



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

A. Gómez-Pérez

Universidad Politécnica de Madrid

asun@fi.upm.es



- 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”



“Red”

Etimology: Del latín “rete”

Gender: “f”

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

Wikilengua del español

“Red”

Norm: UNE 21302-131

English: network

German: Netzwerk

“Red”

Complementary but not connected



“Red_de_computadores”

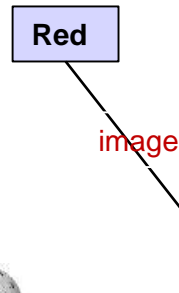
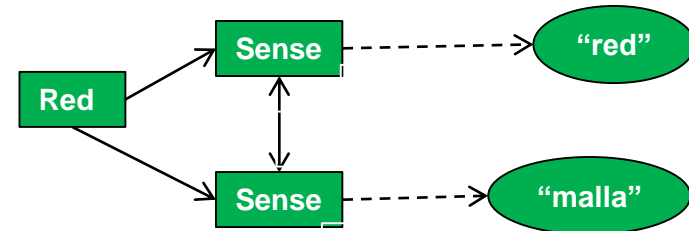
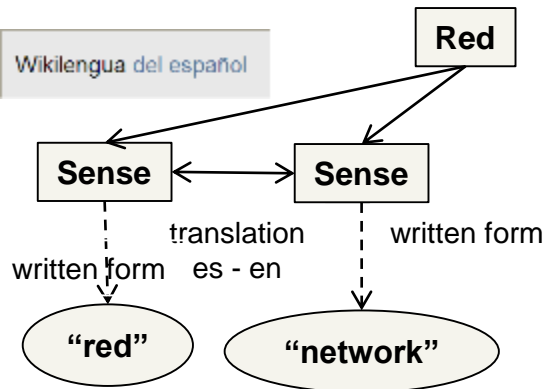
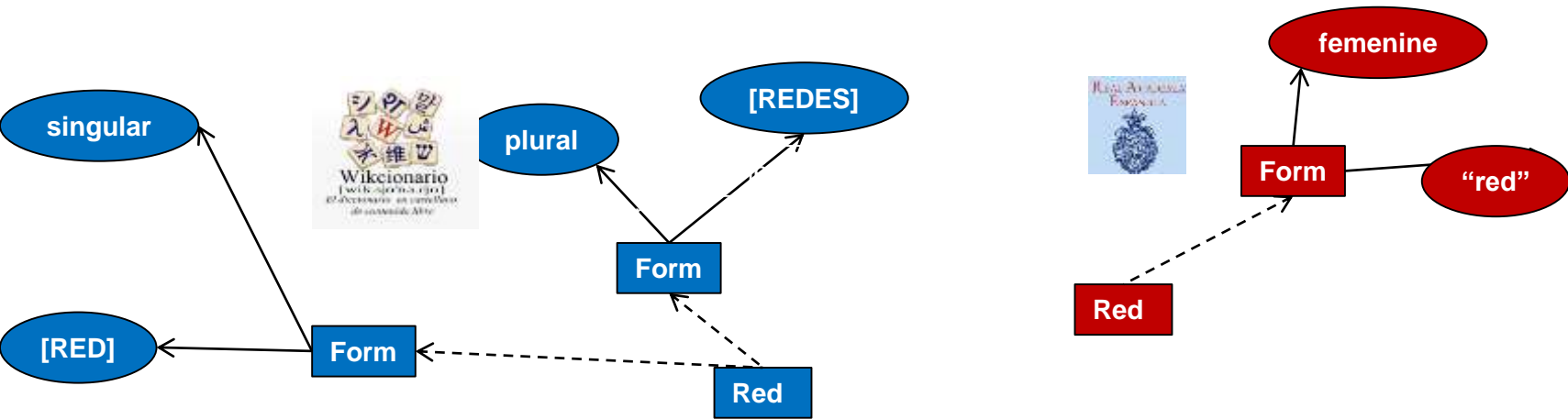
Category: redes informáticas

Image

“Red”

Synonyms: “sistema”, “malla”, “distribución”







