Modelling dictionaries as complex objects using top level ontologies

Fahad Khan

ILC-CNR & CLARIN-IT, Pisa, Italy

fahad.khan@ilc.cnr.it

Nexus Linguarum

H²IOSC Humanities and cultural Heritage Italian Open Science Cloud

MorDigital

Introduction

Presentation of a proposal for an integrated approach to producing digital versions of dictionary texts as linked data. This work places itself at the intersection between e-Lexicography, linked data, ontology engineering and the digital humanities.

In the interests of making this talk as self-contained as possible we will try and assume very little background knowledge and introduce relevant concepts as we go along.

We start by looking at what we mean here by ontologies and explaining what linked data is and why it's important.

Some Background

Ontologies - A Very Quick Overview

The standard definition of the term ontology is as follows:

- "**An ontology is a formal, explicit specification of a shared conceptualization**" (Studer et. al.)

Ontologies allow us to specify the **classes** and **individuals** of interest in a domain, the properties that pertain to them and the **relationships** that hold between them.

They allow us to make the shared assumptions and conceptions that a given community holds towards a **certain domain of knowledge** -- and to make these **processable by a computer**

In many cases we are working with formal languages with nice computational properties so we can reason over ontologies and derive new knowledge

What are ontologies used for?

Ontologies have many different uses. One of the more central purposes for which ontologies are used is to give a formal description of the meanings of the terms in a **controlled vocabulary**.

These controlled vocabularies can subsequently be used to align together datasets that use different (but compatible) categorisations/labels on the basis of the descriptions of what these categorisation *mean*.

It is in the biomedical domain that they have had the greatest success up till now and made the greatest impact to an individual field. They have become part of the **everyday practice** of researchers in the **biomedical sciences**.

What are ontologies used for?

More concretely we have the following specific use cases in the biomedical domain:

- Annotation with standard identifiers, in order to integrate together and **query multiple datasets** (e.g., Gene Ontology)
- As vocabularies for applications relying on **domain-specific terms** (e.g., text mining using ontology labels)
- **Reasoning over ontology annotated datasets** (e.g., determining which protein family a protein belongs to)
- Data Mining and Analysis, **using ontologies as background knowledge** (e.g., Gene Set Enrichment Analysis)

What are ontologies made out of?

Abstracting away from specific formal languages, ontologies are used to describe three kinds of entity:

- **Classes/concepts**, Person, Country, Sheep, Author, etc
- **Properties (or relations)** X child of Y, X lives in Y, X is located in Y, X loves Y
- **Individuals** Fahad Khan, Lisbon, Pisa,

This threefold distinction is echoed in the division of ontologies into a **TBox** (**T**erminological Box) and an **ABox** (**A**ssertion Box) and sometimes a separate **RBox** (**R**ole/relation Box) too.

These 3 kinds of entity are the primitive components out of which formal ontology languages are constructed.

Top Level Ontologies

Important that ontologies for specific domains are **interoperable** with each other.

Top-level ontologies provide a layer of higher level, more abstract concepts that can be (re-)used in domain ontologies. They help ensure consistency, coherence, and accuracy. They also help to ensure we don't need to reinvent our most basic concepts each time (e.g., What we mean by an event? How do we define attributes of things?)

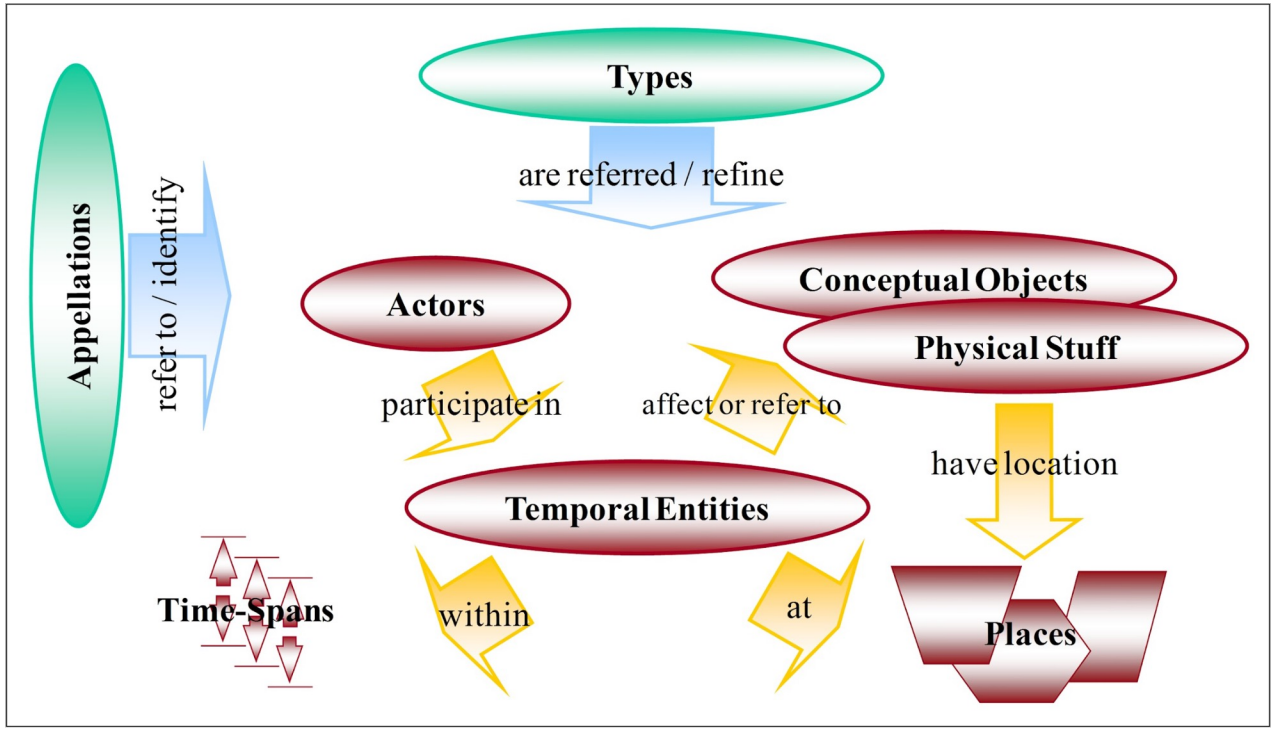
CIDOC CRM

CIDOC CRM is an upper level ontology (CIDOC refers to **International Committee for Documentation of the International Council of Museums** and **CRM**= Conceptual Reference Model) that was originally created for the cultural heritage domain. Published as an **ISO Standard** in 2006

Intended to mediate between cultural heritage datasets in order to facilitate information exchange and data integration

The best known and most widely used top level ontology in the humanities, CIDOC CRM has proven itself an important tool in establishing interoperability between individual resources through making descriptions of objects **semantically transparent** via ontological concepts and properties.

CIDOC CRM


Types
are referred / refine
Appellations
refer to / identify
Actors
Conceptual Objects
Physical Stuff
participate in
affect or refer to
have location
Temporal Entities
Time-Spans
within
at
Places

CIDOC CRM

Object type: tanto; short sword-sheath; menuki; kozuka; hilt; fuchi-kashira; blade
Museum number: 1992,0523.2
Description: Sword blade (tanto); with mounting (short sword-sheath; kozuka; hilt; menuki; fuchi-kashira). Blade: made of steel; signed. Sheath: made of black lacquered wood. Hilt: with gold mekugi; made of wood and skin (ray). Kozuka: crane in high-relief coloured metal inlay on silver ground; inscribed. Menuki: in shape of corn?; made of gilded metal. Fuchi-kashira: made of black lacquered metal. Soshu school blade and Goto school metal fittings.
Producer name: Made by: Goto Ichijo (metal fittings); Made by: Shintogo Kunimitsu (blade)
Culture/period: Meiji Era (metal fittings); Kamakura Period (blade)
Date: 14thC (early; blade); 19thC (late; metal fittings)
Production place: Made in: Japan (Asia,Japan)
Materials: wood; steel; silver; ray skin; metal; lacquer; gold
Technique: lacquered; inlaid ; high relief; gilded; colour
Inscriptions:
Inscription Type: signature
Inscription Script: Japanese
Inscription Position: blade, tang, obverse
Inscription Content: 国光; Inscription Transliteration; Kunimitsu, etc
Curator's comments: Harris 2005 - 'Hira zukuri' tanto blade with the slight 'uchizori' curve of the late Kamakura period. The blade has 'itame' with 'mokume' grain with 'jifu utsuri' and much 'chikei'. The 'suguha hamon' is of fine 'nie' with 'kinsuji'. The maker, Shintogo Kunimitsu, is feted as the founder of the Soshu tradition at Kamakura in the late Kamakura period.
Bibliography: Harris 2005 fig. 11, col. pl. 11, 12 bibliographic details
Location: G93/case10
Exhibition history
Exhibited: 2006 Oct 13-, BM Japanese Galleries, 'Japan from prehistory to the present'
Subjects: arms/armour term details;
Acquisition name: Purchased through: Eskenazi Ltd biography; Purchased from: Christie's biography; Previous owner/ex-collection: Dr Walter A Compton biography
Acquisition date: 1992
Acquisition notes: Bought at Christie's (lot 226) by Eskanazi Ltd at the BM's request. Former collection of Walter A Compton.
Department: Asia
Registration number: 1992,0523.2

CIDOC CRM

Object type: tanto; short sword-sheath; menuki; kozuka; hilt; fuchi-kashira; blade
Museum number: 1992,0523.2
Description: Sword blade (tanto); with mounting (short sword-sheath; kozuka; hilt; menuki; fuchi-kashira). Blade: made of steel; signed. Sheath: made of black lacquered wood. Hilt: with gold mekugi; made of wood and skin (ray). Kozuka: crane in high-relief coloured metal inlay on silver ground; inscribed. Menuki: in shape of corn?; made of gilded metal. Fuchi-kashira: made of black lacquered metal. Soshu school blade and Goto school metal fittings.

