Maximising (Re)Usability of Language Resources using Linguistic Linked Data

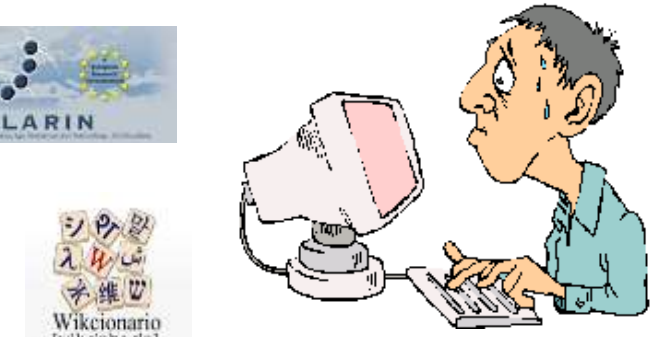
A. Gómez-Pérez

Universidad Politécnica de Madrid

asun@fi.upm.es

Lack of interoperability of Language resources

- Finding and reusing LR in third party applications is manual and time consuming
- Ecosystem of
  - Open and Closed resources
  - Different Languages
  - Silos of LRs
  - Complementary resources
    - Lexicon, Corpora, Dictionaries, Grammars, ….
  - Heterogeneous formats
    - E.g, for Lexicons: Lexinfo, LMF, LIR, Lemon, …
  - Several repositories with different metadata and schemas
  - Many APIs and services for querying



Discovery and reuse LR in third party applications is hard, manual and time consuming

Wikcionario

"Red"

Pronunciation: [red]

Grammar category: sustantivo femenino

Singular: "red"

Plural: "redes"

"Red"

Complementary but not connected

Wikilengua del español

"Red"

Norm: UNE 21302-131

English: network

German: Netzwerk

"Red"

Etimologiy Del latin "rete"

Gender: "f"

Definition.: "Conjunto de ordenadores o de equipos informáticos conectados entre sí…."

"Red_de_computadores"

Category: redes informáticas

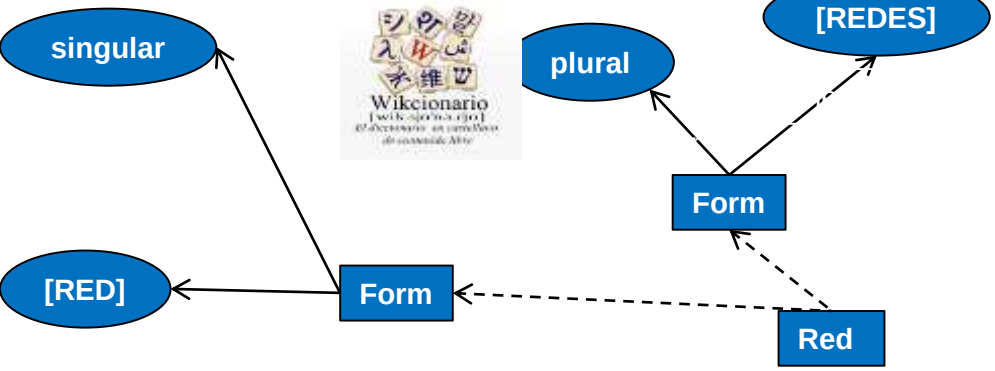
Image

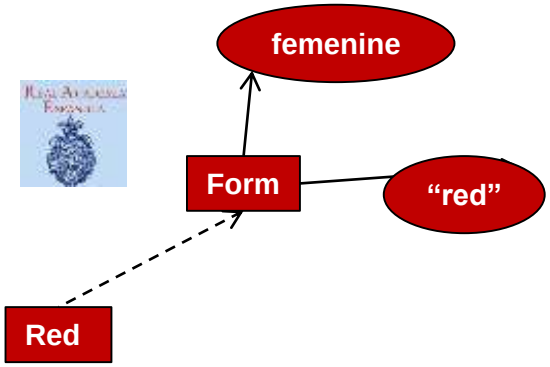
"Red"

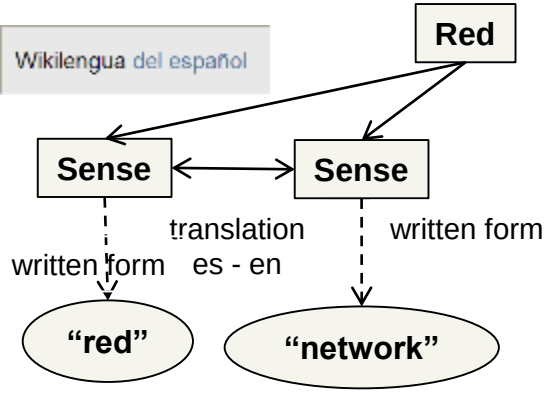
Sinonyms: "sistema", "malla"," distribución"