Linked Data

Linked Data interconnects data from resources

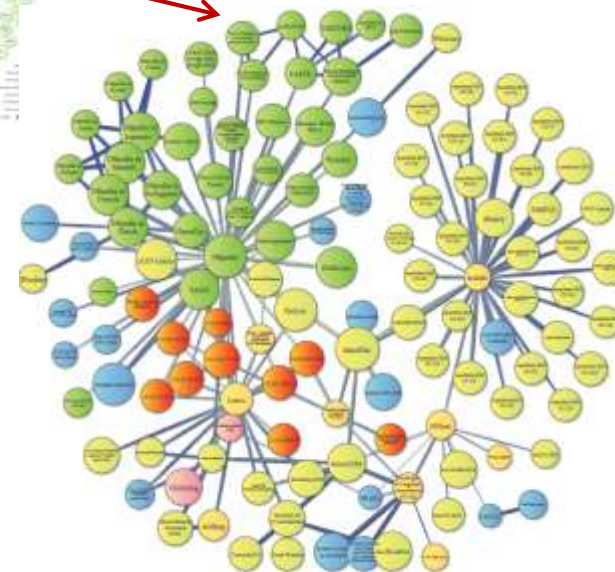
- In many domains
- In many languages
- Open and closed License
- Links with other datasets