Producer name: Made by: Goto Ichijo (metal fittings); Made by: Shintogo Kunimitsu (blade)
Culture/period: Meiji Era (metal fittings); Kamakura Period (blade)
Date: 14thC (early; blade); 19thC (late; metal fittings)
Production place: Made in: Japan (Asia,Japan)
Materials: wood; steel; silver; ray skin; metal; lacquer; gold
Technique: lacquered; inlaid ; high relief; gilded; colour

Inscriptions:
Inscription Type: signature
Inscription Script: Japanese
Inscription Position: blade, tang, obverse
Inscription Content: 国光; Inscription Transliteration; Kunimitsu, etc

Curator's comments: Harris 2005 - 'Hira zukuri' tanto blade with the slight 'uchizori' curve of the late Kamakura period. The blade has 'itame' with 'mokume' grain with 'jifu utsuri' and much 'chikei'. The 'suguha hamon' is of fine 'nie' with 'kinsuji'. The maker, Shintogo Kunimitsu, is feted as the founder of the Soshu tradition at Kamakura in the late Kamakura period.

Bibliography: Harris 2005 fig. 11, col. pl. 11, 12 bibliographic details
Location: G93/case10

Exhibition history
Exhibited: 2006 Oct 13-, BM Japanese Galleries, 'Japan from prehistory to the present'
Subjects: arms/armour term details;
Acquisition name: Purchased through: Eskenazi Ltd biography; Purchased from: Christie's biography; Previous owner/ex-collection: Dr Walter A Compton biography

Acquisition date: 1992
Acquisition notes: Bought at Christie's (lot 226) by Eskanazi Ltd at the BM's request. Former collection of Walter A Compton.
Department: Asia
Registration number: 1992,0523.2

Introduction to Linked Data & the Semantic Web

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', unlike 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.

Linked Data

In 2006, Tim Berners-Lee (the inventor of the WWW) defined the **four guiding principles** for publishing data as linked data. These are:

1. Use **URIs** as names for things
2. Use **HTTP URI**s 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 the data available under an open license then our data becomes **Linked Open Data** (LOD).

Linked Data

These four principles were simplified and distilled by TBL into the following three:

1. All kinds of **conceptual things**, they have names now that start with **HTTP**.
2. If I take one of these HTTP names and I look it up...I will get back some data in a **standard format** which is kind of useful data that somebody might like to know about that **thing**, about that **event**.
3. When I get back that information it's not just got somebody's height and weight and when they were born, it's got **relationships**. And when it has relationships, whenever it expresses a relationship then the other thing that it's related to is given one of those names that starts with HTTP.

Resource Description Framework

As we mentioned in the last slide with linked data we make everything a resource and give it an identifier that we can look up (‘dereference’) in the same way we do with web pages. But we would like to talk about them, describe them and the relationships between them. To say things like, e.g.,

Lisbon is a city, located in Portugal’ or

‘António Costa is the Prime Minister of Portugal’

Where ‘Lisbon’, ‘city’, ‘Portugal’, ‘António Costa’, ‘Prime Minister of Portugal’ as well as the properties ‘is a’, ‘located in’ are all represented as resources with URIs.

The **Resource Description Framework** (RDF) is a model, a kind of standard **way of describing resources and relating them together**.

Resource Description Framework

The idea with RDF is to describe everything using statements of the following form:

Subject- Predicate-Object

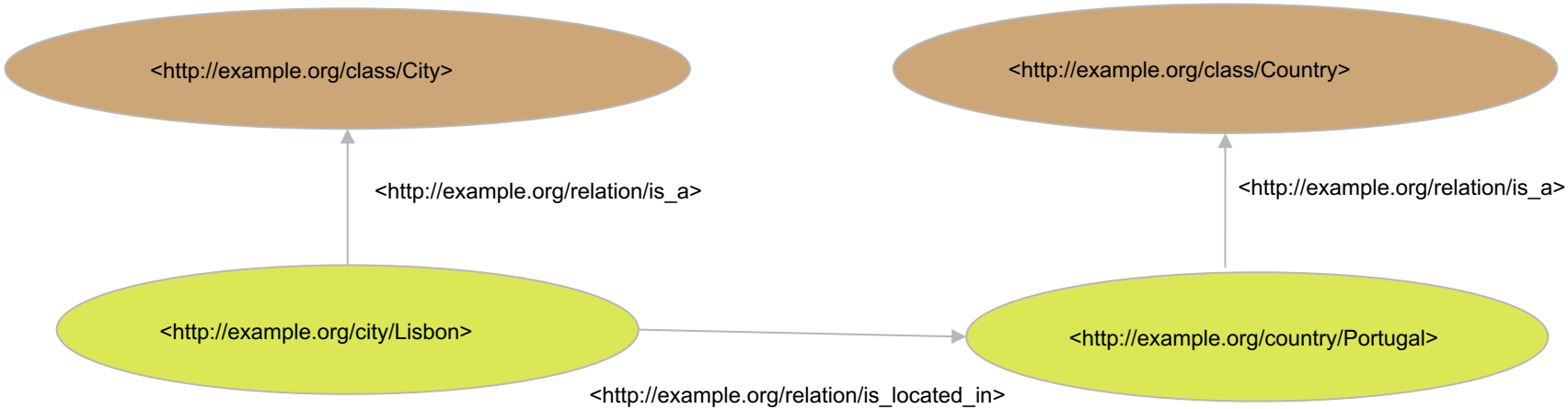
Where the Subject and Object are resources referred to by URIs which are related together by the Predicate which is a ‘property’ also referred to by a URI.

Each such statement is known as an RDF-triple. A linked data dataset consists of a series of such RDF-triples.

There exist a huge number of linked data **vocabularies** covering different topics and subject areas that give us properties and classes that we can use.

Resource Description Framework

With RDF we are essentially creating lots of graphs linking up different resources that can be individuals such as **Lisbon** or **Joe Biden** or classes such as **Country** or **President of the United States of America**.



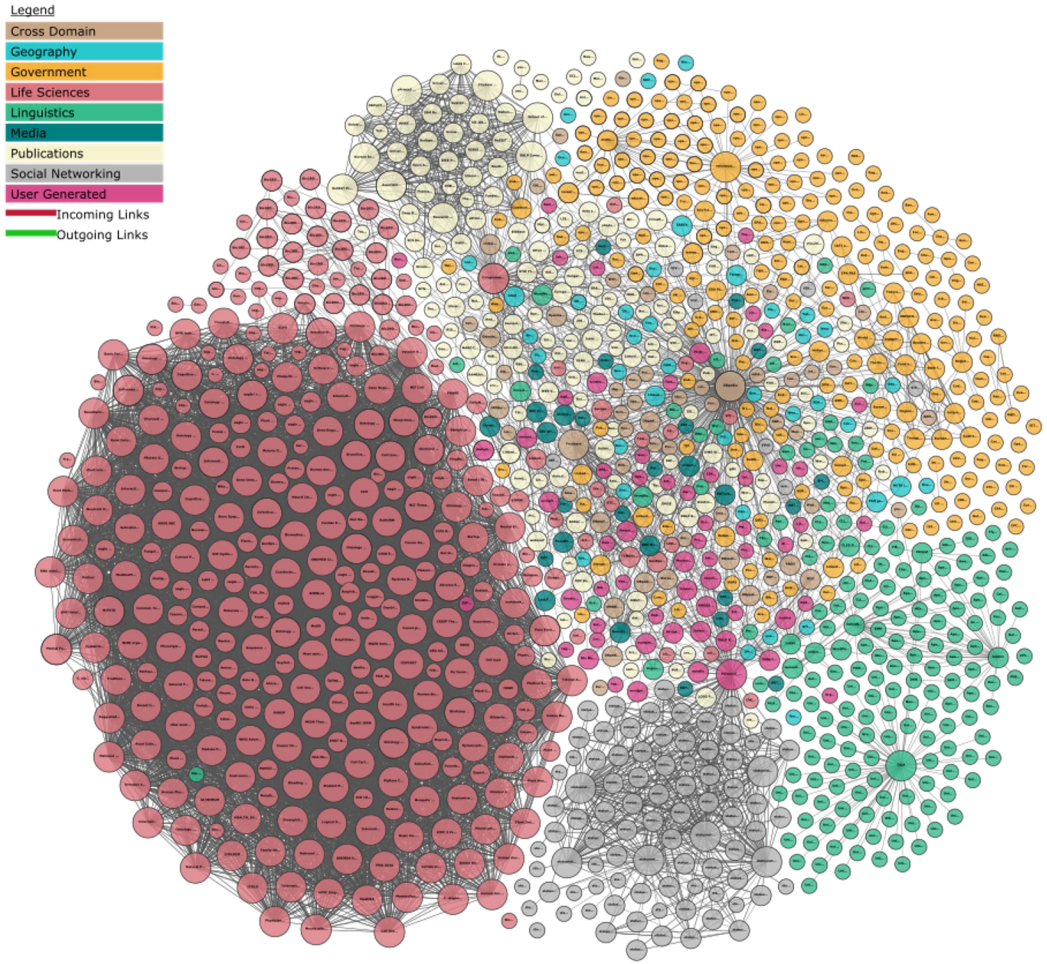
The Linked Open Data Cloud

Taken together the thousands of datasets published as linked open data form the linked open data cloud.