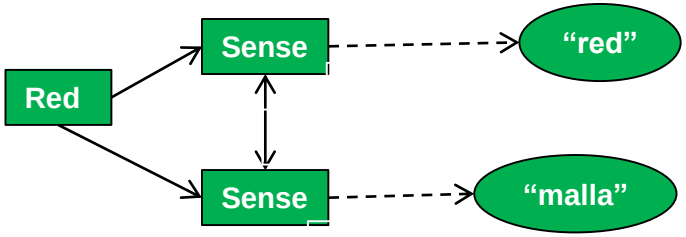
attribution: http://commons.wikimedia.org/wiki/User:Gugerell

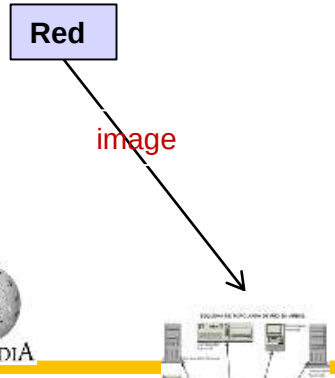
LD allows linguistic data integration





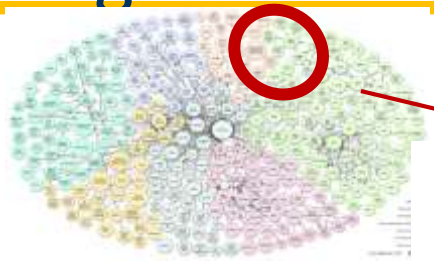




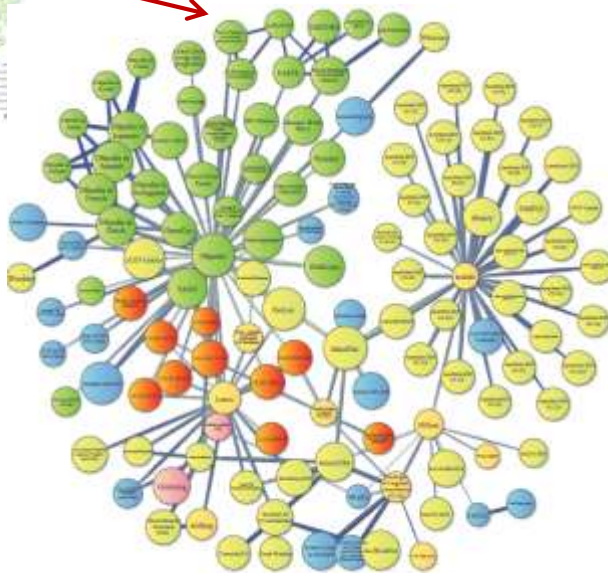


Linked Data

Linked Data and Linguistic Linked Data Clouds



Linguistic Linked Data



Linked Data interconnects data from resources

- In many domains
- In many languages
- Open and closed License
- Links with other datasets

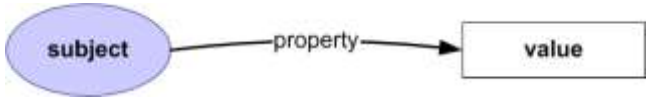
Linguistic Linked Data

- Subset of LOD
- Linguistic domain
- Many type of resources
- Interconnected with other LR
- Enables the lexicalization of data on the web, not necessarily data in the LD format
- Enables a new generation of LD-aware NLP and MT Services

- Corpora
- Terminologies, Thesauri and Knowledge Bases
- Lexicons and Dictionaries
- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases