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

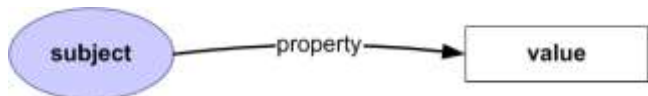


Linguistic Linked Data



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

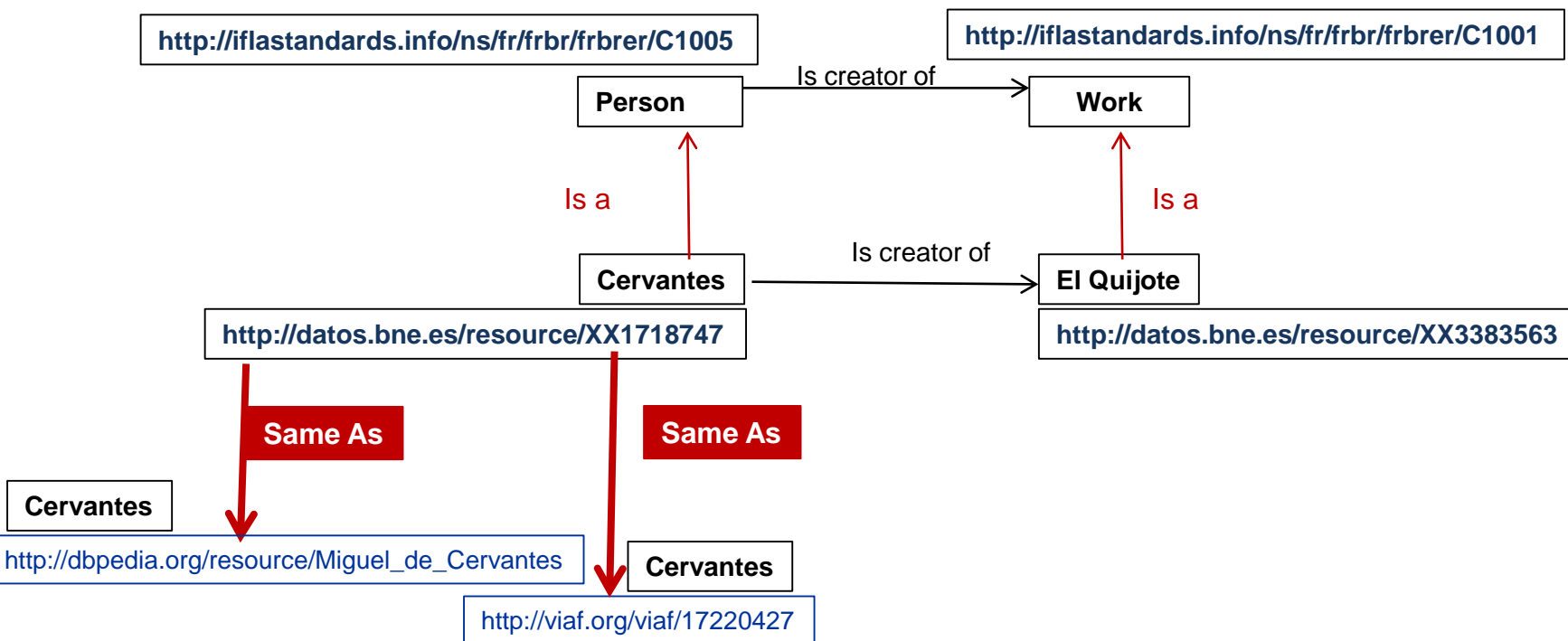
RDF(S) models



Unique identifiers: URI
identify or name a resource

Equivalence links to other datasets
Same As

Data navigation



3LD

Linguistic Linked Licensed Data

Language resources such as:

- Lexica
- Corpora
- Dictionaries ..

