Università Ca'Foscari Venezia

UNIVERSITÀ DEGLI STUDI FIRENZE

Istituto di Linguistica Computazionale "Antonio Zampolli"
Consiglio Nazionale delle Ricerche

Modello dei lessici e ecosistema dei dati

Valeria Quochi - Michela Bandini

Venezia, 20 settembre 2024 - PRIN 2017 Lingue e culture dell'Italia antica: linguistica storica e modelli digitali

Lessici computazionali

- Definizioni:
  - Lessici Computazionali: risorse linguistiche usate per l'elaborazione automatica della lingua.
  - Dizionari Elettronici: dizionari digitalizzati (retro-digitalizzati o nativi) per uso umano e tecnologico.
- Ruolo e Funzioni: sono il “ponte” tra diverse risorse linguistiche e tecnologie:
  - Input per tecnologie di elaborazione del linguaggio naturale → rappresentazione (digitale/formale) permette o facilita l’elaborazione automatica della lingua.
  - Indicizzazione e normalizzazione dei testi.
  - Astrazione linguistica e analisi delle parole.
- Rappresentazione delle parole:
  - Strutturazione granulare di lessemi.
  - Caratteristiche fonetiche, morfologiche, sintattiche, semantiche sono esplicitate.
  - Relazioni lessicali, semantiche ed etimologiche fra lessemi sono esplicitate.

Modelli di rappresentazione dell'informazione lessicale

- La struttura delle informazioni linguistiche dipende dal modello lessicografico applicato dal costruttore del dizionario:
  - Modelli creati e strutturati per facilitare la decodifica da parte dell'applicazione informatica che ne farà uso.
  - Modelli basati su/ispirati a teorie computazionali e/o semantico-lessicali.
  - Modelli mirati alla rappresentazione di fenomeni linguistici specifici.

Principi cardine della rappresentazione lessicale

- **Entrata Lessicale**: i dati lessicali sono solitamente organizzati attorno al concetto di "entrata lessicale", che include informazioni su ortografia, PoS, flessione, e il/i significato/i, pronuncia.
- **Struttura Gerarchica e/o Relazionale**: le banche dati lessicali si basano su queste relazioni con lo scopo di modellare le relazioni tra le parole (e.i. sinonimi, contrari, iponimi/iperonimi, meronimi, etc.) e definire i loro significati e contesti.
- **Proprietà Semantiche e Sintattiche**: principi linguistici comuni, come ruoli semantici, strutture argomentali, categorie sintattiche per fornire profondità alla rappresentazione delle parole.
- **Metadati Standardizzati**: alcune risorse linguistiche seguono standard di metadati o vocabolari controllati e condivisi nella comunità scientifica di riferimento (e.g. TEI, LMF, OntoLex-Lemon, lexinfo, OLiA, etc.) per fornire uniformità nella rappresentazione.

TEI LEX-O & ONTOLEX-LEMON

Dizionario come
testo

- specifica tecnica dello schema `TEI` (*Text Encoding Initiative*)
- raccomandazioni per codifica di dizionari elettronici
- necessità di uniformare processo e formato di creazione di dizionari retro-digitalizzati

Dizionario come
fonte di conoscenza

- standard de-facto per la
- rappresentazione di informazione lessicale nel **Web Semantico**
- modello di rappresentazione di lessici (primariamente) computazionali basati su ontologie e vocabolari controllati che favoriscono l'interoperabilità semantica

Introduzione al web semantico

COSA E' IL WEB SEMANTICO?

- Uno spazio aperto di conoscenza formalizzata e condivisa disponibile via web anche per l'elaborazione automatica.
- Descrizione formale del significato delle risorse attraverso metadati e formalismi (rispetto a `XML`)

SEMANTICA E DATI:

- La **semantica del dato** è esplicita: anche i software possono "comprendere" il significato delle informazioni e manipolarle e integrarle
- Utilizza triple `RDF` (Soggetto - Predicato - Oggetto) per rappresentare e integrare le info