The latest diagram of the cloud can be found here:

http://lod-cloud.net/

The cloud groups together resources in a number of different domains such as **Geography**, **Government**, **Social Media**, **Linguistics**, as well as **Cross Domain** resources. One of the most well connected hubs is DBpedia. The Linked Open Data version of Wikipedia which is regularly harvested from the hypertext version.


Legend
Cross Domain
Geography
Government
Life Sciences
Linguistics
Media
Publications
Social Networking
User Generated
Incoming Links
Outgoing Links

Ontologies on the Semantic Web

Linked data datasets all have the same underlying structure: a set of subject-predicate-object structure statements -- our first fundamental level of interoperability.

With formal languages **RDFS** and the **Web Ontology Language (OWL)** we can create and publish ontologies as linked data on the Semantic Web defining classes and properties. These are the two most well known and popular formal ontology languages.

We also have a very powerful querying language SPARQL which can be used to write complex (graph based) queries remotely on the web.

Ontologies on the Semantic Web

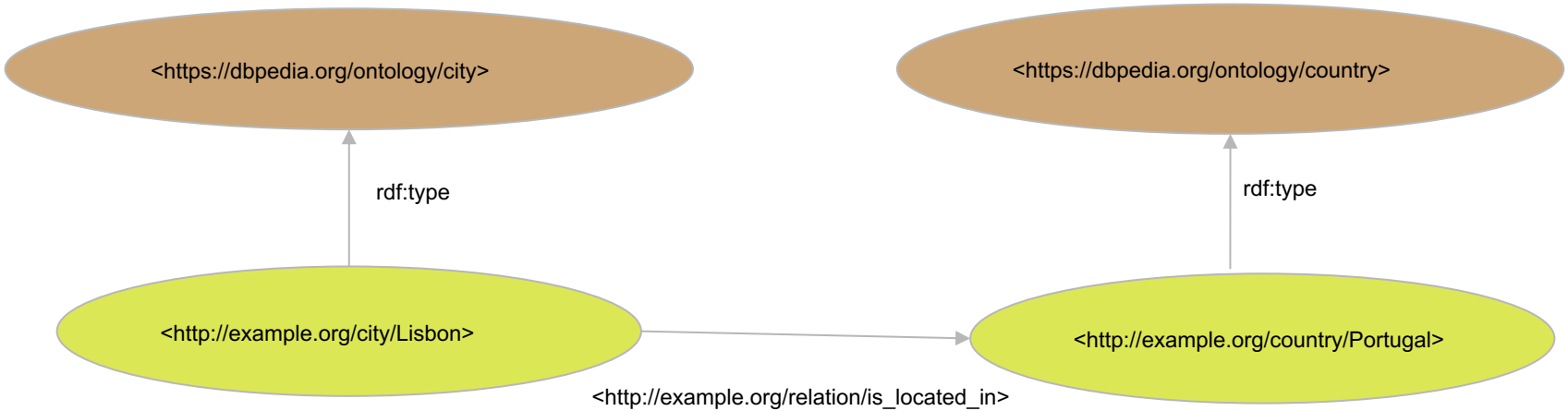
There also exist several reasoning engines and ontology editors for OWL.

When publishing linked data datasets we are encourage to re-use existing linked data ontologies/vocabularies to describe/structure our data.

These can be top level ontologies like the RDFS ontology CIDOC CRM or domain ontologies like OntoLex-Lemon which we will look at next.

Ontologies on the Semantic Web

We can re-use (are encouraged to re-use) classes and properties from pre-existing ontologies/vocabularies to create our datasets. In the next few slides we will look at one such model for creating lexicons.



lemon

Lemon stands for the **Lexicon Model for Ontologies**. It was developed as part of the Monnet project as a collaboration between several European universities and academic institutes. It is closely based on previous standards/models, and in particular on LMF.

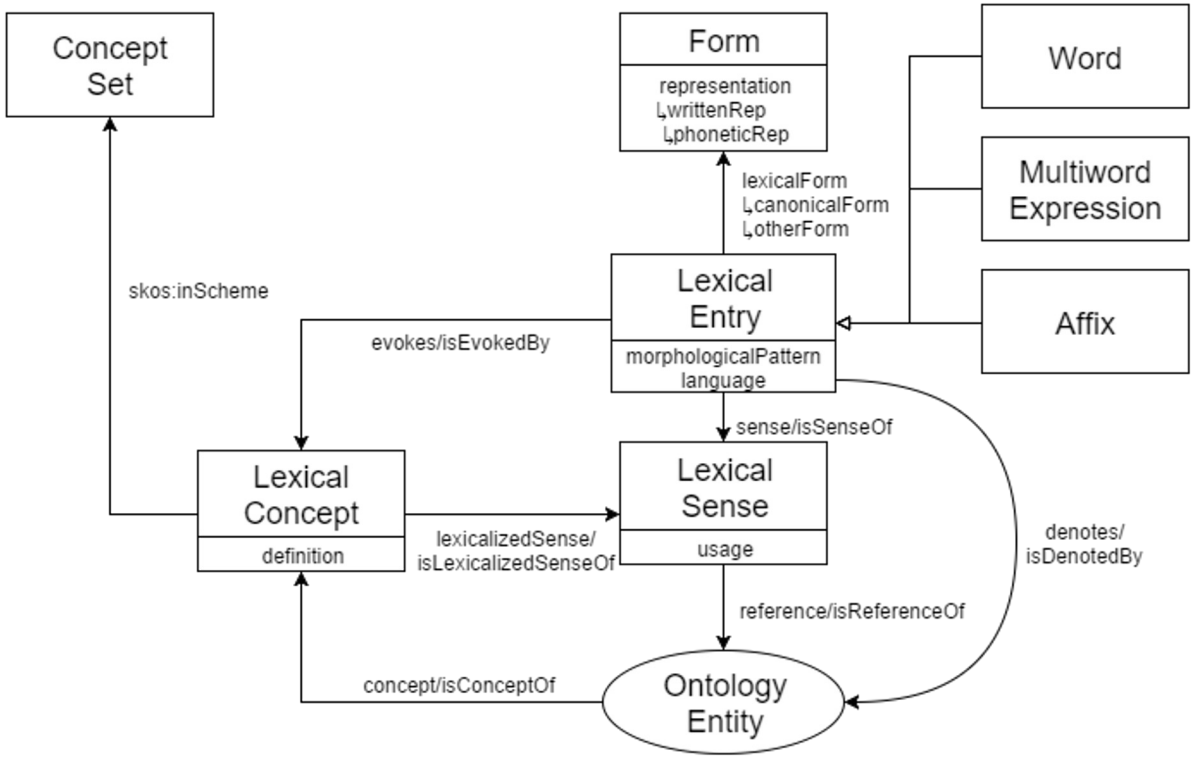
Lemon was originally intended as a model for enhancing and ontologies like DBpedia with linguistic knowledge.

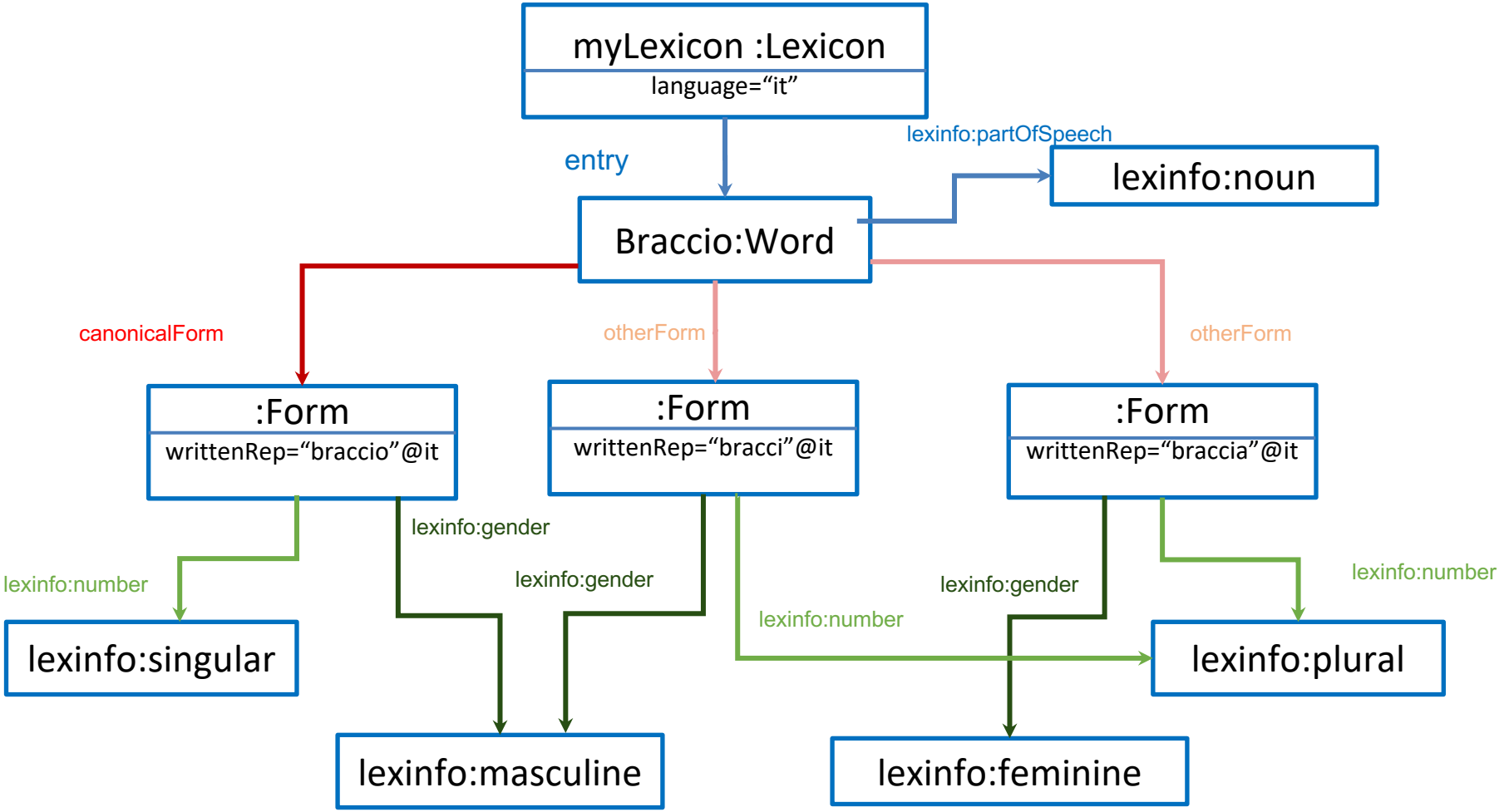
Ontolex-lemon

Lemon soon became the most popular ontology for representing lexicons in RDF, taking on the status of a de facto standard. It has been used to model the Princeton (and other) Wordnets, DBnary (the linked data version of Wiktionary), FrameNet and VerbNet.

