Linking Lexical Resources

The Opportunities and Challenges Offered by the Semantic Web

Part I

Fahad Khan - Institute for Computational Linguistics “A. Zampolli”

elexis
european lexicographic infrastructure

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Introduction

This tutorial is intended as an introduction to the possibilities for inter and (intra) dataset linking that the Semantic Web offers to anyone thinking of publishing their own linguistic resources as linked data. In particular it’s aimed at those who are working with digital lexicons, including wordnets and retrodigitized dictionaries. Our aim in this talk is to **conceptually foreground the notion of a link in lexical linked data resources.**

The session will be broken up into two parts. In the **first part** I will go through some of the basics of the Semantic Web as they pertain to the question of the linking together of resources and/or datasets. I will also explore some more concrete use cases.

In the **second part** Andrea will show some actual examples of linking using some real life examples using the tool which he has developed LexO -- an interface for creating and editing linked data lexicons.

The project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015

Introduction

We will only assume a minimal amount of background knowledge on linked data and the Semantic Web in this talk.

However I will cover quite a bit of material.If anything isn't clear, please feel free to interrupt us and ask questions!

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

A Quick Rundown on Some Semantic Web Basics

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Definitions

Linked Data - a method of publishing structured data so that it can be **interlinked** and become more useful through **semantic queries**. (Source: Wikipedia)

The Semantic Web - a web of datasets that are structured and **linked together using a common set of standards and technologies** so that they can be **more easily processed by computers in terms of what they 'mean' --** in contrast to normal hypertext documents which are designed to be read by humans.

Linked Data is one very important way of making the semantic web a reality.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

The Core Linked Data Principles

In 2006, Tim Berners-Lee stated the four guiding principles for publishing data as linked data. These are:

1. *Use **URIs** as names for things*
2. *Use **HTTP URIs** so that people can look up those names.*
3. *When someone looks up a **URI**, provide useful information, using the **standards** (**RDF, SPARQL**)*
4. *Include links to **other URIs**. so that they can discover more things.*

If, in addition to these four pre-requisites, we make our data available under an open license then it is classified as **Linked Open Data (LOD)**.

These principles are meant to encourage both the **maximum of interoperability** between datasets and to facilitate a **more explicit encoding of meaning** within and between datasets.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Resource Description Framework

The **Resource Description Framework (RDF)** is the standard way of modeling data on the Semantic Web. Crucially RDF makes the constraint that we must describe our data using **only** statements of the form:

Subject- Predicate-Object

Where the **Subject** and **Object** are each resources (the ‘conceptual things’ of the last slide) with their own URIs (the object can also potentially be a literal data value like a string or an integer).

These two subject/predicate resources are related together by the **Predicate,** which is often called a ‘property’ or ‘role’ and which is also a resource with its own URI. These statements are known as **RDF-triples** and a linked data dataset consists of a series of such triples.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Resource Description Framework

The resources referred to in the last slide can be either: **individuals** (*Ljubljana, Slovenia, **Andrea Bellandi**, the Universe, Superman*), **classes** (*the class of events, the class of cities, the class of overcast, dreary Sundays*) or **relations** (*X loves Y, X is bigger than Y, X is a lexicographer*).

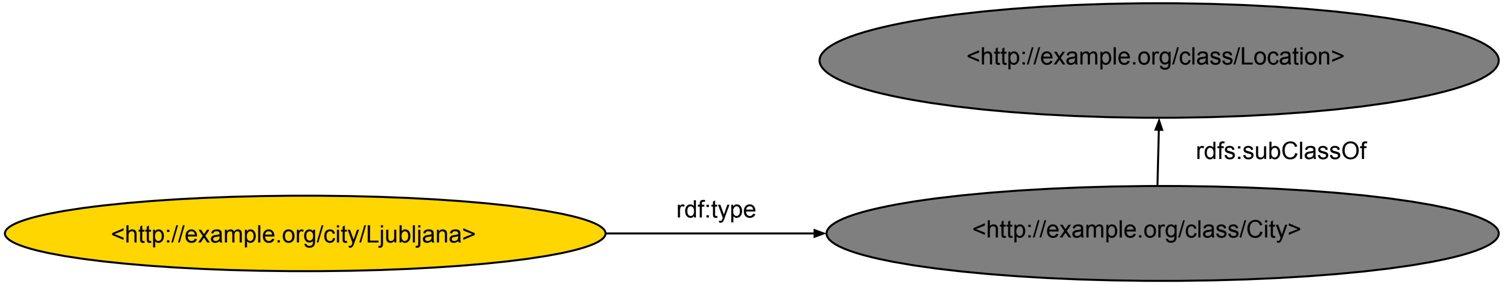
We will often find the classes or individuals or properties we're looking for in **other datasets** and can **re-use** these (if the semantics are the right ones). If we can't find what we need amongst already existing vocabularies then we can create our own.

