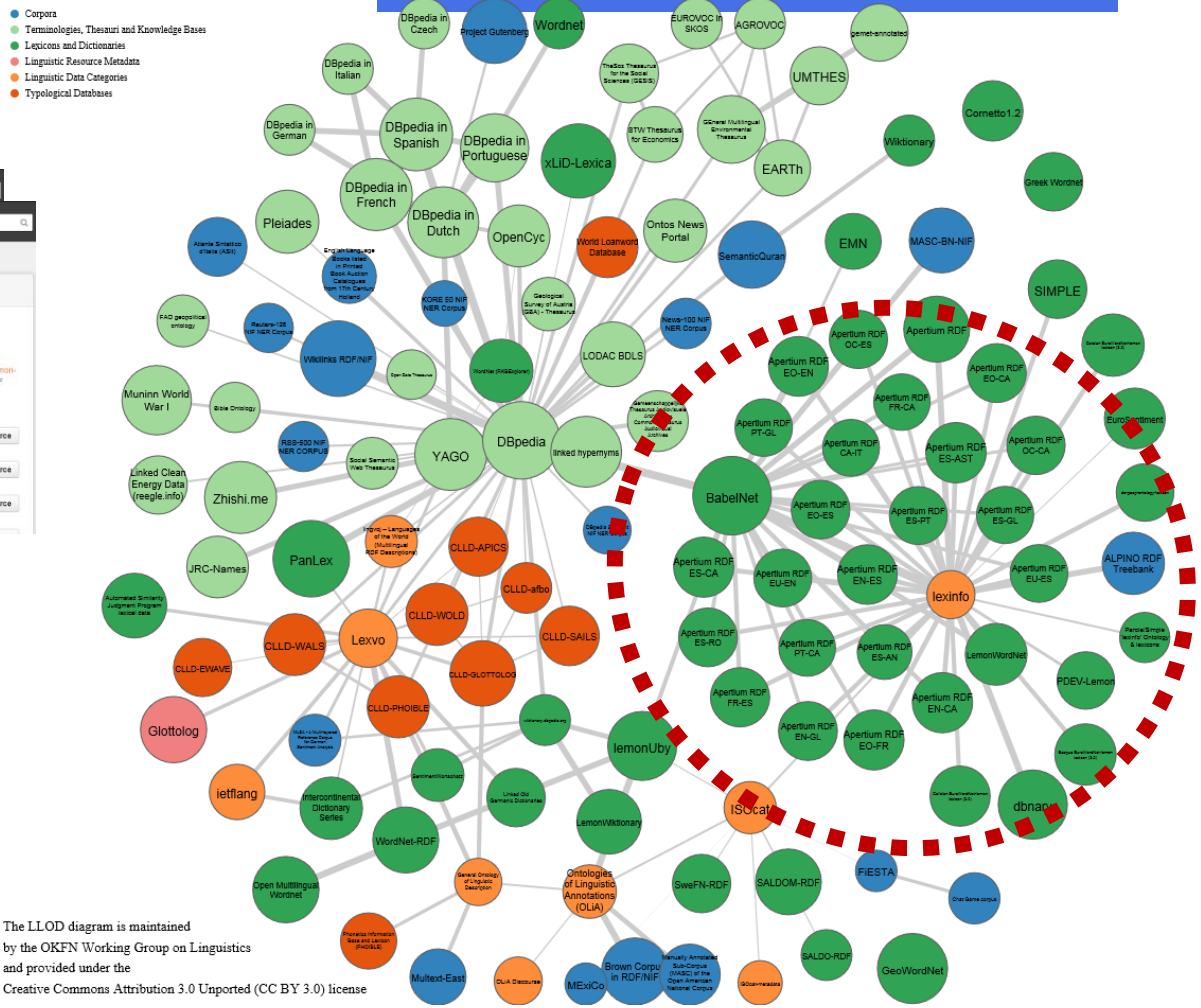
Open data in linguistics

RDF on the Web ★★★★

22 Apertium RDF



Linguistic Linked Data ★★★★★



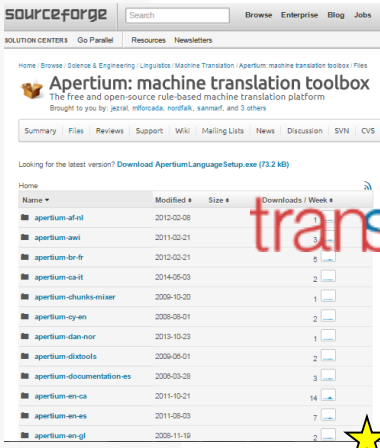
The LLOD diagram is maintained by the OKFN Working Group on Linguistics and provided under the Creative Commons Attribution 3.0 Unported (CC BY 3.0) license