This success led to the development of a new version, Ontolex-lemon, developed by the W3C Ontology-Lexica community group in which anyone is able to participate. Ontolex-lemon was published in December 2016. It consists of a **core module** as well as a **metadata module (lime)**, a **syntax and semantics module (synsem)**, a **decomposition module (decomp)**, and a **variation and translation module (vartrans)**.

Ontolex-lemon Core




myLexicon :Lexicon
language="it"
entry
lexinfo:partOfSpeech
lexinfo:noun
Braccio:Word
canonicalForm
otherForm
otherForm
:Form
writtenRep="braccio"@it
:Form
writtenRep="bracci"@it
:Form
writtenRep="braccia"@it
lexinfo:gender
lexinfo:number
lexinfo:gender
lexinfo:number
lexinfo:gender
lexinfo:number
lexinfo:singular
lexinfo:masculine
lexinfo:feminine
lexinfo:plural

Summary 1/2

- Ontologies are descriptions of the meanings of the key terms in a domain in a formal language that makes them easier to process with computers
- Concepts are represented as classes of things, properties as relationships between individuals of these classes
- Complex concepts are defined using simpler ones using classes and properties
- Top level ontologies give the most fundamental classes, entities and properties (vocabulary elements) that can be re-used in other ontologies
- Q: How can we re-use such elements and those of domain ontologies in practice?
- A: Linked data and the Semantic Web which gives us a way of modelling and publishing data in a way that makes it more **F**indable **A**ccessible **I**nteroperable and **R**eusable

Summary 2/2

- In linked data datasets data is modelled as a series of statements (triples) using resolvable universal identifiers...giving us a mechanism for re-using vocabulary elements
- We also have a Semantic Web based ontology language OWL and numerous ontologies which we can re-use in constructing our own datasets (as well as other datasets which we can link to)
- We can make our datasets available for remote querying using a powerful querying language SPARQL
- Next we look at how we can use ontologies to publish (editions of) texts as linked data

Applying Ontologies to Texts

Using Ontologies to Model Texts

Linked data ontologies already used in modeling **cultural heritage data**:

- E.g., **CIDOC-CRM** has been successfully used in several projects including aligning museum catalogues and archaeological datasets

There already exist linked data ontologies/vocabularies for textual metadata which allow for the description of **bibliographic information for textual works**:

- The project "**Mapping the Manuscript Migrations**" is a good example of the impact that linked data + ontologies can have

However, ontologies like CIDOC-CRM offer the possibility of modeling texts **as complex objects** and integrating seemingly contradictory properties.

Using Ontologies to Model Texts

Modelling texts is **challenging** due to their dual nature as physical and as information.

Texts are associated with a physical support, these physical supports can be located in different geographical locations, as well as being subject to various physical processes, such objects can have a fascinating history in their own right (see the MMM project).

On the other hand they also have an (informational) content that can, e.g., be translated into different languages or adapted in different media.

Ontologies provide **a principled way of describing and reasoning about such entities**.

In the world of ontology engineering we call such kinds of multifaceted entities, **informational entities**. These are complex ontological objects that have **a physical form and carry informational content**.

Using Ontologies to Model Texts

Informational entities are related to ***dot objects*** first proposed by the linguist James Pusetejovksy in order to model phenomena such as **co-predication:**

> *"The blue dictionary has more understandable but less comprehensive definitions than the red one, that's why it's lighter!"*
>
> *“The dictionary is outdated and very often incorrect in its etymological analyses but the definitions can be amusing and it makes a nice doorstop.”*

As well as books, other examples of dot objects include **countries**, **institutions**, **diseases**.

Some ontologists argue for the introduction of **separate complex categories** in ontologies to account for dot objects. These categories could be defined using a modified version of the coincidence relation, used to model situations like those described by the clay and statue paradox.

Using Ontologies to Model Texts

Some aspects of texts are difficult to model using already existing ontologies (and formal ontology languages):

> *What are the **arguments** of the text? What is the **plot** of **a literary work**? What are the main **themes** of a novel? What **literary devices** does it make use of?*

Lack of agreement on shared vocabularies and ontologies for describing such properties is a hurdle to modeling texts using linked data ontologies in general.

However certain types of texts can be modeled using already existing ontologies, and **dictionaries/lexicographic resources** are one such example.

Why Lexicographic Resources?

The creation of digital descriptions/versions of ***any*** kind of text confronts us with the distinction between the **content of a text**, and how the **content** is **presented**. Dictionaries are an interesting case: they tend organise **similar kinds** of (**linguistic) information** in **standardised ways**.

Moreover this (linguistic) content can be represented (in a formal way) much more easily than in other cases, e.g., plays, novels, encyclopedias, etc. This makes them **a useful test case in the modelling of texts using ontologies**.

To a large extent we can combine existing vocabularies to model dictionaries **as complex ontological objects**

Encoding Dictionaries as Structured Datasets

What kinds of things can we potentially encode in a linked data edition of a dictionary using ontologies?

- **Metadata** common to other texts can be encoded using existing vocabularies such as **Dublin Core** and **DCAT**.
- Descriptions specific to legacy printed texts, such as **number of pages** and **fonts used**
- Dictionary entries provide information on morpho-syntactic properties of words, **citations**, **examples**, and **etymologies** which can be represented as knowledge graphs

The **extraction** of this information can be done using machine learning methods; ontologies can be used to create **schemas** 'templates' for the information. But the semantics of this information isn't always straightfoward (challenge of what to encode/leave out). In the next few slides we look at some of the complexities that information organisation in dictionaries can present.

Citation: An Anomalous Example

Citations can be used to *attest* various different properties of a lexical entry, e.g., **orthographic**, **semantic**, **phonetic**. But they can also be used for other purposes.

We will look at the entry for **ἀνώμᾰλος** (anomalos) from the hugely influential Liddell-Scott-Jones ancient Greek-English lexicon (made available online by the **Perseus project**).

An Anomalous Example

ἀνώμα^λ-ος , ον, (ἀ- priv., ὁμαλός)

A.*uneven, irregular,* “χώρα” **Pl.Lg.625d**; “φύσις” **Id.Ti.58a**; “τὸ ἀ. τῆς ναυμαχίας” **Th.7.71** (cj.), cf. **Arist.Pr.885a15**: and in **Sup., Hp.Aër.13**; of movements, **Arist.Ph.228b16**, al.; of periods of time, **Id.GA772b7**; of the voice, **ib.788a1**. Adv. “-λως, κινεῖσθαι”**Id.Ph.238a22**, cf. **Pl.Ti.52e**.

II. of conditions, fortune, and the like , “φεῦ τῶν βροτείων ὡς ἀ. τύχαι” **E.Fr.684**; πόλις, πολιτεία, **Pl.Lg.773b**, **Mx.238e**; “θέα” **Plot.6.7.34**. Adv. “-λως” **Hp.Prog.3**, **Isoc.7.29**; ἀ. διατεθῆναι τὸ σῶμα fall into *precarious* health, **Prisc.p.333** ***D***.

III. of persons, *inconsistent, capricious*, “ὁμαλῶς ἀ.” **Arist.Po.1454a26**; ὄχλος, δαιμόνιον, **App.BC3.42**, **Pun.59**; “πίθηκος” **Phryn. Com.20**; “τύχη” **AP10.96.** Adv. “-λως” **Isoc. 9.44**.