IDENTIFICAZIONE E STRUTTURA DATI

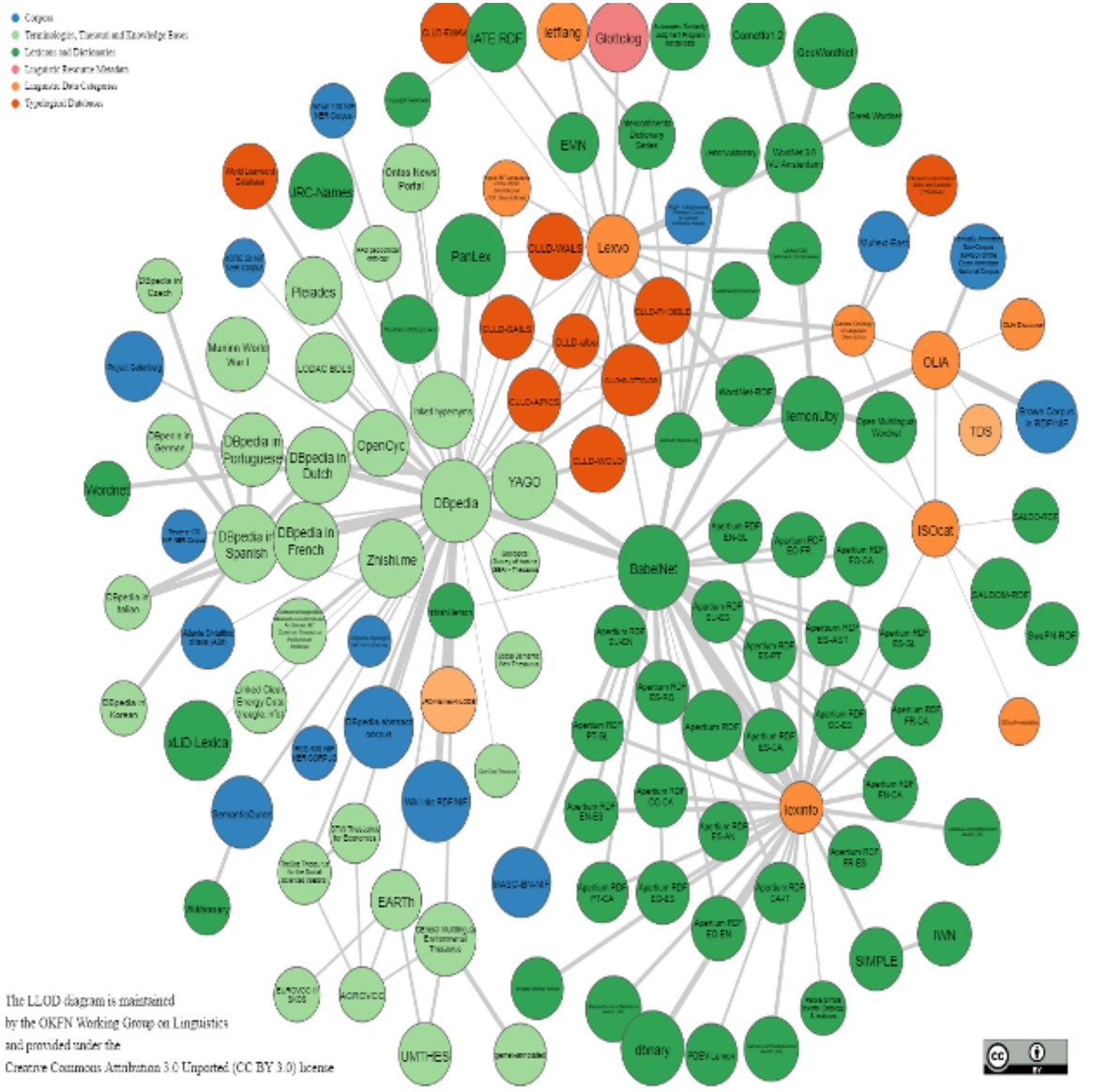
- L'informazione è atomizzata: ogni elemento, proprietà e relazione ha un **identificativo univoco dereferenziabile** (`URI`)
- Insiemi strutturati di informazioni e di **regole d'inferenza** (con le ontologie) che permettono alla macchina di **automatizzare i ragionamenti** (Grafi orientati con archi etichettati)

ACCESSO ALLA CONOSCENZA E RIUSO

- La diffusione di internet permette di **accedere a tutta la conoscenza** distribuita nella rete, purchè sia collegata
- Garantita l'**interoperabilità** tra risorse diverse e **riuso** di informazioni usando vocabolari comuni e condivisi

Linked (Open) Data e applicazioni

- Modalità di pubblicazione sul web di dati strutturati e collegati basati su standard del Web Semantico: `RDF`, `URI` o `HTTP`.
- La pubblicazione sfrutta anche l'utilizzo di **Ontologie** e **Vocabolari** condivisi
- Ogni informazione codificata può essere riusata da altre risorse disponibili sul web (**interoperanbilità**)
- → Si ha una rete di risorse indipendentemente descritte ma collegate disponibili sul web
- Il Web Semantico e Linked Open Data hanno rivoluzionato l'approccio alla conoscenza: adottato rapidamente anche dalle **Digital Humanities**.
- Grandi potenzialità per le Digital Humanities, con applicazioni significative come l'**epigrafia digitale**, che sfrutta le tecnologie semantiche per migliorare l'analisi e la gestione dei testi antichi.