In essence then RDF allows us to **build up graphs** linking together resources. These graphs are distributed across different web locations and we can access them using the HTTP protocol.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

RDF and RDFS

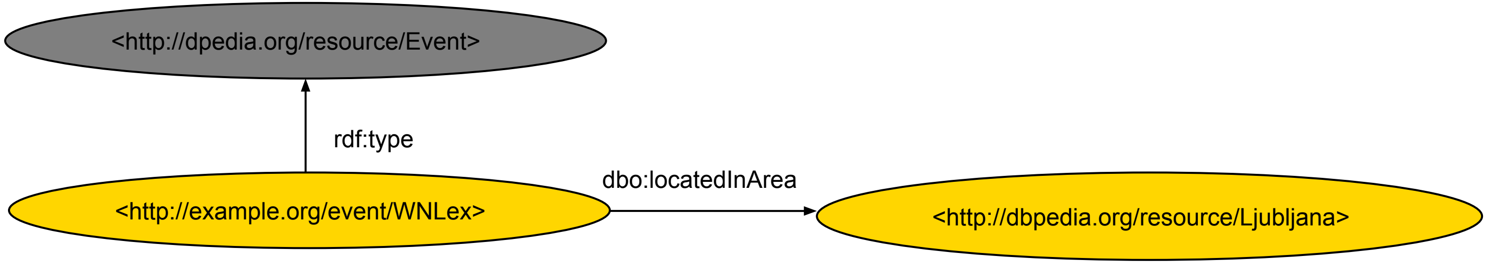
These built-in properties essentially **put a constraint on how we build our graphs**. They allow for more concision in our data too. For instance, using a reasoner on the following graph we will be able to derive the fact that **Ljubljana is a Location**, even if that isn't explicitly stated in the data itself.


<http://example.org/class/Location>
rdfs:subClassOf
rdf:type
<http://example.org/city/Ljubljana>
<http://example.org/class/City>

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Re-using Individuals and Classes

In the previous diagrams we defined our own individuals and classes but of course as we mentioned before it is *preferable to* **reuse** those that have already been defined in other datasets and vocabularies. This way we help to ensure that the Semantic Web remains as **interconnected as possible**. It has the effect though that the **meaning of concepts is often 'deferred'** or rather distributed across the network.


<http://dpedia.org/resource/Event>
rdf:type
dbo:locatedInArea
<http://example.org/event/WNLex>
<http://dbpedia.org/resource/Ljubljana>

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Interoperability and Shared Vocabularies

By using **shared vocabularies** and **standard identifiers** to describe our data it becomes much easier to integrate with other datasets (since these datasets are using the **same names** to refer to **the same things**) thus opening up the possibility of easily querying large numbers of datasets at the same time. Note that in SPARQL we have a powerful query language for linked data datasets which we can make publically queryable by exposing them using SPARQL endpoints.

The website:

http://lov.okfn.org/dataset/lov/

is hosted by the **Open Knowledge Foundation** and provides a search engine facility for linked data vocabularies

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Specifying the ‘Meanings’ of Properties

In addition to the built in properties mentioned above RDFS allows us to:

- Specify that **one property is a subproperty of another** using **rdfs:subPropertyOf**. This allows for the easy creation of taxonomies of properties
    - *i.e., **:hasSister rdfs:subPropertyOf :hasSibling***
- Specify the **domains and ranges of properties** (that is the classes of things that a property can potentially relates together) using the properties **rdfs:domain** and **rdfs:range**
    - *i.e., the domain and range of **:hasSibling** is the class **:Human**, on the other hand we could, if we wanted, set the domain of **:hasMother** as **:Human** and the the range to be the subclass, **:FemaleHuman** of **:Human***

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Specifying the ‘Meanings’ of Properties

Together with RDF/S’s other built-in properties that deal with classes and individuals, and those relating to so-called datatype and annotation properties, we are already able to do quite a lot when it comes to formally specifying (non trivial aspects of) the meanings of properties.