An Anomalous Example

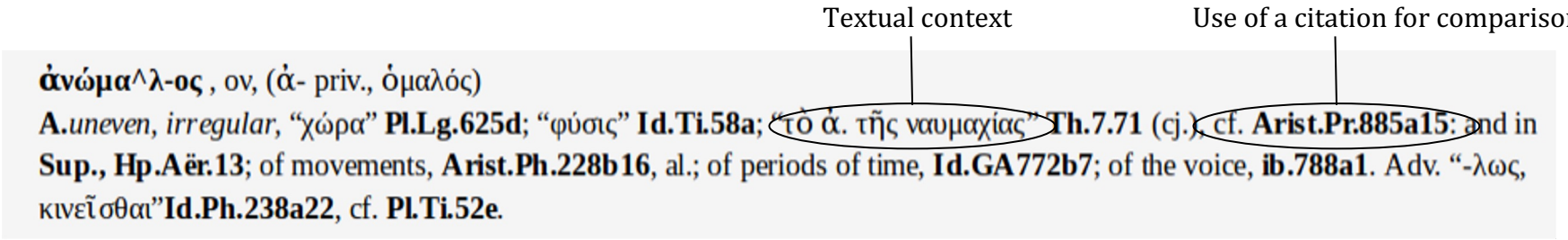
ἀνώμα^λ-ος , ον, (ἀ- priv., ὁμαλός)

A.*uneven, irregular,* “χώρα” **Pl.Lg.625d**; “φύσις” **Id.Ti.58a**; “τὸ ἀ. τῆς ναυμαχίας” **Th.7.71** (cj.), cf. **Arist.Pr.885a15**: and in **Sup., Hp.Aër.13**; of movements, **Arist.Ph.228b16**, al.; of periods of time, **Id.GA772b7**; of the voice, **ib.788a1**. Adv. “-λως, κινεῖσθαι”**Id.Ph.238a22**, cf. **Pl.Ti.52e**.

II. of conditions, fortune, and the like , “φεῦ τῶν βροτείων ὡς ἀ. τύχαι” **E.Fr.684**; πόλις, πολιτεία, **Pl.Lg.773b**, **Mx.238e**; “θέα” **Plot.6.7.34**. Adv. “-λως” **Hp.Prog.3**, **Isoc.7.29**; ἀ. διατεθῆναι τὸ σῶμα fall into *precarious* health, **Prisc.p.333** ***D***.

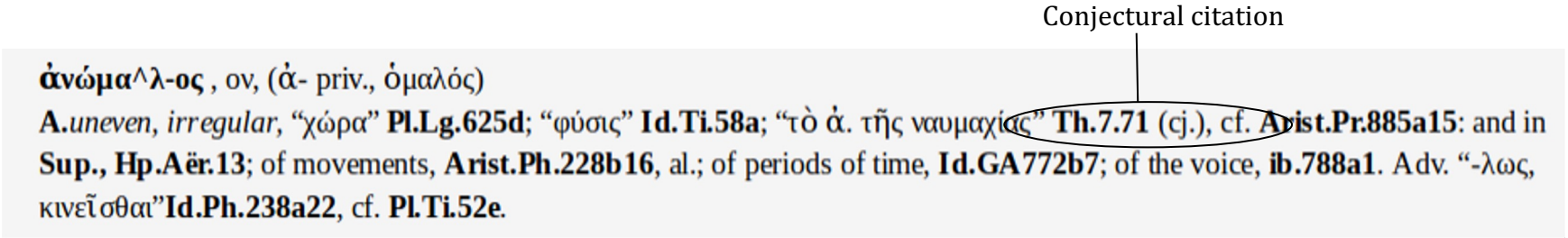
III. of persons, *inconsistent, capricious*, “ὁμαλῶς ἀ.” **Arist.Po.1454a26**; ὄχλος, δαιμόνιον, **App.BC3.42**, **Pun.59**; “πίθηκος” **Phryn. Com.20**; “τύχη” **AP10.96.** Adv. “-λως” **Isoc. 9.44**.

An Anomalous Example



Most of the citations in the example are used to *attest* to different shades of meaning of the word in question, with the **textual context** of an attestation **explicitly given** in one case. In other cases citations are used to contrast with other citations: without necessarily attesting to the word sense being dealt with. This use of the citation is annotated by the abbreviation ‘**cf.**’.

An Anomalous Example



It is also interesting to note that one of the citations, **'Th.7.71'**, is marked with a **'(cj.)'** meaning that it is conjectural -- i.e., it is based on **a reconstruction of the original text**. In this case we can say that the entry cites the text (from the corpus of works attributed to Thucydides) even though the original text might not have actually attested the sense itself.

An example etymological entry

GIRL, a female child, young woman. (E.) ME. gerle, girle, gyrle, formerly used of either sex, and signifying either a boy or girl. In Chaucer, C.T. 3767 (A 3769) gerl is a young woman; but in C.T. 666 (A 664), the pl. girles means young people of both sexes. In Will. of Palerne, 816, and King Alisander, 2802, it means ‘young women;’ in P. Plowman, B.i.33, it means ‘boys;’ cf. B. x. 175. Answering to an AS. form *gyr-el-, Teut. *gur-wil-, a dimin. form from Teut. base *gur-. Cf. NFries. gor, a girl; Pomeran. goer, a child; O. Low G. gor, a child; see Bremen Wortebuch, ii. 528. Cf. Swiss gurre, gurrli,a depriciatory term for a girl; Sanders, G. Dict. i. 609, 641; also Norw. gorre, a small child (Aasen); Swed. dial. garra, guerre (the same). Root uncertain. Der. girl-ish, girlish-ly, girl-ish-ness, girl-hood.

An example etymological entry

GIRL, a female child, young woman. (E.) ME. gerle, girle, gyrle, formerly used of either sex, and signifying either a boy or girl. In Chaucer, C.T. 3767 (A 3769) gerl is a young woman; but in C.T. 666 (A 664), the pl. girles means young people of both sexes. In Will. of Palerne, 816, and King Alisander, 2802, it means ‘young women;’ in P. Plowman, B.i.33, it means ‘boys;’ cf. B. x. 175. Answering to an AS. form *gyr-el-, Teut. *gur-wil-, a dimin. form from Teut. base *gur-. Cf. NFries. gor, a girl; Pomeran. goer, a child; O. Low G. gor, a child; see P[illegible]28. Cf. Swiss gurre, gurrli,a depriciatory term for a [illegible] also Norw. gorre, a small child (Aasen); Swed. [illegible]e). Root uncertain. Der. girl-ish, girlish-ly, girl-ish-ness, girl-hood.

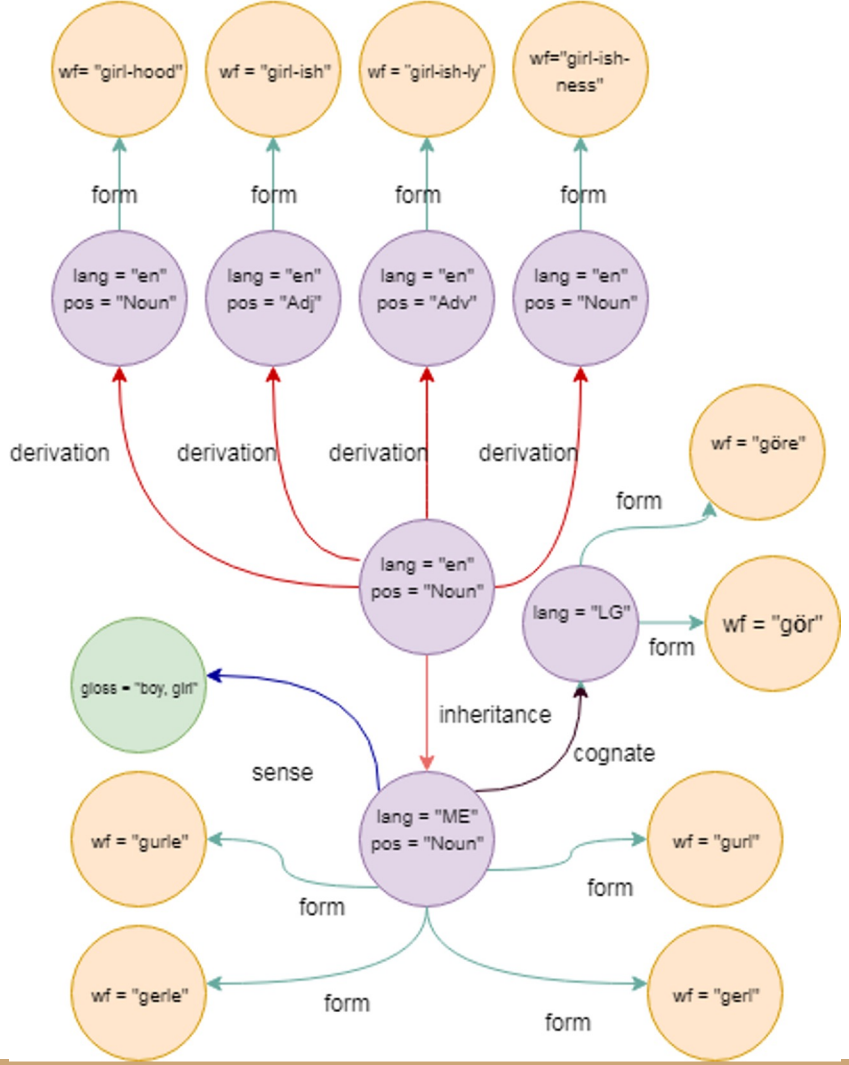
Description of the history and development of the word

An example etymological entry

girl

Three different hypotheses for the origin of the same word

, whence *gir*[illegible] *gerle*, *gurle*: o.o.o.: perh of C origin: cf Ga and Ir *caile*, EIr *cale*, a girl; with Anglo-Ir *girleen* (dim *-een*), a (young) girl, cf Ga-Ir *cailin* (dim -in), a girl. But far more prob, *girl* is of Gmc origin: Whitehall postulates the OE etymon **gyrela* or **gyrele* and adduces Southern E dial *girls*, primrose blossoms, and *grlopp*, a lout, and tentatively LG *goere*, a young person (either sex). Ult, perh, related to L *puer*, *puella*, with basic idea '(young) growing thing'.