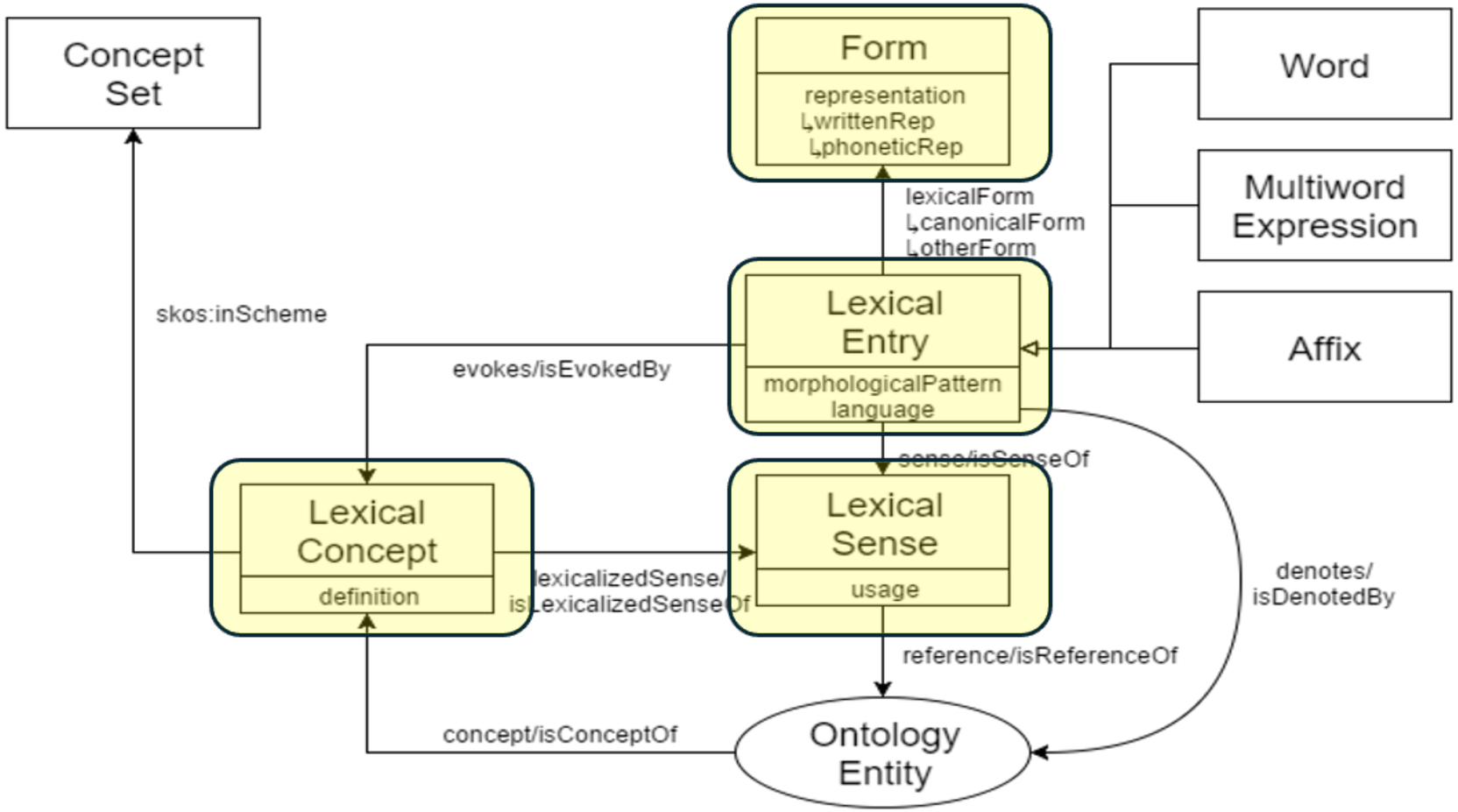


CC-BY. "Open Linguistics Working Group (2016), The Linguistic Linked Open Data cloud diagram, version of 2016-05-24, http://linguistic-lod.org/llod-cloud"

ONTOLEX-LEMON

- *Lexicon Model for Ontologies*: formato de-facto standard **W3C** per la rappresentazione e pubblicazione di lessici (“dictionary-like information”) come triple `RDF` sfruttando i Linked Data.
- Modello di rappresentazione di lessici (primariamente) computazionali, in connessione a ontologie, basato su 5 principi fondanti:
  - **tecnologia Linked Data**: formati `OWL`, `RDF`
  - **multilingue**: rappresenta info di molte lingue
  - **semantica referenziale/denotazionale**: il significato delle parole è dato dalla presenza di un’ontologia
  - **formato aperto**: gratuito e flessibile, estendibile
  - riuso di standard esistenti
- Per i lessici computazionali esiste già una comunità attiva dal 2006, modelli e vocabolari già testati da poter riutilizzare (i.e, estensioni)

ONTOLEX-LEMON: Core Model



cfr.: McCrae&Declerc “Introduction to the OntoLex-Lemon Model”- http://esslli2018.folli.info/wp-content/uploads/Introduction-to-the-OntoLex-Lemon-Model.pdf

Il lessico multilingue di ItAnt

- Progettato in base alle specificità delle **Restsprachen** e ai requisiti degli storici, definiti progressivamente attraverso cicli di confronto.
- Conforme allo **standard OntoLex-Lemon** e a sue estensioni non ufficiali ma già in uso (`lemonEty`, `FrAC`).
- Utilizza, ove possibile, altri vocabolari e modelli standard per massimizzare l'**interoperabilità** e il riuso futuro.
- Promuove la codifica di collegamenti a **risorse esterne** (e.g. *LiLa*)