Linked Data Foundations

RDF(S) models



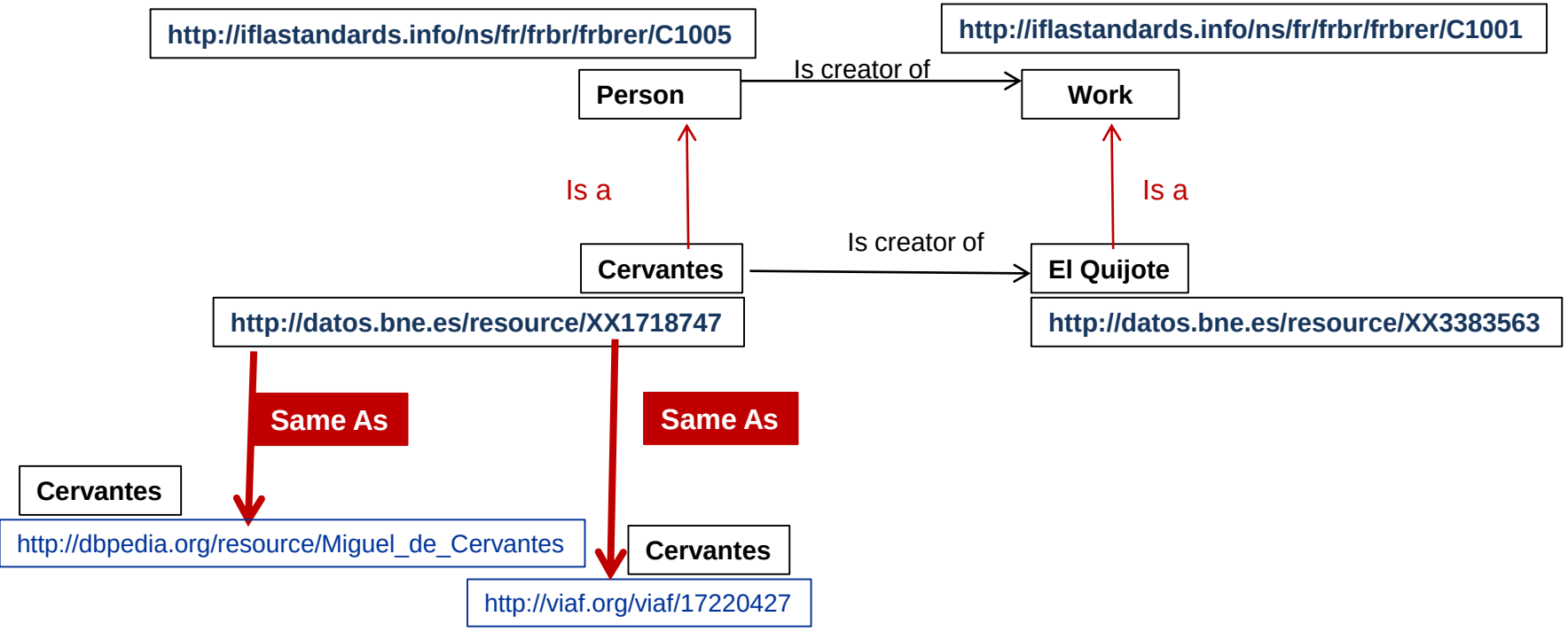
Unique identifiers: URI

identify or name a resource

Equivalence links to other datasets

Same As

Data navigation



Linguistic Linked Licensed Data

3LD

Linguistic **Linked** **Licensed** **Data**

Language resources such as:

- *Lexica*
- *Corpora*
- *Dictionaries ..*

*Using **RDF** and **standard data models** (vocabularies):*

- *Lexica* lemon
- *Corpora* **NIF** NLP Interchange Format

*Published along with a **machine-readable license***

ODRL
Open Digital Rights Language

creative commons

Linked data for machines: a world of URIs

Language

http://iflastandards.info/ns/fr/frbr/frbrer/C1002

translation

work

Is creator of

Person

Año

http://iflastandards.info/ns/fr/frbr/frbrer/C1001

http://iflastandards.info/ns/fr/frbr/frbrer/C1005

Publication date

birthPlace

Has subject

http://geo.linkeddata.es/ontology/Municipio

Located in

Biblioteca

http://xmlns.com/foaf/0.1/Organization

Ontology

Data

Catalán

http://datos.bne.es/resource/XX1924295

http://geo.linkeddata.es/resource/Alcalá de Henares

translation

Don Quijote de la Mancha

Cervantes Saavedra, Miguel de

Es autor

birthPlace

1960

http://datos.bne.es/resource/XX3383563

http://datos.bne.es/resource/XX1718747

Publication date

Has subject

Located in

http://datos.bne.es/resource/bimo0002045496

BNE

Vida de Miguel de Cervantes Saavedra

http://datos.bne.es/#

Linked Data without ontologies

http://www.server1.org/resource/Cervantes

Same as

http://d-nb.info/gnd/11851993X

Same as

http://datos.bne.es/resource/XX1718747

Same as

http://www.server2.es/resource/Cervantes

Same as

http://geo.linkeddata.es/page/resource/Municipio/Cervantes

URI
URI
URI
URI
URI

Cervantes

Author

D. Quijote

Date of Birth

1547

#People

1547

Size

276,4 km²

Phone

914 296 093

Linked Data needs ontologies

rdf:type → Retaurant

http://www.server1.org/resource/Cervantes

URI
UR.
URI
URI
URI

Cervantes (Person)