wf= "girl-hood"
wf = "girl-ish"
wf = "girl-ish-ly"
wf="girl-ish-ness"
form
form
form
form
lang = "en"
pos = "Noun"
lang = "en"
pos = "Adj"
lang = "en"
pos = "Adv"
lang = "en"
pos = "Noun"
derivation
derivation
derivation
derivation
wf = "göre"
form
lang = "en"
pos = "Noun"
lang = "LG"
wf = "gör"
form
gloss = "boy, girl"
inheritance
cognate
sense
lang = "ME"
pos = "Noun"
wf = "gurle"
wf = "gurl"
form
form
wf = "gerle"
wf = "gerl"
form
form

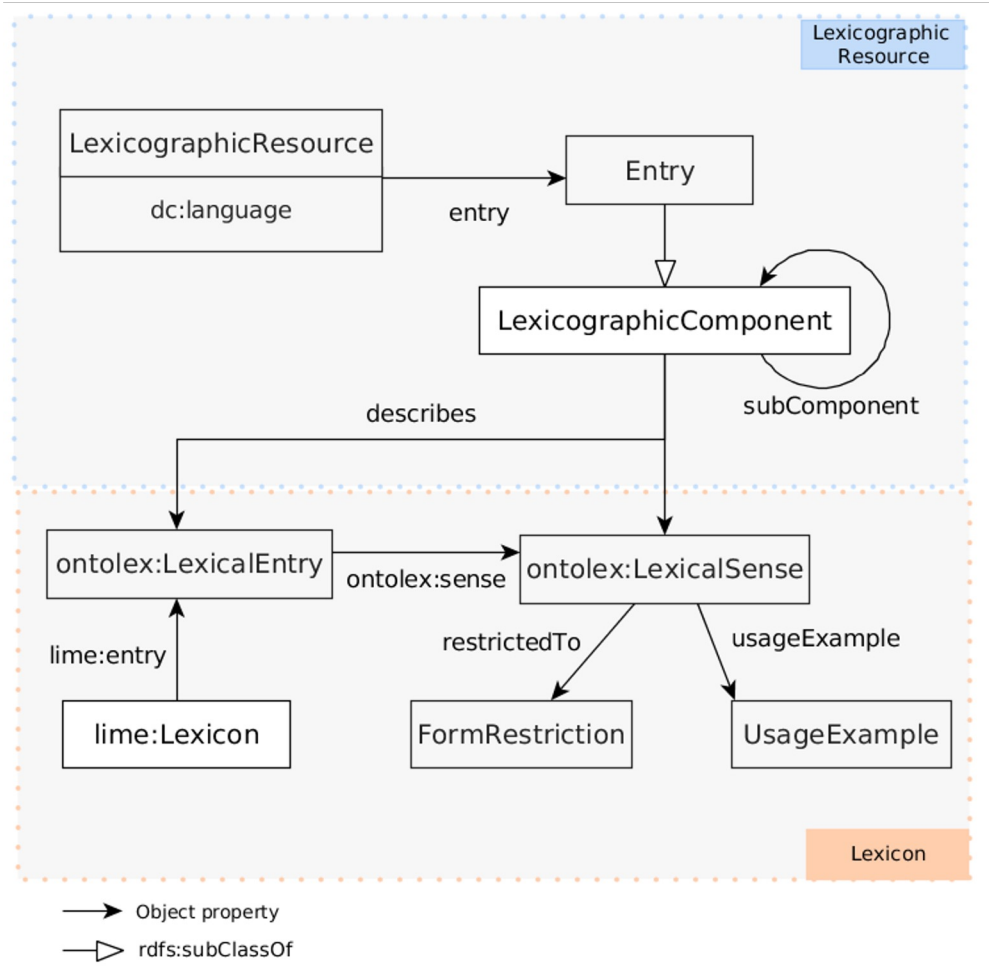
Encoding Dictionaries using Semantic Web Ontologies

Lexicog

The **OntoLex-Lemon Lexicography Module (lexicog)** developed by the W3C OntoLex group to represent some of the structural information “lost” in OntoLex-Lemon.

It defines new classes such as **Lexicographic Resource** (complementing OntoLex **Lexicon**) which consists of single **Entry** individuals which represent lexicographic articles and which can be realised by OntoLex **Lexical Entry** elements.

Entry is a subclass of **Lexicographic Component** which represents elements which describe the structuring of lexicographic articles.


Lexicographic Resource
LexicographicResource
dc:language
entry
Entry
LexicographicComponent
subComponent
describes
ontolex:LexicalEntry
ontolex:sense
ontolex:LexicalSense
lime:entry
restrictedTo
usageExample
lime:Lexicon
FormRestriction
UsageExample
Lexicon
Object property
rdfs:subClassOf

Dictionaries as Textual/Material Objects

OntoLex-Lemon + Lexicog however **still aren't** sufficient to represent all the different aspects we might be potentially interested in.

- Who **compiled** the dictionary, is it based on **previous works**?
- What about the **publishing history** of the text itself, its **different editions** (with different entries, definitions, etc), its **translations**, **manuscripts**, what about **individual copies in libraries**?
- What about the **texts/corpora** that are **cited as attestations**, citations to scholarly works?
- For some of these there already exist generic vocabularies (**Dublin Core**, **Prov-O**, **CITO**) which can provide solutions, others have to be adapted to the dictionary domain.

In fact we still need a conceptual framework for integrating together different levels of description. FRBR will provide this...and this will eventually bring us back to CIDOC-CRM

FRBR

- Stands for **Functional Requirements for Bibliographic Records**: an entity relationship model intended for the classification of intellectual products in **bibliographic databases** and **library catalogues**.
- It introduced an important distinction in terms of how we can describe intellectual products. We can refer to such products at four different levels of description. Namely, at the level of **Work**, **Expression**, **Manifestation**, and **Item**.
- We use the version of this distinction given in the **CIDOC-CRM aligned LRM** ontology.

Work and Expression

- **Work**: *"[C]omprises distinct intellectual ideas conveyed in artistic and intellectual creations such a poems stories or musical compositions. A work is the outcome of an intellectual process of one or more expressions."*
  - Note that in the case of dictionaries this would encompass the **TEI lexical view**.

- **Expression**: *"[C]omprises the intellectual or artistic realisations of works in the form of identifiable immaterial objects, such as texts, poems [...] or any combination of such forms. The substance of F2 Expression is signs."*
  - In the case of dictionaries we claim that this description encompasses the **TEI editorial view**.

Manifestation and Item

- **Manifestation**, ”[C]omprises products rendering one or more Expressions. A Manifestation is defined by both the overall content and the form of its presentation. The substance of F3 Manifestation is not only signs, but also the manner in which they are presented to be consumed by users, including the kind of media adopted[...] An instance of F3 Manifestation typically incorporates one or more instances of F2 Expression representing a distinct logical content and all additional input by a publisher such as text layout and cover design”
  - In the case of dictionaries F3 Manifestation encompasses the **TEI typographic view**
- The **Item** class: “*[C]omprises physical objects*” such as specific physical copies of dictionaries kept at libraries or academic institutions.
  - This class is associated with the kind of metadata information that is usually contained within the **TEI header element**.

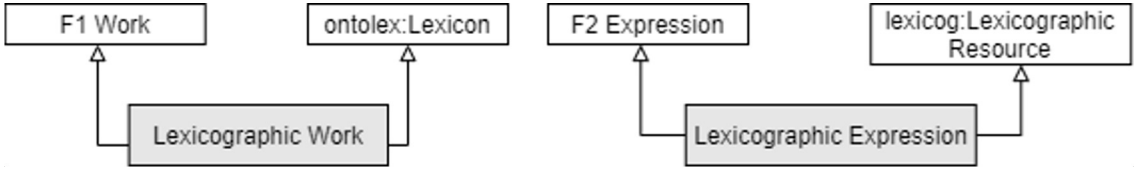
Bridging FRBRoo and OntoLex

- We propose a number of new classes and properties to bridge together LRM (and CIDOC-CRM) and OntoLex-Lemon/Lexicog.
- **Lexicographic Work:** A subclass of the FRBRoo class **F1 Work** and the Ontolex-Lemon class **Lexicon**. It comprises concepts or combinations of concepts for representing/describing the lexicon for a given language community or communities or domain.
    - As **F1 Work** is a subclass of the CIDOC-CRM class **E89 Propositional Object** we can view individuals of **Lexicographic Work** as sets of **propositions about lexemes and related linguistic concepts belonging to a lexicon.**

Bridging FRBRoo and OntoLex

- **Lexicographic Expression:** A subclass of the FRBRoo class **F2 Expression** and the lexicog class **Lexicographic Resource:** The class comprises an intellectual realisation of the description of a lexicon as a structured text.
  - In other words it is a text viewed apart from **a specific typographic realisation**: a sequence of words that has an **additional organisation** in terms of entries, senses (defined as a sub-part of a lexicographical article that discusses a meaning of a lexical unit), forms, etc.

Bridging FRBRoo and OntoLex



Asserting the Lexical View

- In our approach, we view a lexicographic entry as a series of statements making claims about different linguistic phenomena, about the lexicon of a language, as well a structural component of a text. In this we elaborate on previous work in both OntoLex and in CIDOC/FRBRoo.
- By modelling a dictionary as consisting of different levels of information, we can explicitly represent these as **hypotheses** (using named graphs or nanopublications).
- This comes in especially useful when it comes to combining together **etymologies**.

Modelling Citations and Annotations

By forcing us to **explicitly model our data** in terms of Subject-Predicate-Object triples RDF encourages us to think in terms of simple **declarative truth claims**: i.e., they make the preceeding considerations more salient. This is even more true wrt RDFS and OWL as these are much more expressive formal languages (OWL is a of description logic) and enable us/encourage us to make the meanings of our data much more '**explicit**'

The advantage of making this distinction is that it makes these different kinds of information more easily findable and queryable using the Semantic Web Query Language **SPARQL** for example.