Entrata lessicale

Il lessico multilingue di ItAnt

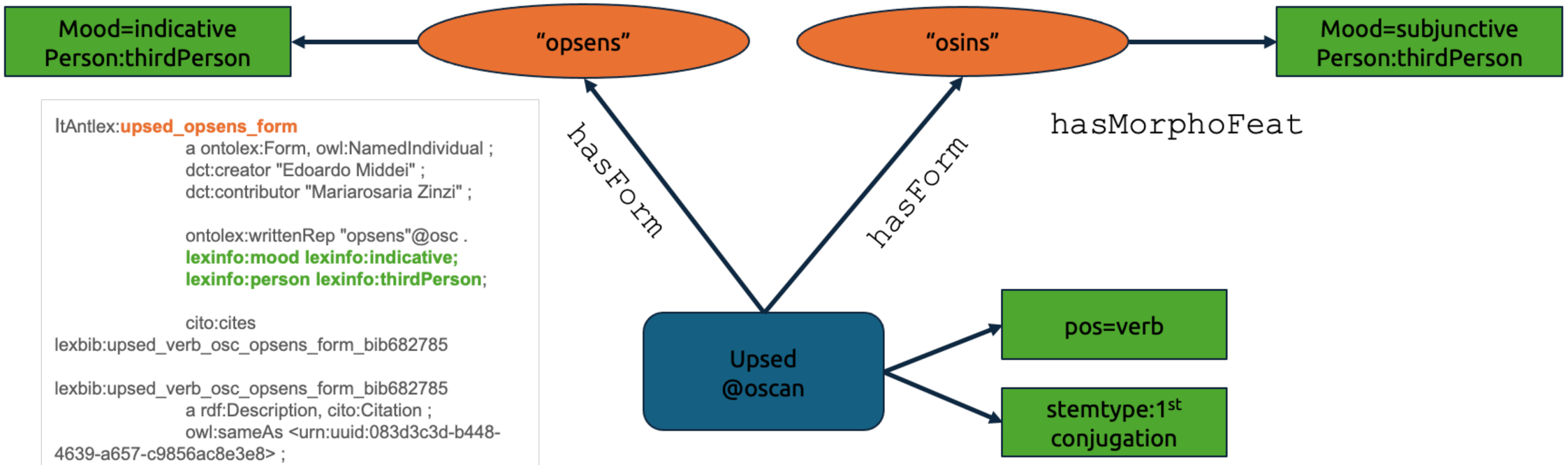
- Rappresenta una unità di analisi del lessico ed è costituita / descritta da:
  - una categoria grammaticale;
  - un eventuale paradigma morfologico;
  - una etimologia;
  - una o più forme grammaticalmente correlate;
  - un insieme di significati di base (nei lessici ItAnt 1 Entrata : 1 Senso);
  - può essere una parola monorematica, polirematica o un affisso.

```
ItAntlex:upsed_entry
        a ontolex:Word;
        dct:creator "Edoardo Middei" ;
        dct:contributor "Mariarosaria Zinzi" ;
        ns:term_status "draft";
        rdfs:label "upsed"@osc ;
        lime:language "osc" ;
        lexinfo:partOfSpeech lexinfo:verb ;
        itant:stemtype "1 conjugation" ;
        ontolex:sense ItAntlex:upsed_sense1 ;
        ontolex:evokes
ItAntlex:toWorkToil_semfield_concept ;
 <!-- Forms list -->
        ontolex:lexicalForm
ItAntlex:upsed_opsens_form,
ItAntlex:upsed_osins_form,
ItAntlex:upsed_upsed_form  .
```

Forma e proprietà grammaticali

Il lessico multilingue di ItAnt

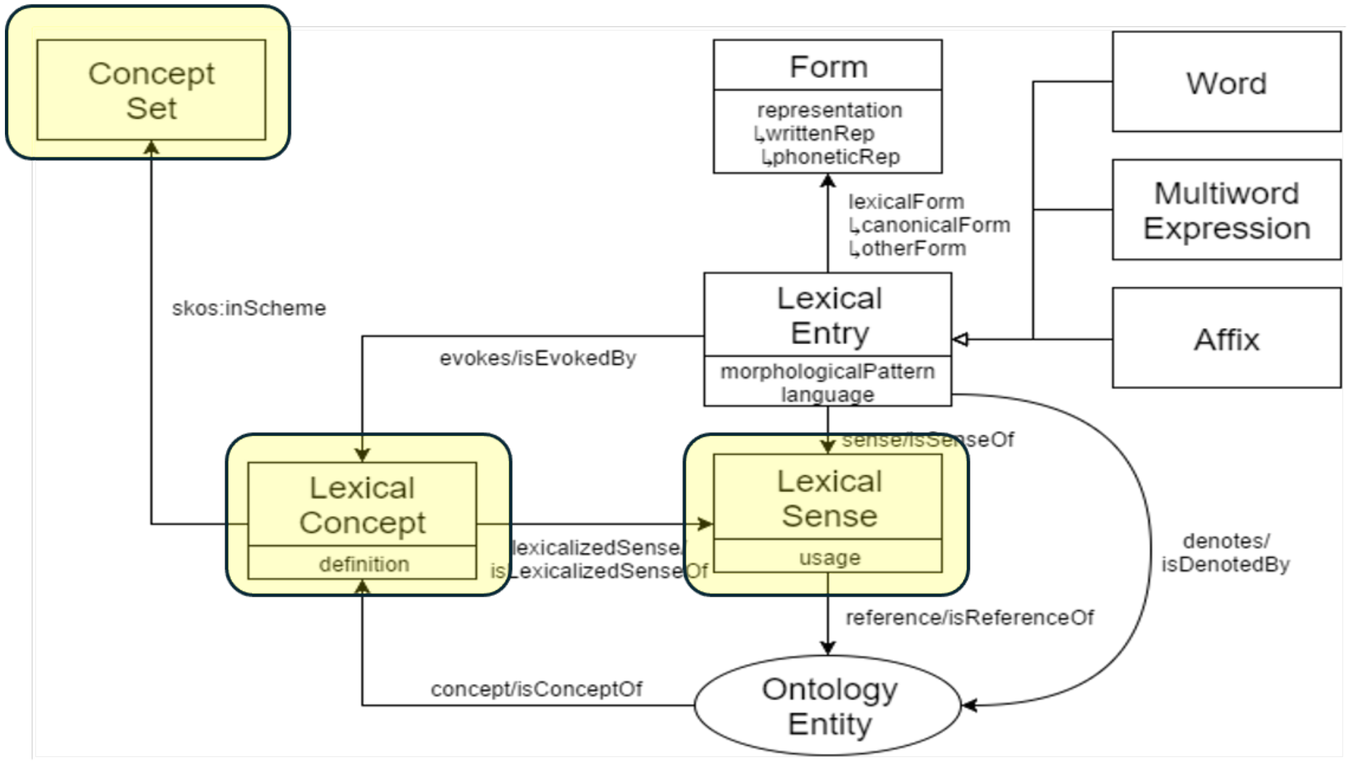
osco *upsed*




```
ItAntlex:upsed_opsens_form
                a ontolex:Form, owl:NamedIndividual ;
                dct:creator "Edoardo Middei" ;
                dct:contributor "Mariarosaria Zinzi" ;

                ontolex:writtenRep "opsens"@osc .
                lexinfo:mood lexinfo:indicative;
                lexinfo:person lexinfo:thirdPerson;

                cito:cites
lexbib:upsed_verb_osc_opsens_form_bib682785

lexbib:upsed_verb_osc_opsens_form_bib682785
                a rdf:Description, cito:Citation ;
                owl:sameAs <urn:uuid:083d3c3d-b448-
4639-a657-c9856ac8e3e8> ;
                dct:contributor "Mattiocco" ;
                dct:title "Centri fortificati vestini" ;
                dct:date "1986" .
```

