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
“Red”

Pronunciation: [red]

Grammar category: sustantivo femenino

Singular: “red”

Plural: “redes”

Complementary but not connected

Norm: UNE 21302-131

English: network

German: Netzwerk

“Red_de_computadores”

Category: redes informáticas

Image

“Red”

Etimologíy Del latin “rete”

Gender: “f”

Definition.: “Conjunto de ordenadores o de equipos informáticos conectados entre sí....”

Sinonyms: “sistema”, “malla”, “distribución”
LD allows linguistic data integration

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

Asunción Gómez-Pérez

CEF - Towards a Connected Multilingual Europe @ Riga Summit. 29th April, 2015. Riga
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

---

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

CEF - Towards a Connected Multilingual Europe @ Riga Summit. 29th April, 2015. Riga
RDF(S) models

Unique identifiers: URI
identify or name a resource

Equivalence links to other datasets
Same As

Data navigation

http://iflastandards.info/ns/fr/frbr/frbrer/C1005
Is creator of
Person

http://iflastandards.info/ns/fr/frbr/frbrer/C1001
Is creator of
Work

http://datos.bne.es/resource/XX1718747
Is a
Cervantes

http://datos.bne.es/resource/XX3383563
Is a
El Quijote

http://iflastandards.info/ns/fr/frbr/frbrer/C1005
Is creator of
Person

http://iflastandards.info/ns/fr/frbr/frbrer/C1001
Is creator of
Work

http://datos.bne.es/resource/XX1718747
Is a
Cervantes

http://datos.bne.es/resource/XX3383563
Is a
El Quijote

http://iflastandards.info/ns/fr/frbr/frbrer/C1005
Is creator of
Person

http://iflastandards.info/ns/fr/frbr/frbrer/C1001
Is creator of
Work

http://datos.bne.es/resource/XX1718747
Is a
Cervantes

http://datos.bne.es/resource/XX3383563
Is a
El Quijote

http://iflastandards.info/ns/fr/frbr/frbrer/C1005
Is creator of
Person

http://iflastandards.info/ns/fr/frbr/frbrer/C1001
Is creator of
Work

http://datos.bne.es/resource/XX1718747
Is a
Cervantes

http://datos.bne.es/resource/XX3383563
Is a
El Quijote

http://iflastandards.info/ns/fr/frbr/frbrer/C1005
Is creator of
Person

http://iflastandards.info/ns/fr/frbr/frbrer/C1001
Is creator of
Work

http://datos.bne.es/resource/XX1718747
Is a
Cervantes

http://datos.bne.es/resource/XX3383563
Is a
El Quijote

http://iflastandards.info/ns/fr/frbr/frbrer/C1005
Is creator of
Person

http://iflastandards.info/ns/fr/frbr/frbrer/C1001
Is creator of
Work

http://datos.bne.es/resource/XX1718747
Is a
Cervantes

http://datos.bne.es/resource/XX3383563
Is a
El Quijote
Linguistic Linked Licensed Data

3LD

Linguistic Linked Licensed Data

Language resources such as:
- Lexica
- Corpora
- Dictionaries ..

Using RDF and standard data models (vocabularies):
- Lexica
- Corpora

Published along with a machine-readable license
ODRL
Open Digital Rights Language

NIF
NLP Interchange Format

RDF

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

CEF - Towards a Connected Multilingual Europe @ Riga Summit. 29th April, 2015. Riga
Linked data for machines: a world of URIs

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

CEF - Towards a Connected Multilingual Europe @ Riga Summit. 29th April, 2015. Riga

Asunción Gómez-Pérez
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

D. Quijote

Date of Birth

1547

Size

276.4 km²

#People

1547

Phone

914 296 093

Author

Cervantes
Linked Data needs ontologies

 RDF: type

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

 RDF: type

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

 RDF: type

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

 RDF: type

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

 RDF: type

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

 RDF: type

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

 RDF: type

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

 RDF: type

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

 RDF: type

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

 RDF: type

 Cervantes (Person)

 RDF: type

 Author

 RDF: type

 D. Quijote

 RDF: type

 Date of Birth

 RDF: type

 1547

 RDF: type

 Municipality

 RDF: type

 Street

 RDF: type

 Person

 RDF: type

 Restaurant
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
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)


Culture
Geographical
Smart Cities
How do we represent license information?

Rights expression in ODRL:
Who [can|cannot|must] act what in which resource how
Maximising (Re)Usability of Language Resources using Linguistic Linked Data

Computers understand each other and do business in the DSM

Asunción Gómez-Pérez