Conclusions

The work presented though based on numerous case studies is still largely theoretical. The idea now is to move beyond the proof of concept stage and work with real use cases.

I am currently collaborating on a Portuguese national project called MorDigital, led by Professor Costa for digitising the Morais dictionary in TEI and OntoLex where I think the approach presented here could be useful.

These themes will also be explored in the Italian national PNRR project H2IOSC.

Obrigado

Thanks to Rute for organising this event and hosting me in Lisbon. To Nexus Linguarum for funding my short term scientific mission to Portugal!

More Details on the Semantic Web

The features of OWL

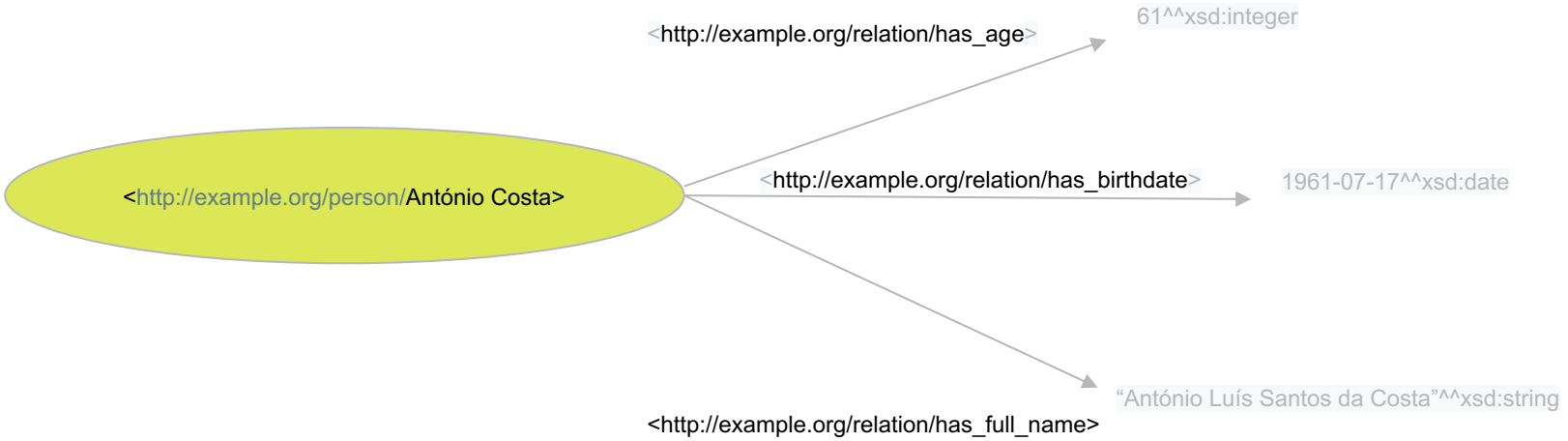
OWL allows us to **add constraints** to the definition of classes and properties that correspond to the features of the description logics we've looked at.

It also gives us a number of very useful properties:

- owl:sameAs (two individuals are the same) and owl:differentFrom (two individuals are different)
  - dbr:Leonardo_da_Vinci owl:sameAs dbpedia-ja:レオナルド・ダ・ヴィンチ
  - dbr:Leonardo_davincii owl:differentFrom dbr:Leonardo_da_Vinci
- owl:equivalentClass

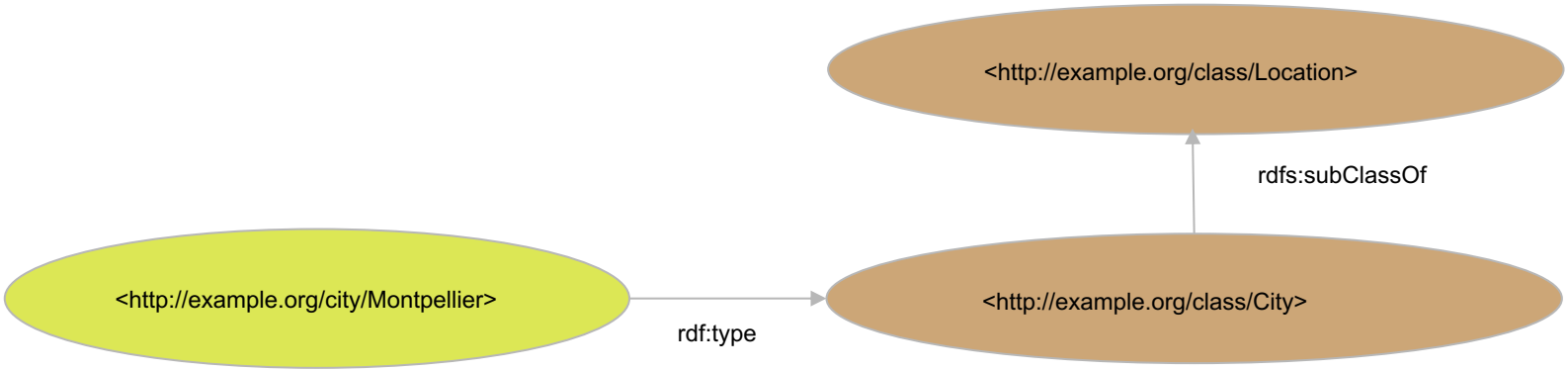
Resource Data Framework

RDF also allows us to assign literals, e.g., strings and numerical values to resources.



RDF and RDFS

RDF gives us a number of useful ‘built in’ classes and properties such as **rdf:Property**, **rdf:type**. RDF has been extended by a further standard **Resource Description Framework Schema** that gives us additional classes/properties such as **rdfs:Class**, **rdfs:subClassOf.**



*’rdf:Property’ is just an abbreviated way of writing the full address <http://www.w3.org/1999/02/22-rdf-syntax-ns#Property>

Web Ontology Language (OWL)

OWL is a knowledge representation language for the semantic web that is built on top of RDF

...or rather it's a family of such languages

The latest version is called OWL2 and was released as a W3C recommendation in late 2009

It is the most ontology language for the Semantic Web, it happens to be well known and popular language for writing, publishing and reasoning with ontologies

The features of OWL

- OWL allows us to specify two types of properties:
  - Object properties: binary relations holding between instances of classes
    - **dbr:Pisa dbo:region** dbr:Tuscany
  - Datatype properties: binary relations between class instances and RDF literals and XML Schema datatypes
    - **dbr:Pisa dbp:name “Pisa”@en**
    - **dbr:Pisa dbo:populationTotal "90834"^^xsd:nonNegativeInteger**

The features of OWL

We can specify transitivity and symmetry of properties

- We can also ensure that roles are functional (*if aRb and aRc then b=c*) and inverse functional (*if bRa and cRa then b=c*),

  dbpedia:John_Lennon ex:isFatherOf dbpedia:Julian_Lennon

We can also specify that one property is the inverse of another (if aRb then $bR^{-}a$)

- e.g., hasFather and isFatherOf (=$hasFather^{-}$).

The features of OWL

OWL also allows us to encode the sort of property restrictions that we saw in the section on description logics

- **allValuesFrom** and **someValuesFrom** encode ∀ and ∃ respectively
- **owl:cardinality**, **owl:maxCardinality**, and **owl:minCardinality** encode the quantifiers **≤ *n***, **≥*n***

OWL Tools

There exist numerous reasoners (**FaCT++**, **HermiT**, **Pellet**, and **Racer**) and ontology editors including the **NeOn toolkit**, **TopBraid** (a commercial product) and the **Fluent Editor** (which uses Controlled Natural Language in its interface) for OWL

To date the most popular tool for OWL is Stanford University's **Protégé**, a free open source **ontology editor**

If at first you don't succeed...

This example involves **Dante Alighieri**. It is intended to show how two authoritative lexical sources can **disagree** on the *meaning* of a citation. It revolves around the following two Italian homonyms:

- *riprovare* 'to try something again' (from *provare* 'to try' and the prefix *ri-* which adds the sense of repetition); call this **riprovare1.**
- *riprovare* 'to scold, rebuke' (in this sense it is cognate with the English verb *reprove*); call this **riprovare2**.

If at first you don't succeed

- The **motto** of the 16th century ***Accademia del Cimento*** "provare e riprovare", *try and try again*, captured the spirit of scientific endeavour promoted by that organisation. I.e., **riprovare1** *is attested by* **the AdC motto**
- **Dante's Paradiso** (Par. III, 1-3) contains a passage *attesting* to **riprovare2**, i.e.,
  - 'Quel sol che pria d'amor mi scaldò 'l petto,
    di bella verità m'avea scoverto,
    provando e riprovando, il dolce aspetto'
    (*That Sun, which erst with love my bosom warmed/ Of beauteous truth had unto me discovered/By proving and reproving, the sweet aspect*.)