Sensi e campi semantici

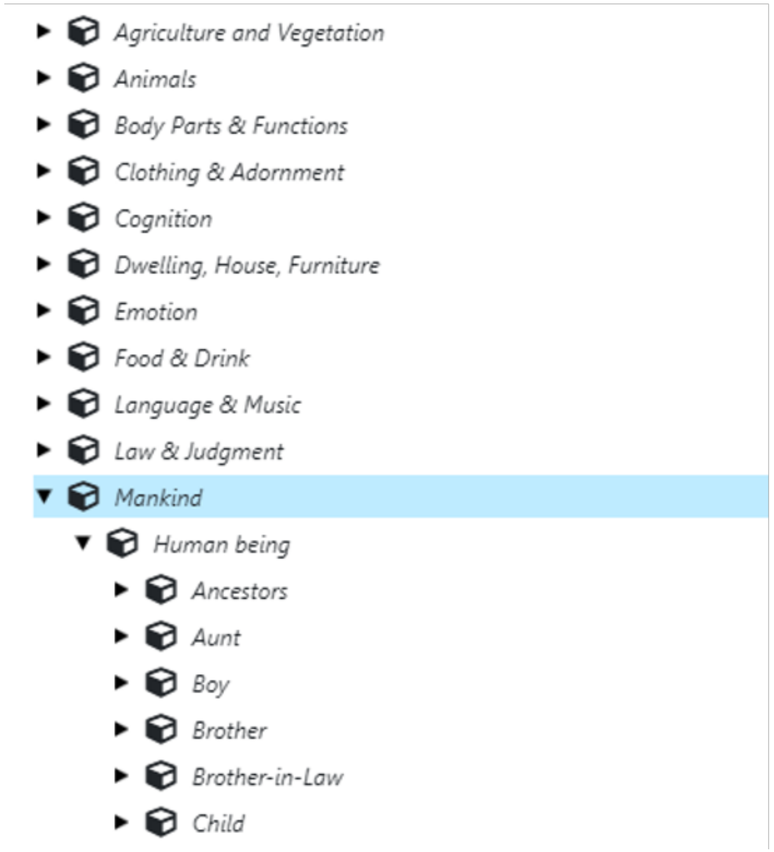
Il lessico multilingue di ItAnt



- Campi semantici formalizzati come `LexicalConcept`
- Lista di `LexicalConcept` è rappresentata come una tassonomia **SKOS** (*Simple Knowledge Organization System*)
- Standard sviluppato da **W3C** per rappresentazione di vari schemi di classificazione (e.i. tassonomie, thesauri)

Tassonomia dei campi semantici (Buck 1949)

Il lessico multilingue di ItAnt



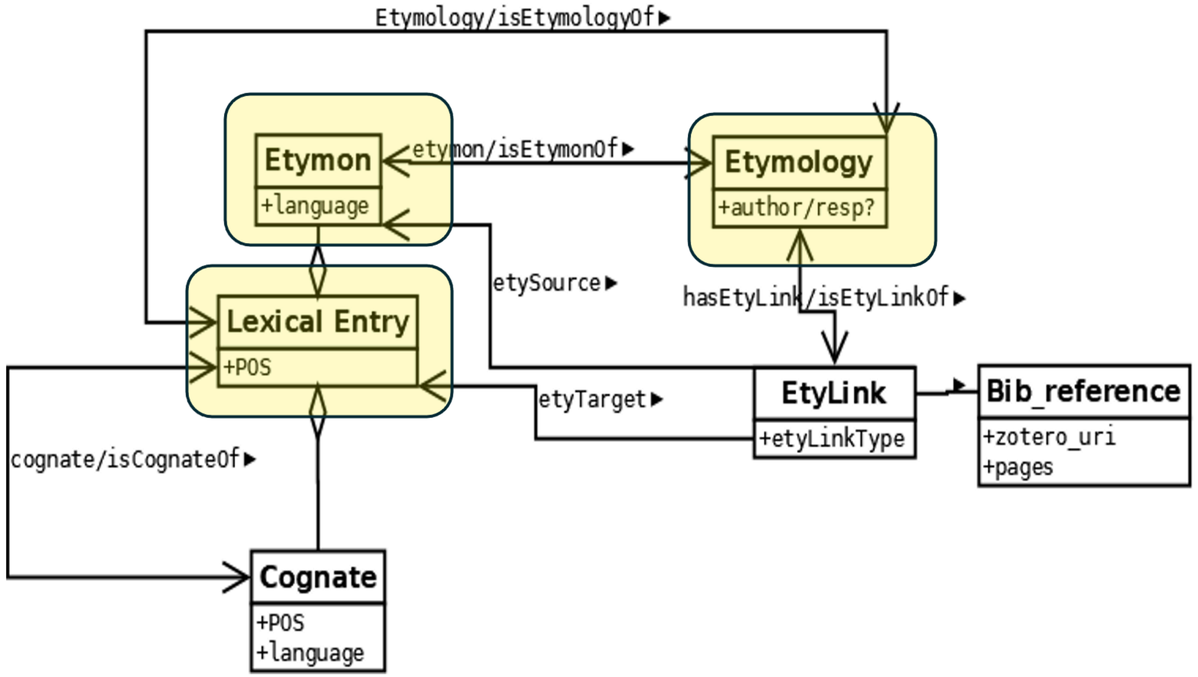

```
semfield:LexO_2023-10-2307_38_43_802
        a owl:Meaning, skos:Concept, ontolex:LexicalConcept ;
        dcterms:created "2023" ;
        dcterms:creator "vquochi" ;
        skos:prefLabel> "Mankind"@en;
        owl:sameAs <https://lrc.la.utexas.edu/lex/semantic/category/MK>;
        skos:broaderTransitive semfield:LexO_2023-10-2307_39_21_048,
                LexO_2023-10-2307_39_58_622, LexO_2023-10-2307_40_00_376,
                LexO_2023-10-2307_40_01_735, LexO_2023-10-2307_41_04_479,
                LexO_2023-10-2307_41_50_396, LexO_2023-10-2307_42_49_164,
                LexO_2023-10-2307_42_59_495, LexO_2023-10-2307_43_31_279,
                LexO_2023-10-2307_43_52_459, LexO_2023-10-2307_43_54_483,
                LexO_2023-10-2307_43_55_819, LexO_2023-10-2307_43_57_127,
                LexO_2023-10-2307_45_14_356, LexO_2023-10-2307_45_15_78,
                #...;
```