It should be clear by now how different the basic conception of modelling data in linked data is from a standard like TEI. Our datasets consist of **a series of triples representing declarative subject predicate object statements** and should be viewed as **graphs with edges with meaningful labels** -- rather than as trees as again in TEI. Of course this means it’s harder (or more verbose) to do things such as creating digital editions of texts in linked data, but it makes other things much easier (anything where we’re basically representing networks!).

The project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015

SKOS and OWL

We have even more expressivity with the **OWL (Web Ontology Language)** which is a formal knowledge representation language that builds on top of RDFS and that allows us to break down classes/properties into more basic components.

Simple Knowledge Organization System (SKOS) is a more lightweight option for those who are more interested in creating taxonomies of concepts that follows on from RDFS's built in properties. But going into more detail on either these would take us too far afield.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

rdfs:seeAlso and owl:sameAs

It’s important to know about the properties **seeAlso** and **sameAs.** The first of these is a built-in property of RDFS and the other a built in property of OWL. They’re two of the most generic properties that we can use to link together linked data resouces :

- **seeAlso**: this property allows us to state the fact that the object of the property contains information that’s (somehow) relevant to the subject.
    - <http://dbpedia.org/resource/Ljubljana> rdfs:seeAlso <http://dbpedia.org/resource/Slovenia>
- **sameAs**: this property simply states that two URIs actually refer to the same thing:
    - <http://dbpedia.org/resource/Ljubljana> owl:sameAs <http://it.dbpedia.org/resource/Lubiana>

The project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015

Summing Up

Summing up then we have the following:

- The Semantic Web and Linked Data enable the creation of distributed knowledge graphs in which the links between concepts/resources across datasets can be assigned meanings using formal languages of different levels of expressivity.

This should suggest the necessity of a change of mindset when it comes to modelling datasets as linked data at least in comparison with standards like **TEI**. In fact the Semantic Web seems to draw us towards a more classical **AI/Knowledge Engineering** approach. In the rest of my part of the talk I want to focus on the opportunities and challenges which the modeling of lexical resources as linked data throws up. In particular I want to look at how we can enrich individual lexical resouces by linking them to other resources, lexical or not.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Linking Lexical Datasets on the Semantic Web

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Moving onto Lexical Datasets

Let’s move on now and look at how we can apply some of the things we’ve just seen to lexical datasets. That is, let’s focus on the questions: **What kinds of links would be particularly interesting or helpful for consumers of lexical RDF datasets**? **How can we best augment lexical datasets with the kinds of links we saw earlier**?

In what follows I’m going to list a number of what I think are the most interesting and usefuls kinds of links that we can add to lexical datasets. I’m going to do this by highlighting two sets of use cases. The first set of use cases deals with links between lexical datasets. The second set of use cases deals with links between lexical datasets and other kinds of datasets.

The project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015

Links to Lexical Linked Datasets

Unfortunately due to time constraints I won't be able to talk about the very important topic of links **to** linked data lexicons from other datasets e.g., as in the case of documents annotated with lexical entries or word senses from linked data datasets.

However this is a very important application of lexical linked data datasets.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Enriching Lexical Linked Datasets

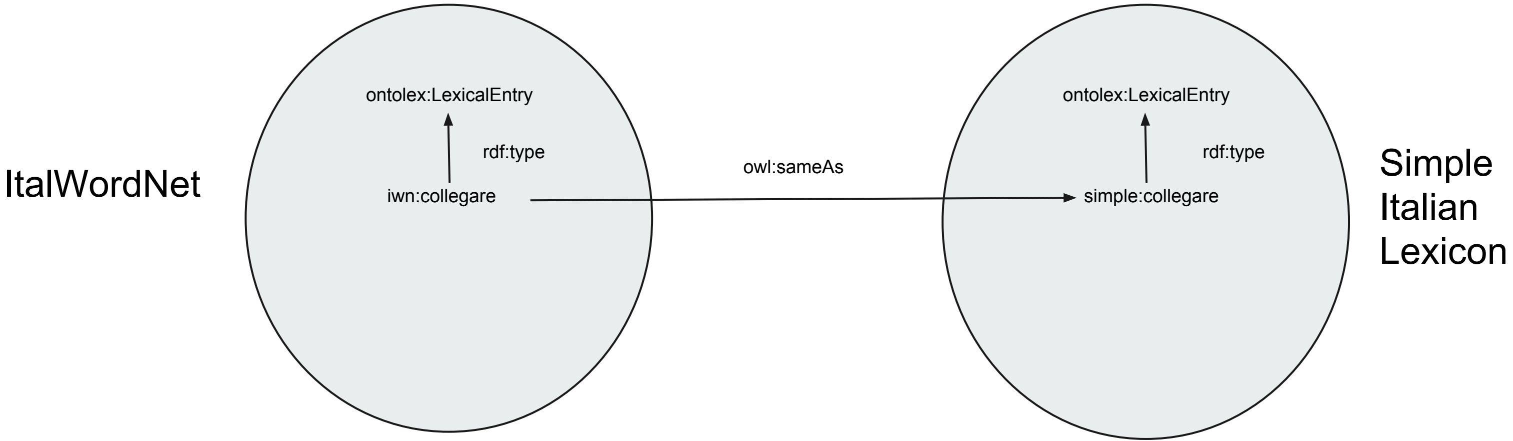
What then are some interesting use cases for linking individual lexical datasets together? (Not an exhaustive list)