Treccani v. Battaglia

The two authoritative Italian-language lexicographic resources, *il vocabolario Treccani* and *il Grande Dizionario della Lingua Italiana* (GDLL) treat these homonyms and the previous sources as follows:

- *Treccani's entry for **riprovare1** cites the AdC motto as attesting to the entry (see http://www.treccani.it/vocabolario/riprovare1)*
- *Treccani's entry for **riprovare2** cites Par. III, 1-3. as attesting to the entry (see http://www.treccani.it/vocabolario/riprovare2)*
- *GDLL's entry for **riprovare1** cites Par. III, 1-3. and the AdC motto as attesting to the entry*

Attestations and Citations

There are a number of truth claims here that we can list as follows:

1. Treccani's entry for **riprovare1** cites the AdC motto
2. Treccani's entry for **riprovare2** cites Par. III, 1-3.
3. GDLL's entry for **riprovare1** cites Par. III, 1-3.
4. **riprovare1** is attested by Par. III, 1-3.
5. **riprovare2** is attested by Par. III, 1-3.
6. **riprovare1** is attested by AdC

Attestations and Citations

There are a number of truth claims here that we can list as follows:

1. Treccani's entry for **riprovare1** cites the AdC motto
2. Treccani's entry for **riprovare2** cites Par. III, 1-3.
3. GDLL's entry for **riprovare1** cites Par. III, 1-3.
4. **riprovare1** is attested by Par. III, 1-3.
5. **riprovare2** is attested by Par. III, 1-3.
6. **riprovare1** is attested by AdC

Statements 1-3 describe *citations* at the level of bibliography.

Attestations and Citations

There are a number of truth claims here that we can list as follows:

1. Treccani's entry for **riprovare1** cites the AdC motto
2. Treccani's entry for **riprovare2** cites Par. III, 1-3.
3. GDLL's entry for **riprovare1** cites Par. III, 1-3.
4. **riprovare1** is attested by Par. III, 1-3.
5. **riprovare2** is attested by Par. III, 1-3.
6. **riprovare1** is attested by AdC

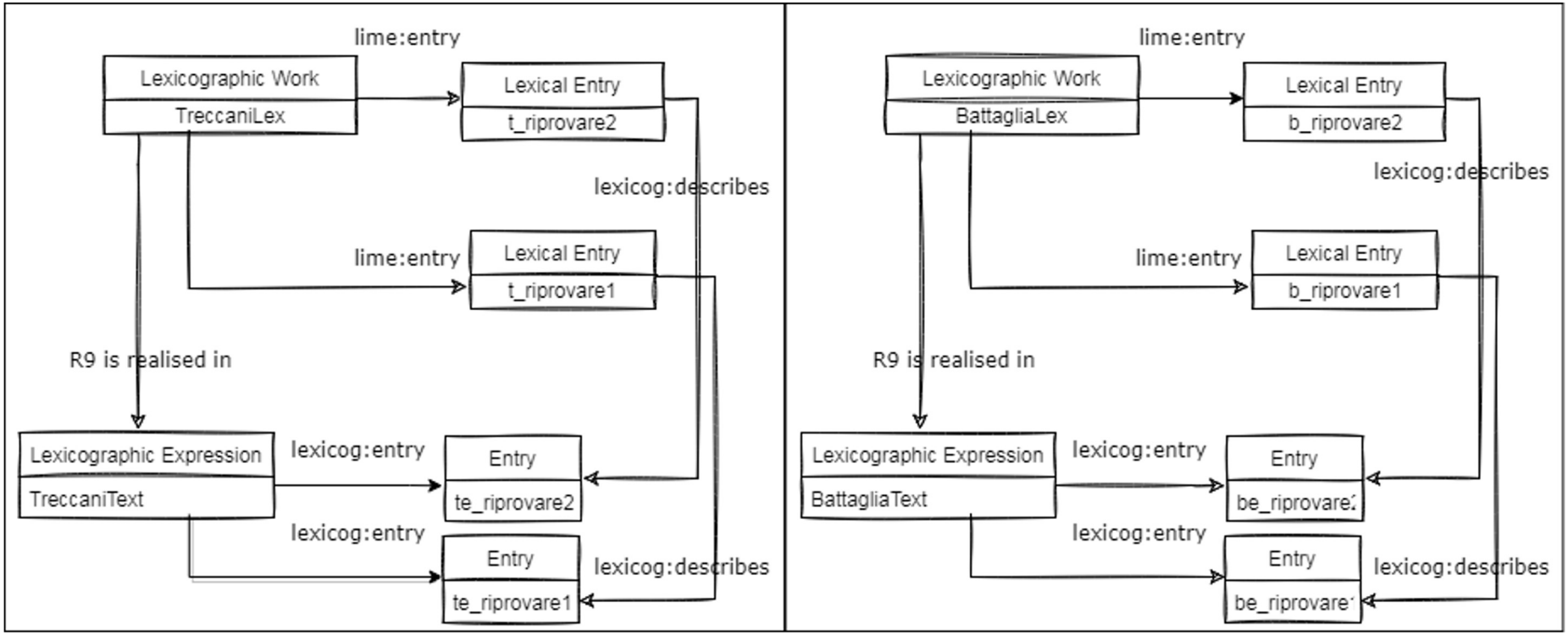
Statements 4-6 describe *attestations* at what we might call a *lexical* level.

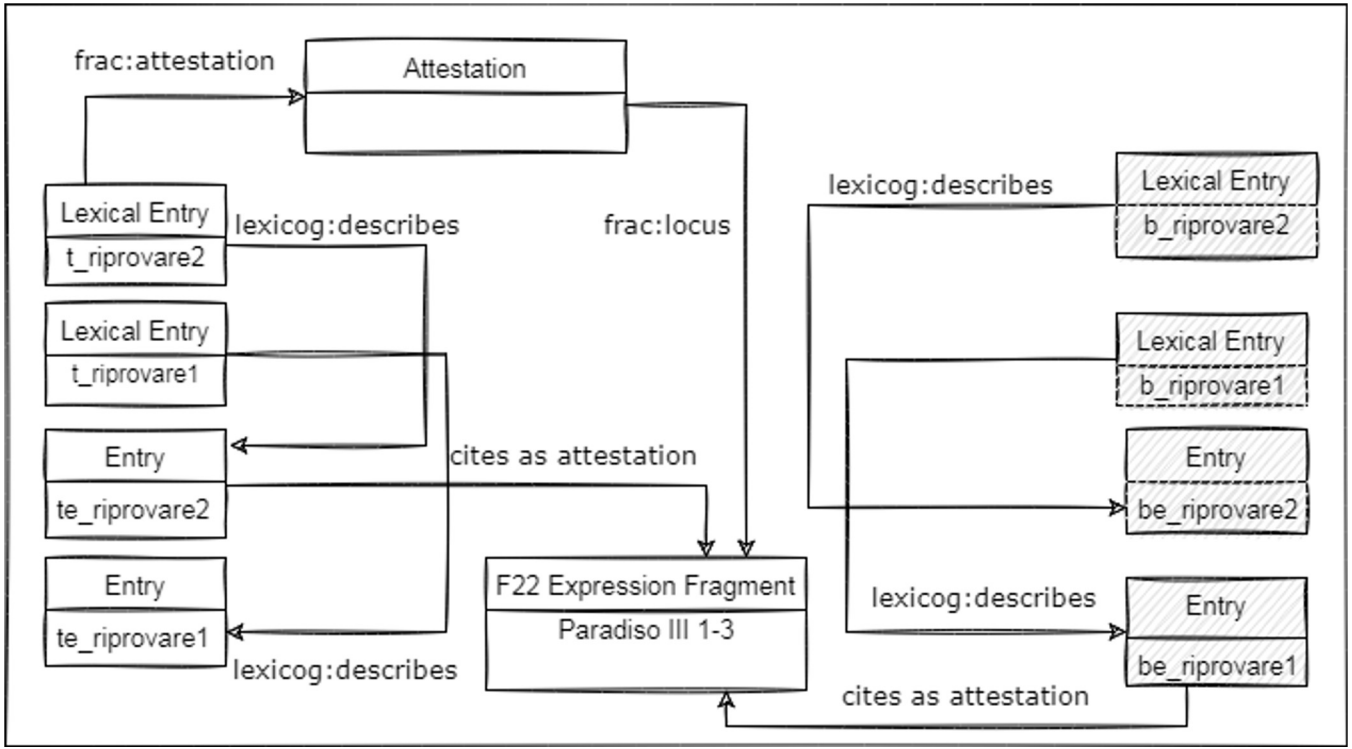
Attestations and Citations

There are a number of truth claims here that we can list as follows:

1. Treccani's entry for **riprovare1** cites the AdC motto
2. Treccani's entry for **riprovare2** cites Par. III, 1-3.
3. GDLL's entry for **riprovare1** cites Par. III, 1-3.
4. **riprovare1** is attested by Par. III, 1-3.
5. **riprovare2** is attested by Par. III, 1-3.
6. **riprovare1** is attested by AdC

Statement 3 is true, but its corresponding lexical claim, its related truth content, Statement 4 is false. Both these levels may be interesting independently of one another.


lime:entry
Lexicographic Work
TreccaniLex
Lexical Entry
t_riprovare2
lexicog:describes
lime:entry
Lexical Entry
t_riprovare1
R9 is realised in
Lexicographic Expression
TreccaniText
lexicog:entry
Entry
te_riprovare2
lexicog:entry
Entry
te_riprovare1
lexicog:describes
lime:entry
Lexicographic Work
BattagliaLex
Lexical Entry
b_riprovare2
lexicog:describes
lime:entry
Lexical Entry
b_riprovare1
R9 is realised in
Lexicographic Expression
BattagliaText
lexicog:entry
Entry
be_riprovare
lexicog:entry
Entry
be_riprovare
lexicog:describes


frac:attestation
Attestation
Lexical Entry
t_riprovare2
lexicog:describes
frac:locus
lexicog:describes
Lexical Entry
b_riprovare2
Lexical Entry
t_riprovare1
Lexical Entry
b_riprovare1
Entry
te_riprovare2
cites as attestation
Entry
be_riprovare2
Entry
te_riprovare1
F22 Expression Fragment
Paradiso III 1-3
lexicog:describes
Entry
be_riprovare1
lexicog:describes
cites as attestation