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

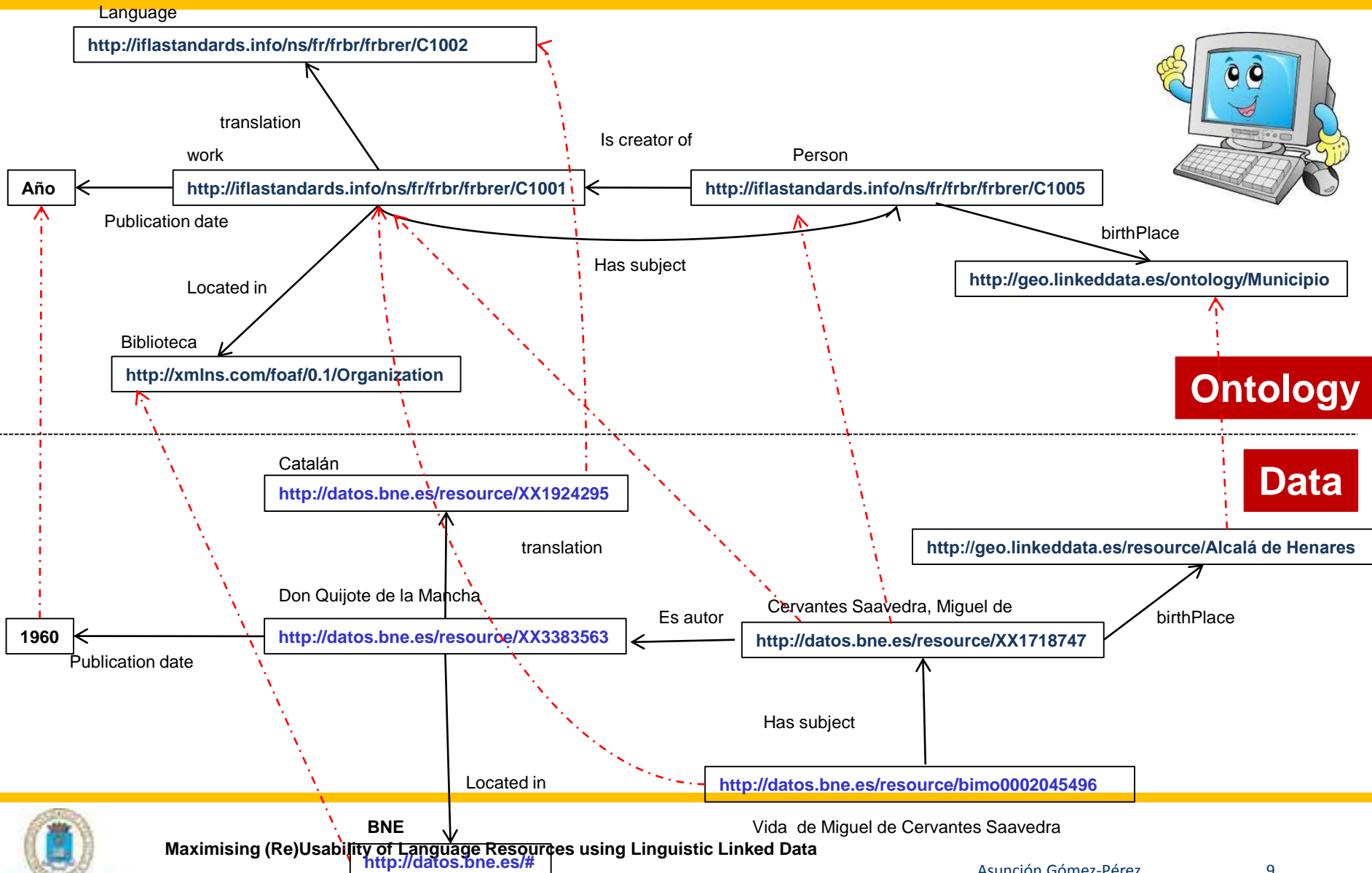


- Lexica 
 - Corpora 
- NIF Interchange Format

Published along with a machine-readable license

ODRL
Open Digital Rights Language





Ontology

Data

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



Cerveceria Cervantes



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

Same as



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

Same as



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

Same as

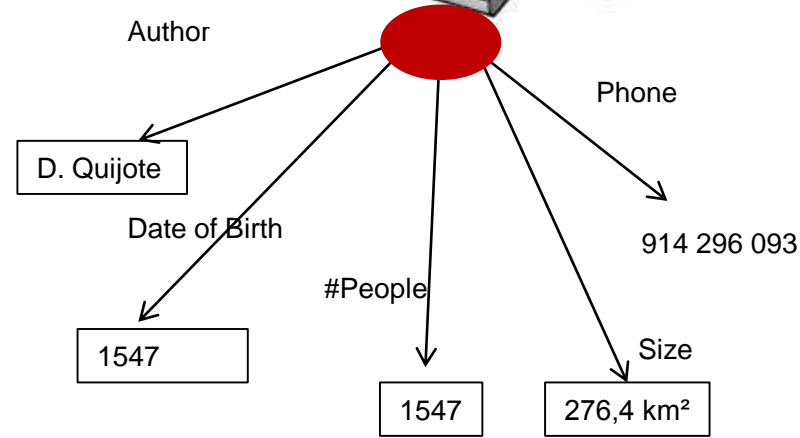
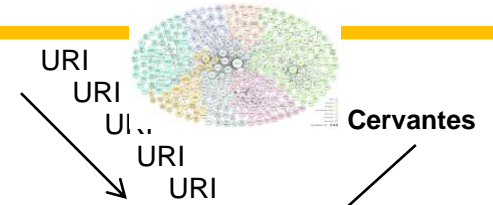


<http://www.server2.es/resource/Cervantes>

Same as



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



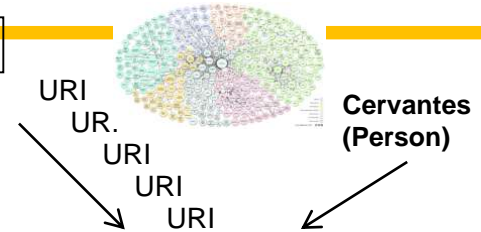
City of Language Resources using Linguistic Linked Data