- **Creating links between entries for the same word in different lexicons in order to make additional linguistic information accessible**; useful for wordnets which (usually) do not include much in terms of phonetic or morpho-syntactic information.
- **Creating bi/multi-lingual lexical resources by linking together senses that are translations/equivalents of one another in two or more lexicons**.
- **Modelling etymological data and referencing words in other languages as part of individual word histories**.
- **Tracing the evolution of lexical entries over different versions of the same dictionary or lexicon** See for instance: http://nenufar.huma-num.fr/ (now live though not complete)

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Lexicon-Lexicon Linking

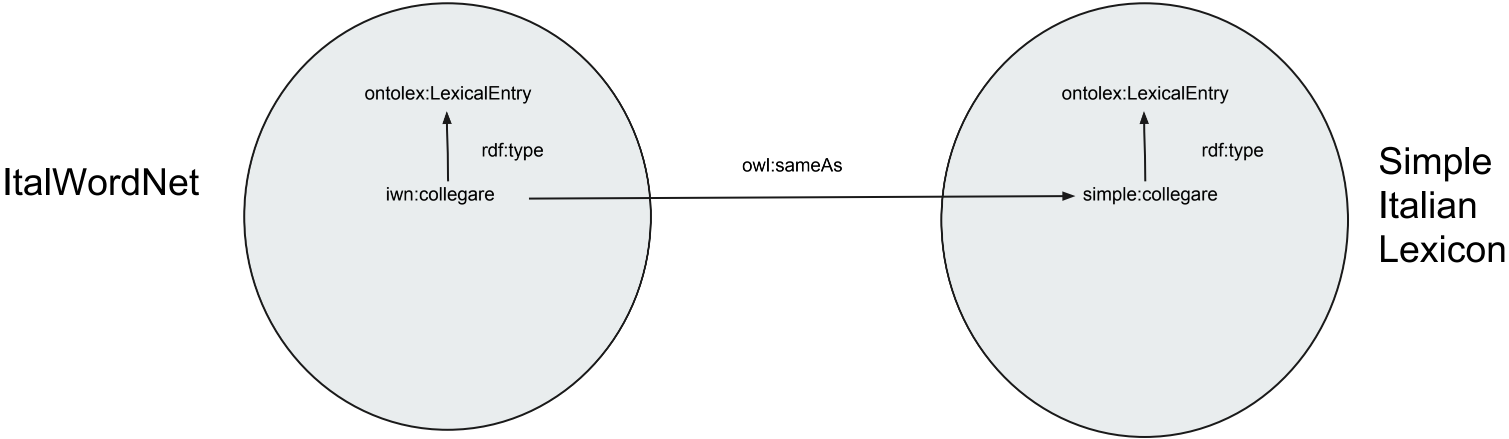
Perhaps the simplest kind of lexicon to lexicon linking that we can carry out is linking together entries for the same word but across different lexicons. In this case we can use the **sameAs** relation to do the linking.