rdf:type → Person

http://d-nb.info/gnd/11851993X

rdf:type

Same as

http://datos.bne.es/resource/XX1718747

rdf:type → Street

http://www.server2.es/resource/Cervantes

Author

D. Quijote

Date of Birth

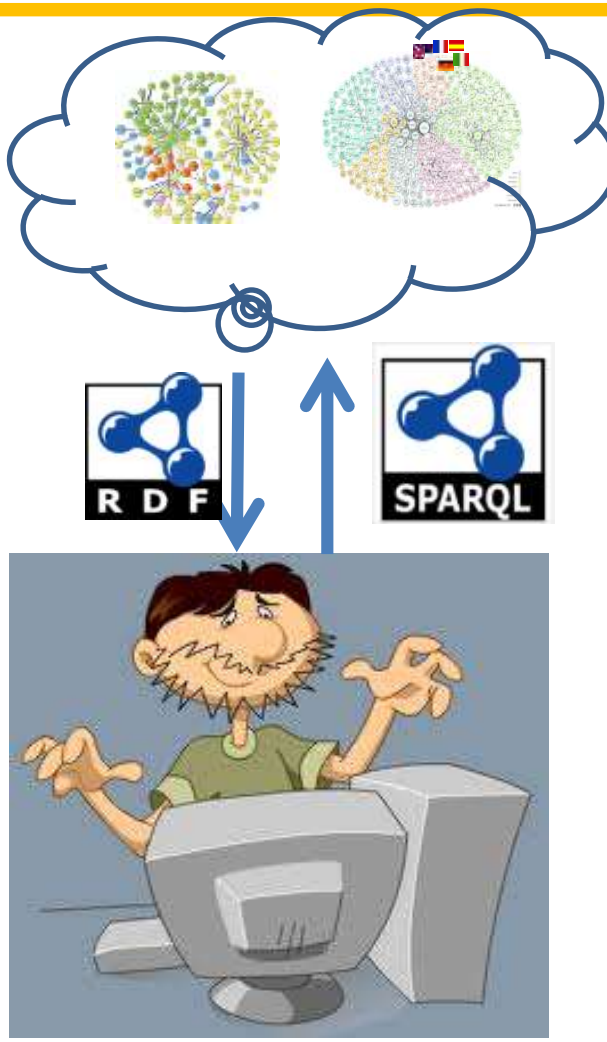
1547

rdf:type → Municipality

http://geo.linkeddata.es/page/resource/Municipio/Cervantes

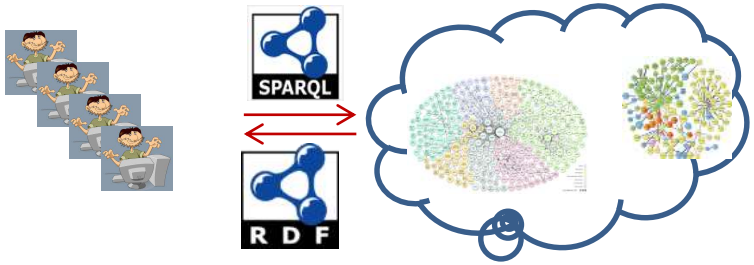
Linguistic Linked Licensed Data

1. **Agree on vocabularies** for describing
   - Domain metadata and data
   - Language resource metadata and content
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
6. Lexicalized data by using Linguistic Linked data