cfr. https://lrc.la.utexas.edu/lex/semantic

Etimologia e Cognates: LemonEty

Il lessico multilingue di ItAnt

→ **`LemonEty`** *OntoLex* (unofficial) extension

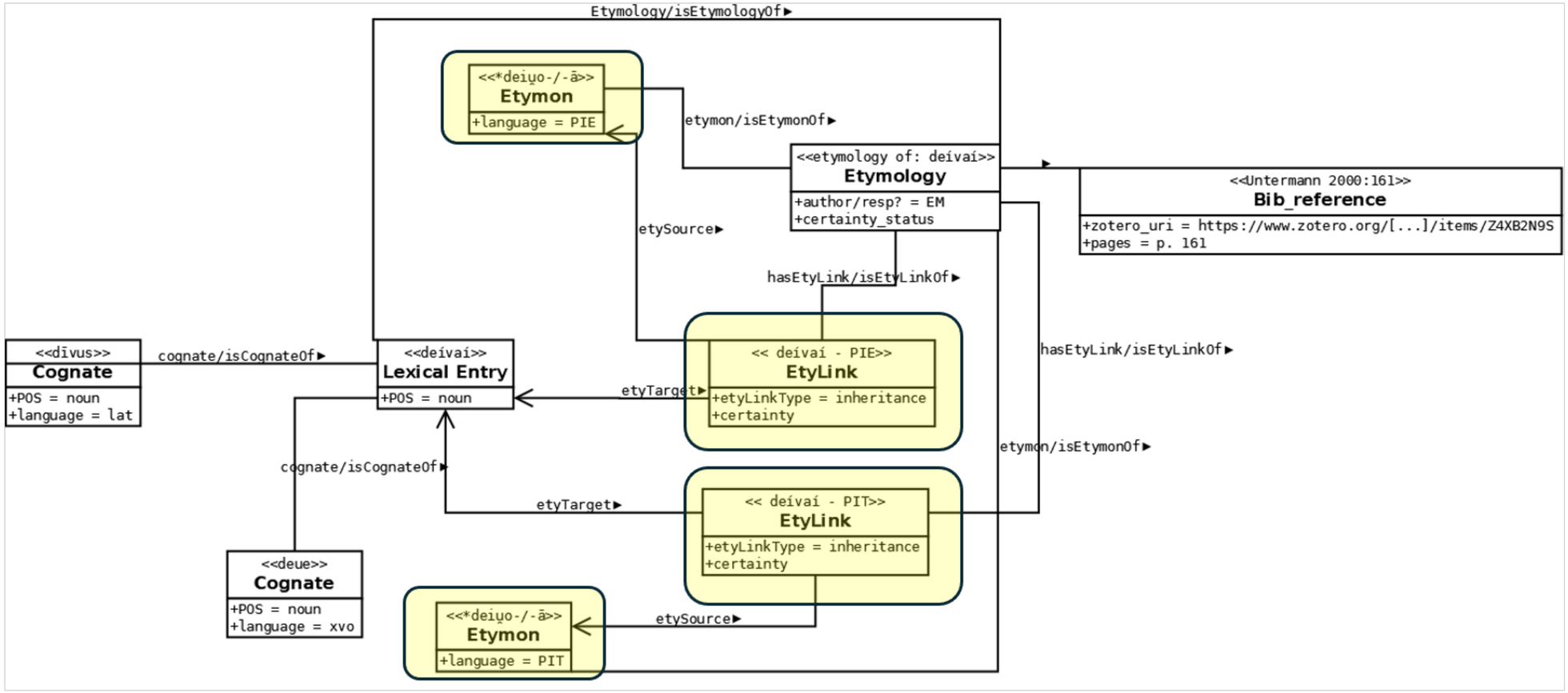


- I lessici ItAnt **non rappresentano etimologie complete**, ma codificano radici nelle proto-lingue all'interno della risorsa lessicale multilingue come entrate minimali
- **Non** si "linka" **un solo dizionario etimologico** esistente
- Ogni elemento lessicale è introdotto al solo scopo di descrivere la storia di una parola e non appartiene al lessico della lingua in oggetto ma è un membro della classe `Etymon` (sottoclasse di `ontolex:lexicalEntry`)