mylider

www.lider-project.eu

Open data in linguistics in many formats

+40 Apertium Open bilingual Dictionaries proprietary format



transducens

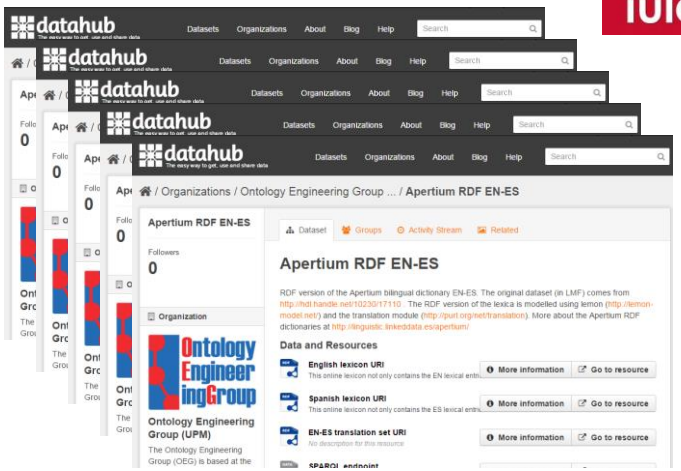
22 Apertium bilingual Dictionaries in LMF

upf. iula

690K links with LexInfo
277K links with Babelnet

Ontology Engineering Group

22 Apertium RDF



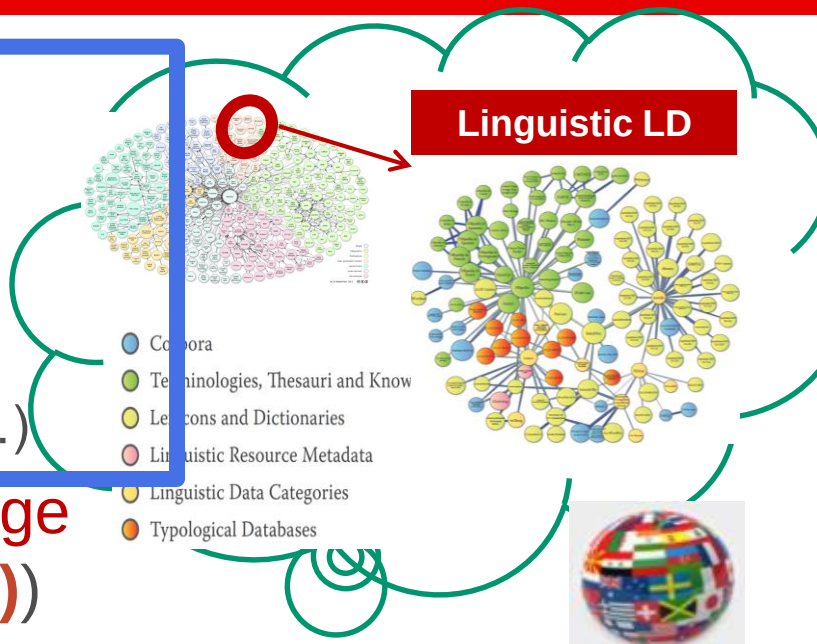
upf. iula Ontology Engineering Group

Muchas decisiones de diseño se tienen que hacer cuando se publiquen, consuman, enlacen datos abiertos de interés lingüístico con datos abiertos de interés general

Por ejemplo …

Ventajas de tener datos enlazados

1. **Agree on vocabularies** for describing
   - Domain datasets
   - LR metadata
   - LR content (Lemon-Ontolex, NIF, …)
2. Unified and standardized language for describing resources (**RDF(S)**)
3. Unified and standardized **query language** (SPARQL)
4. Standardized non-proprietary APIs
5. Links to other resources



mylider
www.lider-project.eu

Requirements for Language infraestructure

1. Platform development based on powerful set of **APIs, metadata, corpus, language resources in different formats** to ease the development of applications in multiple sectors

2. Evaluation facilities for **automatic benchmarking** any kind of technologies

3. **Technology Recommendation Framework** to select the best in class algorithm for solving a particular problem