The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Lexicon-Lexicon Linking

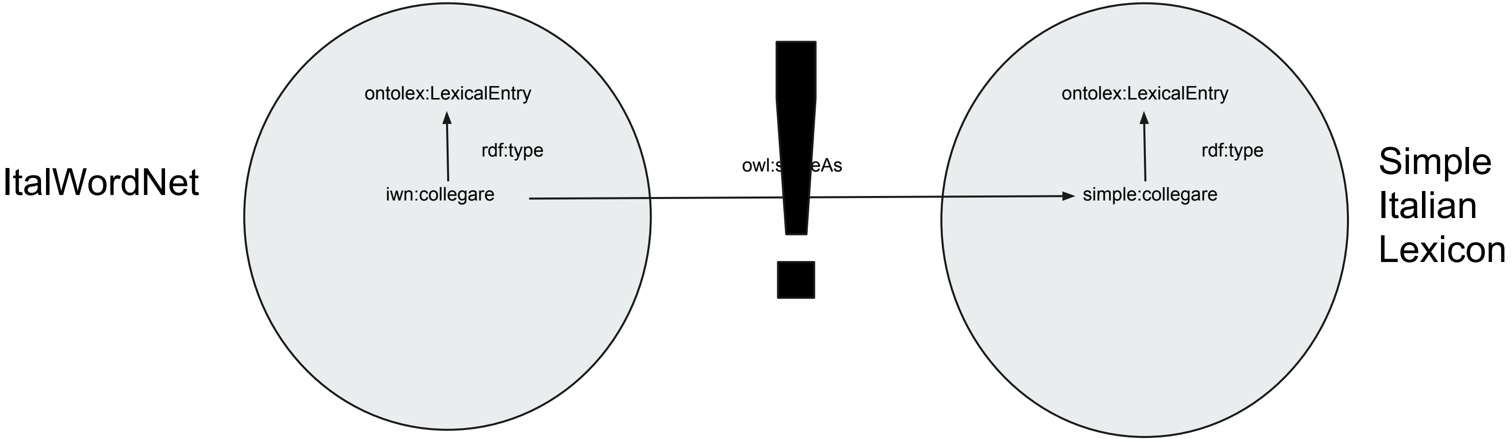
For instance linking *ItalWordNet* to a lexicon like *Simple* will allow us to access more comprehensive morpho-syntactic information contained in the latter from the former, and vice versa wrt the semantic information contained in IWN.



The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Lexicon-Lexicon Linking (Caveat Emptor)

...however the semantics of **owl:sameAs** may not always be appropriate in the case when the two lexical sources being linked together contain differing or contradictory information about the same entry (*referential opacity*). Careful curation is necessary and/or the use of another property (*sameEntry*?).



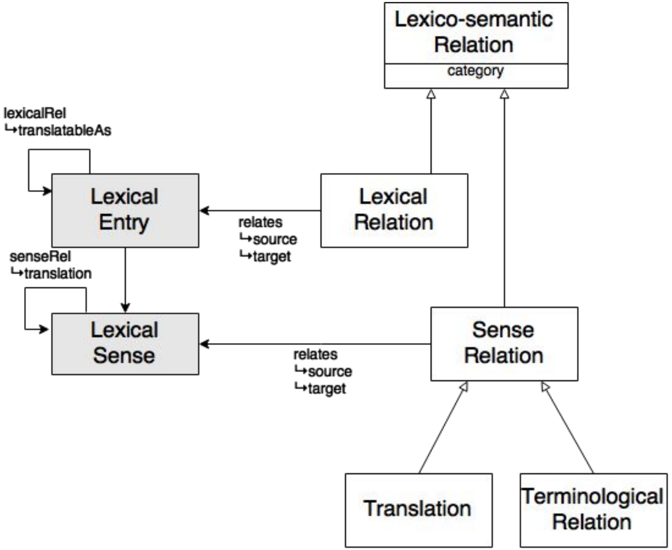
The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Lexicon-Lexicon Linking: Translations

We can also, in effect, create **bilingual and multilingual resources** by linking together the senses contained in monolingual lexicons (semi-)automatically.

Fortunately the **ontolex-lemon** vocabulary already provides a host of properties and classes to do this: https://www.w3.org/community/ontolex/wiki/Final_Model_Specification#Variation_.26_Translation_.28vartrans.29

Used in the linking together of various Apertium dictionary language pairs (Gracia et. at. 2017)



The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Lexicon-Lexicon Linking: Etymologies

Etymologies of words are **essentially descriptions of graphs,** they also very often involve different kinds of **historical** and more generally **non linguistic** information which makes them perfect candidates for being modelled as linked data, e.g., representing metaphorical sense shifts between meanings of words we can re-use vocabularies dealing with different types of **metaphor** and **metonymy.**

Generally, in order to represent the etymology of a word we need to create links between the entry for that word and entries in other lexicons or in the same lexicon, links that are typed for the specific kind of historical relationships obtaining between the different words (e.g., **inheritance, borrowing)**.