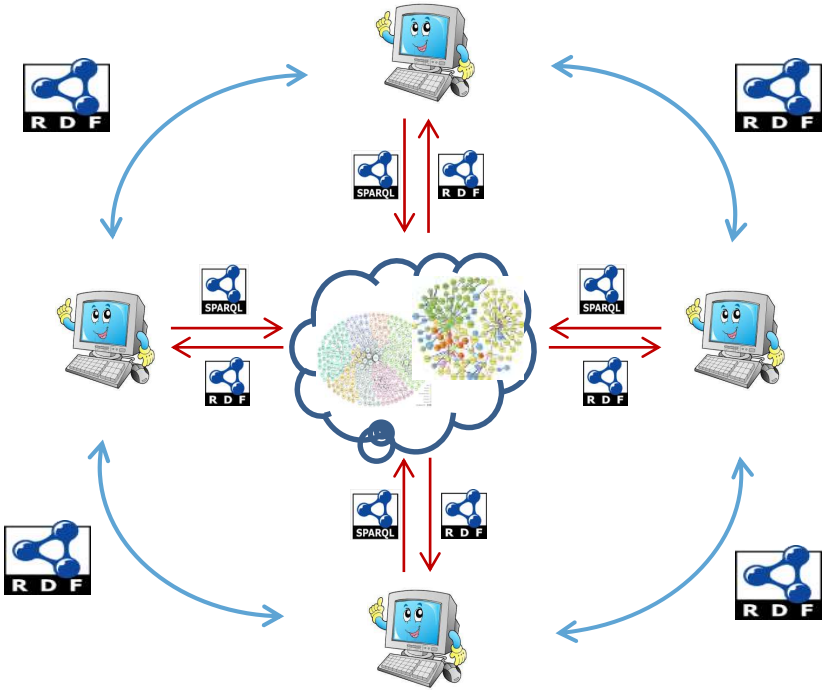
Uses and Users of Linked Data

1. Programmers built applications making queries in SPARQL and get RDF

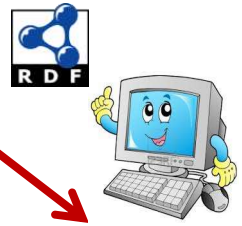


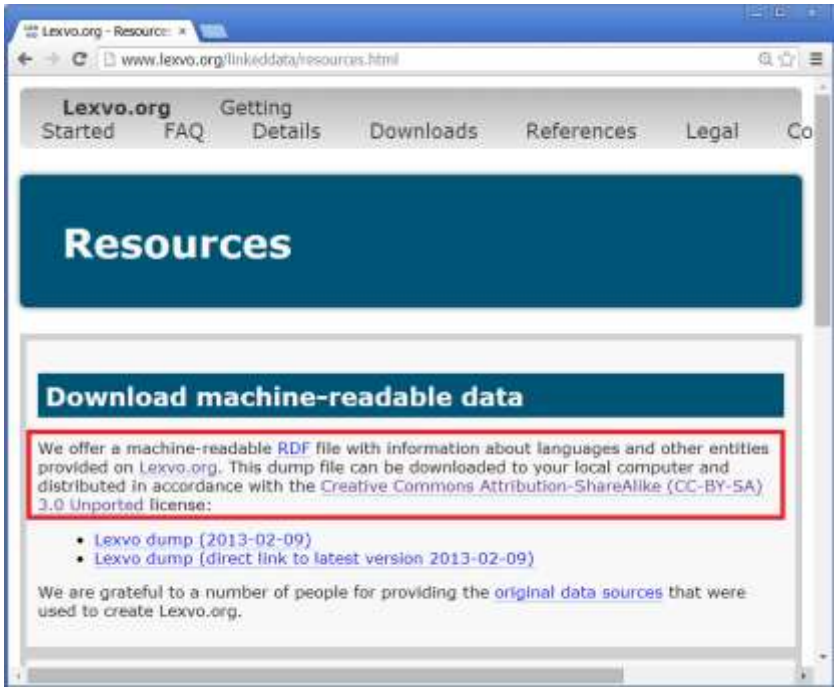
2. Citizens/Users access LD through a user interface (they do not see RDF)

3. Machine – Machine data exchange and semantic interoperability in RDF

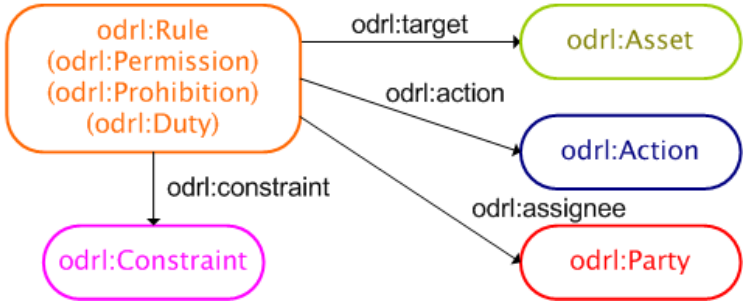


How do we represent license information?





Rights expression in ODRL:
Who [can|cannot|must] act what in which resource how



Computers understand each other and do business in the DSM