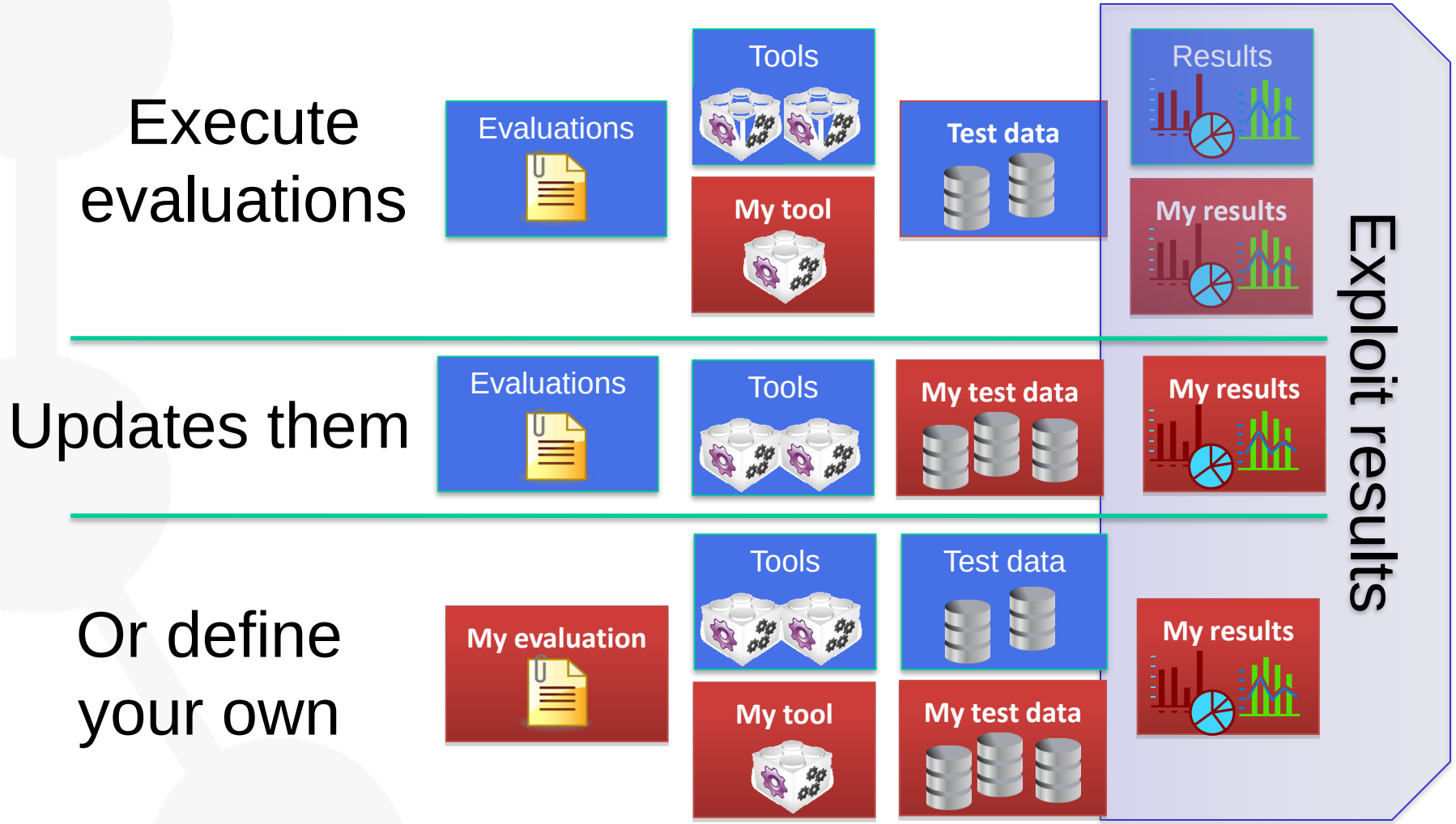
4. **Virtual lab for training** undergraduate, master and Ph.D students using the platform

5. **Building a community** of researchers, students, developers and end-users adopting and contributing to the platform

6. **Flexible business models** based on conditional access to services, knowledge and data (IPR, licenses, combination of open data and closed data, privacy, etc.

7. **Accelaration programm** to support SMEs in the use of the platform

Benchmarking: Evaluation of linguistic resources and services at scale



Datos abiertos de Interés Lingüístico

Prof. Dr. Asunción Gómez-Pérez

Artificial Intelligence Department

Universidad Politécnica de Madrid
Campus de Montegancedo sn
28660 Boadilla del Monte, Madrid

http://www.oeg-upm.net

asun@fi.upm.es
Phone: 34.91.3367439
Fax: 34.91.3524819