Etymological WordNet (De Melo 2014) an example of an already existing dataset that uses the principles of linked data to model etymological networks.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Lexicon-Lexicon Linking: Etymologies

The creation of such etymological links will be greatly assisted by the current push towards publishing lexical resources for historical languages such as **ancient Greek** and **Latin** in linked data. Wordnets for **Latin, ancient Greek**, and **Sanskrit** should already be available or will be made publically available soon(ish).

At the moment the ontolex-lemon group is working on developing an etymology module for lemon. Anyone interested in contributing to this module is cordially invited to participate by joining the ontolex mailing list.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Enriching Lexical Linked Datasets

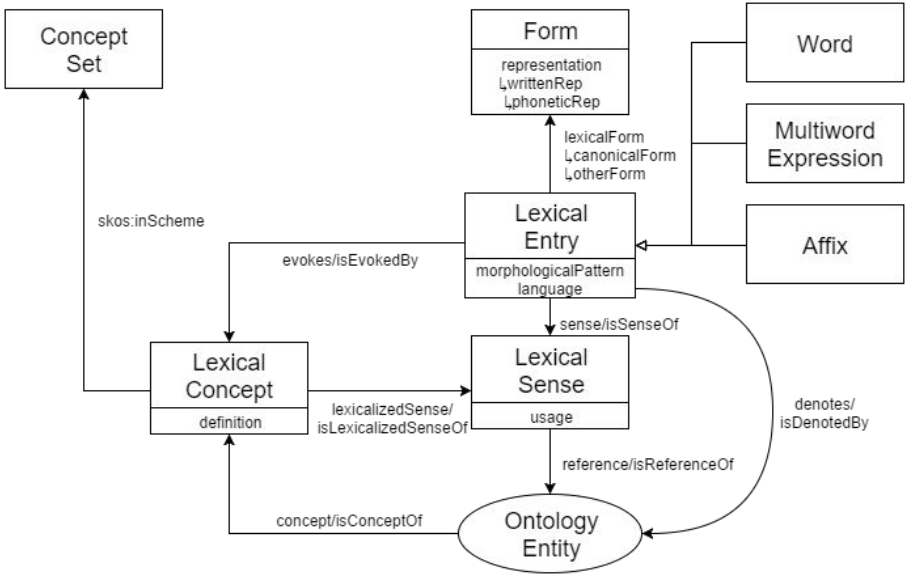
What are some interesting use cases for linking lexical datasets to other kinds of (linked data) datasets? (*Definitely* not an exhaustive list)

- The linking of word senses to **ontological/taxonomical concepts** to describe their extensions (e.g., the ontolex **reference** property) .
- Adding **encyclopedic (dbpedic?)/world knowledge** to entries, e.g., **named entities**, **geographical**, **bibliographical**, **historical** information.
- The linking of lexicons to **registries of linguistic categories**, or other kinds of linguistic ontologies
- The linking of lexical entries to **texts** through **canonical identifiers** for sections of text (**CTS-URNs**) in a corpus. Linking to corpora and tree banks.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Lexicon-Dataset linking (Semantics)

Semantics by reference was one of the core ideas behind the original **lemon** model (the predecessor of the current ontolex-lemon model). The definition of the external linking part also separates lemon from the otherwise very similar **Lexical Markup Framework** (LMF) model.



The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Lexicon-Dataset linking (Encyclopedic)

This category covers a broad variety of cases. As in other kinds of resources we can link **named entities** such as **people**, **locations**, **events** to **semantic web knowledge bases**, e.g., authority files, etc. Again it is useful to use **standard identifiers** for these **named entities** as it makes **individual datasets much more interoperable with others**. Using RDF properties to link to these means we are able to specify the relationship that these NEs have with what is being described.

For instance we can augment information about the use of variants which are limited to specific geographical regions by linking them geographical datasets (**wikidata, geonames**, **getty**, **pleiades/pelagios** for places in the ancient world) using salient properties.