Etimologia e Cognates: LemonEty

Il lessico multilingue di ItAnt

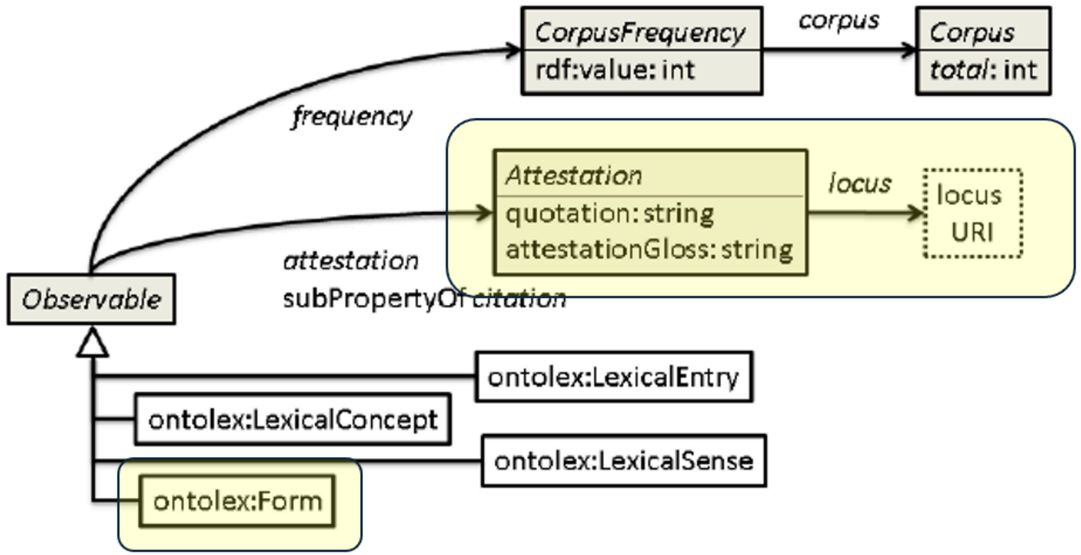
osco *deívaí*



Attestazioni: collegamento al corpus

Il lessico multilingue di ItAnt

→ **FrAC** *OntoLex* (unofficial) extension



- Collega una forma a una iscrizione o una parola (sequenza di caratteri) nell'edizione del corpus ItAnt
- Rappresenta altre info:
  - autore responsabile dell'attribuzione
  - forma ricostruita secondo convenzioni Leida
  - riferimenti bibliografici

Attestazioni: collegamento al corpus

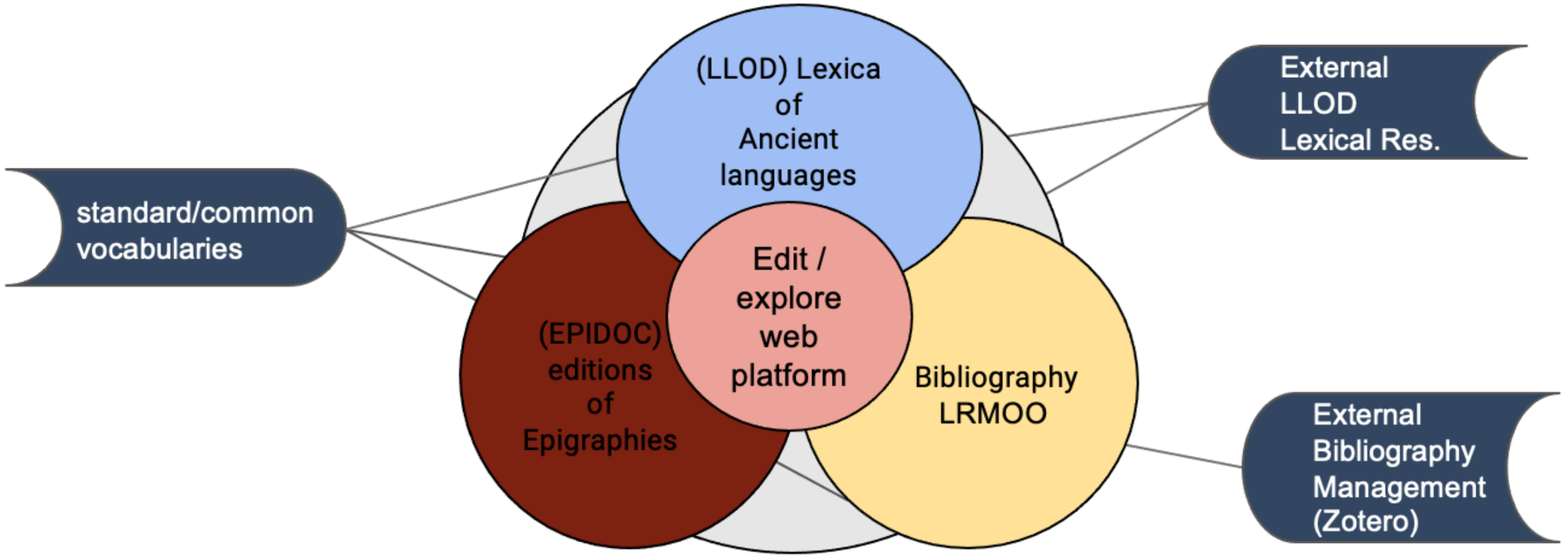
Il lessico multilingue di ItAnt

- Per il progetto ItAnt, esistono poche risorse Linked Open Data (5 stelle) utili a questo scopo.
- Due risorse lessicali interessanti e utili per specificare le informazioni etimologiche, per il Latino:

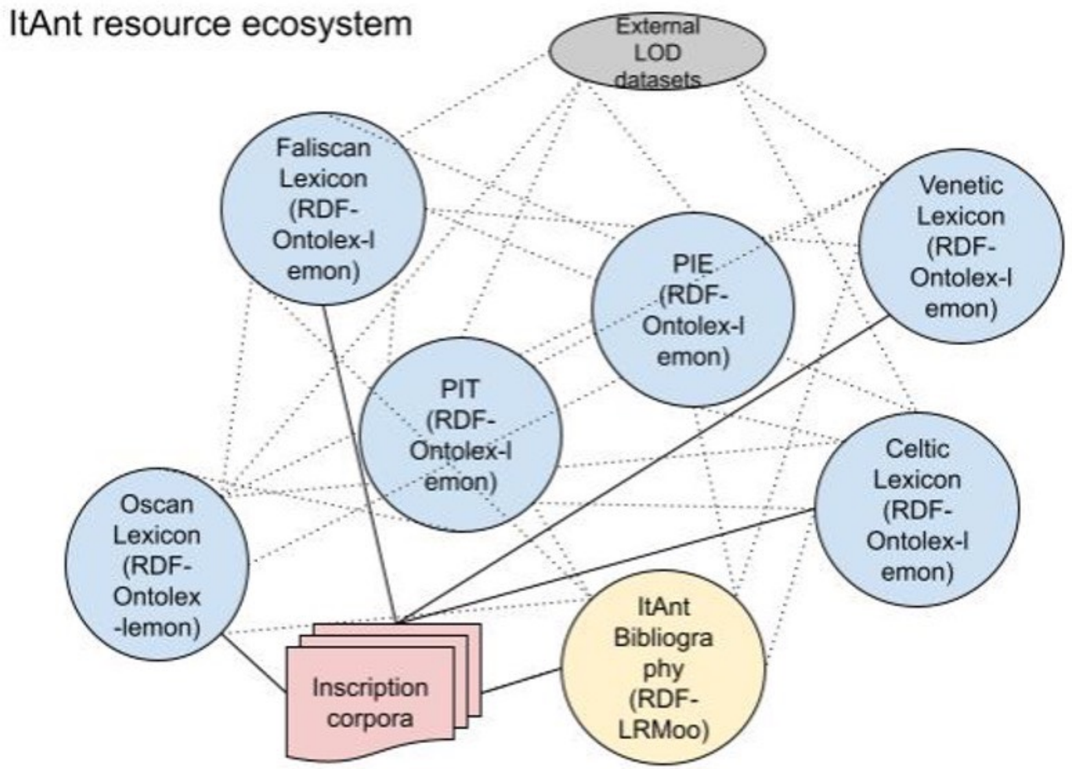
- **LiLa Lemma Bank**
  - http://hdl.handle.net/20.500.11752/OPEN-532
  - collezione di lemmi con proprietà morfosintattiche
  - usato per collegare i Cognates e le entrate@lat
- **The Etymological Dictionary of Latin and the other Italic Languages**
  - http://hdl.handle.net/20.500.11752/OPEN-532
  - usato per collegare gli etimi e le etimologie

Ecosistema Dati (Dig)ItAnt

- Un Ecosistema di dati connessi (lessici, testi, metadati, riferimenti bibliografici) e strumenti per la loro manipolazione e visione integrata.



Verso una “LOD-ificazione” totale



- Il **lessico computazionale** è allineato, sfruttando modello OntoLex-Lemon e rappresentazione `RDF`.
- Ma la **pubblicazione di corpora testuali** (in `TEI/EpiDoc`) sul Semantic Web è ancora in fase iniziale: si sta ancora decidendo quale approccio/modello seguire.

Verso una “LOD-ificazione” totale

- Abbiamo dunque svolto una **rvisione sistematica della letteratura** (Bandini, M. & Quochi, V. 2024) perché:
  - la maggior parte delle risorse linguistiche nella rete dei Linguistic Linked Open Data sono dizionari, lessici, thesauri, terminologie e vocabolari controllati;
  - la pubblicazione di corpora testuali come Linked Open Data è ancora **dibattuta e controversa**;
  - alcuni progetti recenti hanno iniziato a esplorare **diversi approcci e modelli per convertire e rappresentare i corpora testuali** come Linked Open Data: `NIF` (Hellmann et. ali, 2013) `POWLA`, (Chiarcos, 2022), `CoNLL-RDF` (Chiarcos, 2020);
- → per valutare i vantaggi e identificare il modello più adatto per convertire o rappresentare le iscrizioni in lingua antica in Linked Open Data: `POWLA`.

Bibliografia

- Bellandi, Andrea. 2022. «Le Risorse Linguistiche nell'era del Web Semantico. Un insieme di servizi informatici per la ge-stione di lessici e terminologie». AIDAinformazioni: Rivista di Scienze dell'Informazione 1–2.
- Bellandi, Andrea, Fahad Khan, Monica Monachini, e Valeria Quochi. 2022. «A LexO-server use case: Languages and Cultures of Ancient Italy». In , 16–17. Vilnius: Mykolas Romeris University.
- Quochi, Valeria, Andrea Bellandi, Fahad Khan, Michele Mallia, Francesca Murano, Silvia Piccini, Luca Rigobianco, Alessandro Tommasi, e Cesare Zavattari. 2022. «From Inscriptions to Lexicon and Back: A Platform for Editing and Linking the Languages of Ancient Italy». In LREC 2022 Workshop Language Resources and Evaluation Conference. Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022). Proceedings, 59–67. Marseille: European Language Resources Association.