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

rdf:type

Restaurant



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

rdf:type

Person

Same as



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

rdf:type



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

rdf:type

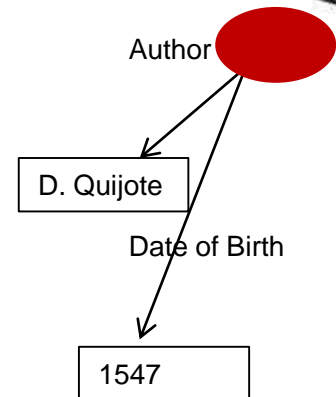
Street



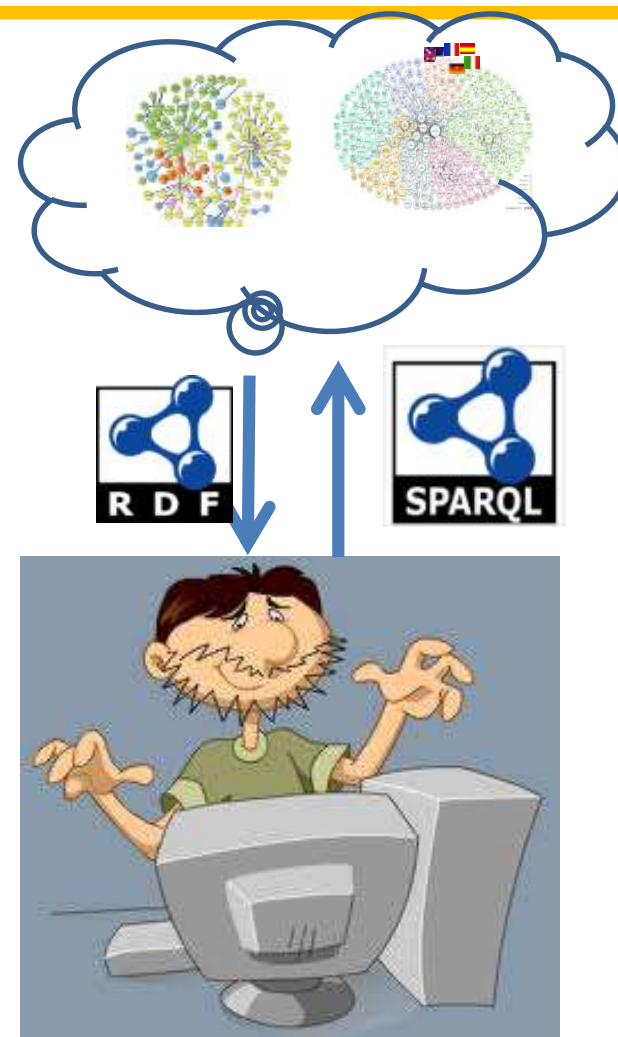
<http://geo.linkeddata.es/page/resource/Municipio/Cervantes>

rdf:type

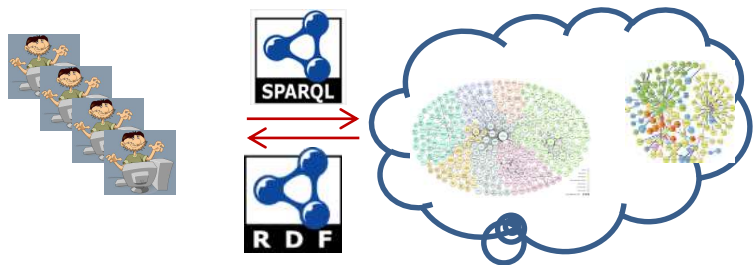
Municipality



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



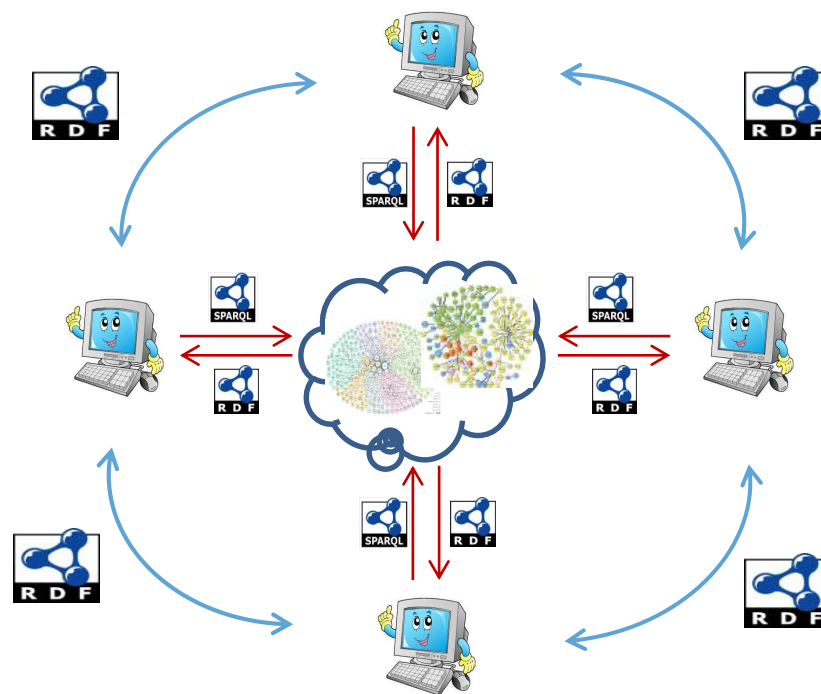
Geographical



Smart Cities



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