The take home message is once more that *linked data makes it simple to link together heterogeneous datasets for purposes of mutual enrichment.*

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Lexicon-Dataset linking (Bibliographic)

One type of extra-linguistic knowledge that is especially important here is **bibliographic knowledge**. **What sources does a lexicon cite and which specific editions of works**? **For which purposes are other works cited** (e.g., *in order to provide attestations? As providing other kinds of evidence about word meaning or as making alternative hypotheses?*). There already exists a vocabulary for describing citation links between documents (CITO) although this is mostly focused on scientific documents. Citations are very often used to describe lexical attestations (**shameless plug**: I will be discussing attestations in more detail at my talk on Saturday.)

FRBR-influenced/aligned vocabularies like FABIO also allow us to refer to the different levels at which texts/works can be referred to (e.g., **work, expression, manifestation, item)**.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Some Useful LoD datasets

- *Generic resources* (from Wikipedia): **DBPEDIA** / **Wikidata**; people, places, works of art, concepts, ...
- *Geo-resources*: **GEONAMES** (current places and spatially determined objects)
- *Geo-historical resources*: **PELAGIOS** ecosystem (ancient places)
- *Heritage*: **Getty thesaurus** - works of art
- *Bibliographic resources* - **VIAF** (person's names), **data.bnf.fr** (BNF catalogue in LOD), **biblissima**, **sdbm** (manuscripts), …

(Thanks to Francesca Frontini for these links)

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Linking Lexical Resources

The Opportunities and Challenges Offered by the Semantic Web

Part II

Andrea Bellandi - Institute for Computational Linguistics "A. Zampolli"

elexis
european lexicographic infrastructure

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Outline

Overview and main features of LexO

Example #1: lexicon-lexicon linking in LexO

Example #2: attestation in LexO

Conclusion

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

LexO

In many cases, scholars are forced to adopt ontology editors such as Protégé to formalize their lexical or terminological resources. LexO takes into account the following aspects:

- ***Ease of use***: the editor is intended to be used mainly by humanists and, thus, hides all the technical complexities related to markup languages, language formalities and other technology issues.
- ***Collaborativeness***: LexO, being a web application, makes collaborative editing possible. The collaborative construction process of lexical resources offers very promising research opportunities in the context of e-lexicography.
- ***Sharing and linking***: the editor adheres to international standards for representing lexica and ontologies in the Semantic Web (such as ontolex-lemon and OWL), so that lexical resources can be shared easily or specific entities can be linked to existing datasets.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

LexO is already used in several contexts

Lexicon of Saussurian Terminology (For a digital edition of Ferdinand de Saussure's manuscripts)

- Type: multilingual terminology
- Languages: French, German, Italian, Spanish, Russian, Polish.
- Domain: linguistics

Totus Mundus (A Virtual Journey Around the World Atlas by Matteo Ricci, SJ (1602))

- Type: bilingual terminology
- Languages: Classical Chinese, Italian
- Domain: geography, toponomastics, cosmology, astronomy

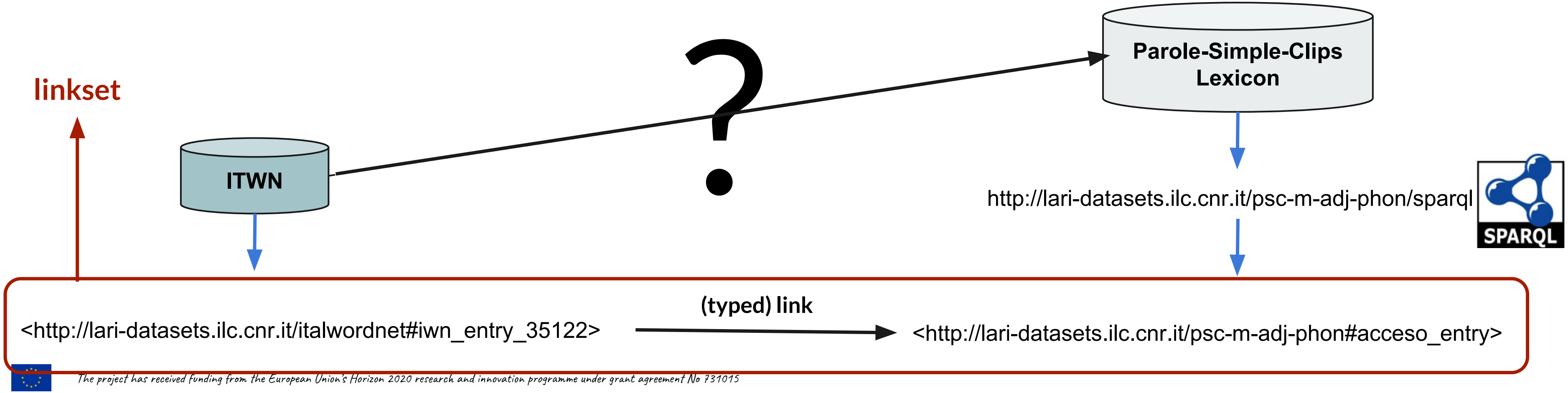
Dictionary of Old Occitan medico-botanical terminology (DiTMAO - funded by the DFG, Deutsche Forschungsgemeinschaft)

- Type: multilingual terminology
- Languages: Ancient Occitan, Hebrew, Arabic, Latin, etc.
- Domain: medieval medico-botanical

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Example #1: linguistic datasets linking

Suppose we are constructing a lexical resource and have to describe an entry that already exists in another dataset. If we trust that dataset, we can refer to it in order to augment the entry description.



linkset

ITWN

?

Parole-Simple-Clips Lexicon

http://lari-datasets.ilc.cnr.it/psc-m-adj-phon/sparql

SPARQL

(typed) link

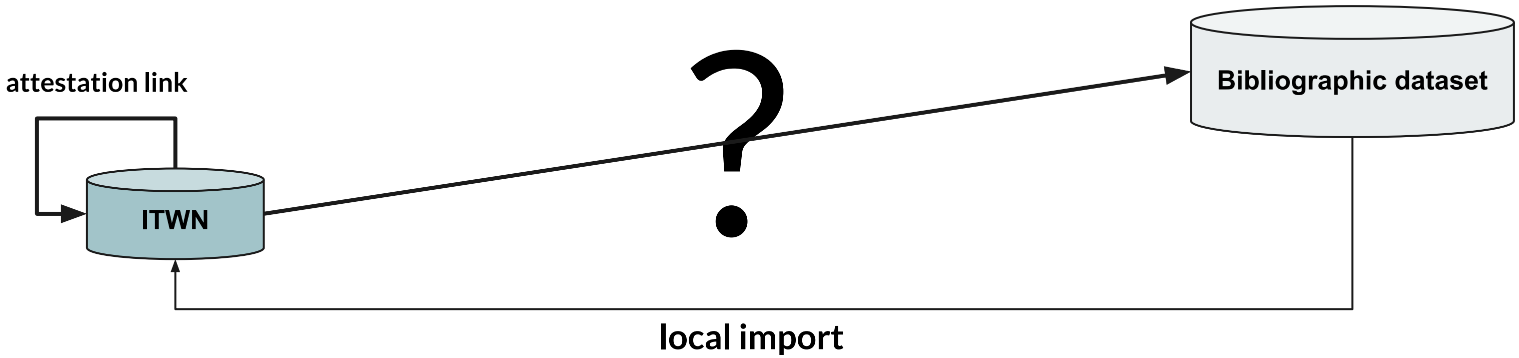
<http://lari-datasets.ilc.cnr.it/italwordnet#iwn_entry_35122>

<http://lari-datasets.ilc.cnr.it/psc-m-adj-phon#acceso_entry>

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Example #2: attestation

Let’s suppose we want to attest a lexical form in a bibliographic description of a text contained in a (remote) dataset.



The project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015

Conclusion

We have presented the advantages of using semantic web technologies, in representing lexical and terminological resources

more theoretical part: semantic web basic concepts, entity linking, opportunities and challenges
more practical part: the importance of making these technologies accessible and usable by lexicographers, scholars and humanists communities

If you want to find out more we are giving the following presentations/demos/posters @ Euralex 2018:

- **On Attestations**:
  - Khan, F., Boschetti, F. ***The Representation of Citations in Lexical Resources in Linked Data***.
- **On LexO**
  - Bellandi, A., Giovannetti, E. Piccini, S. ***Collaborative Editing of Lexical and Termino-ontological Resources: a quick introduction to LexO.***
  - Piccini, S, Bellandi, A., Giovannetti, E.. ***A Semantic Web Approach to Modelling and Building a Bilingual Chinese-Italian Termino-ontological Resource***

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015

Acknowledgements

Thanks to **Francesca Frontini** for her help with the slides. Thanks to our colleagues at ILC-CNR including **Monica Monachini** and **Emiliano Giovannetti**. Thanks to John and David for organising the event and inviting us...and thanks for listening!