Linking, publishing and evaluating

Linked Open Data for language resources

Francesco Mambrini
francesco.mambrini@unicatt.it

SCS Annual Meeting | Washington, DC | January 3, 2020

Table of Contents

Treebanks!

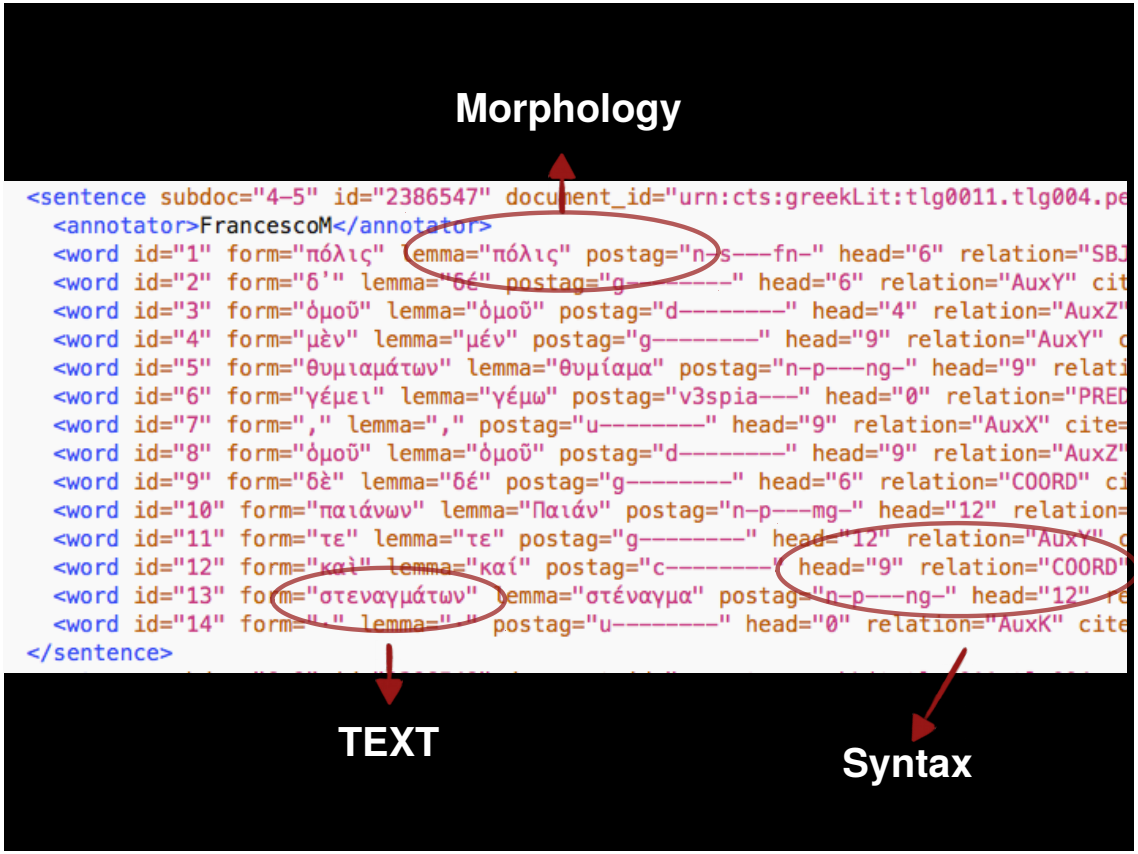


Figure: Morphosyntactic information stored in a XML file of the Ancient Greek and Latin Dependency Treebank.

Treebanks: a success story?

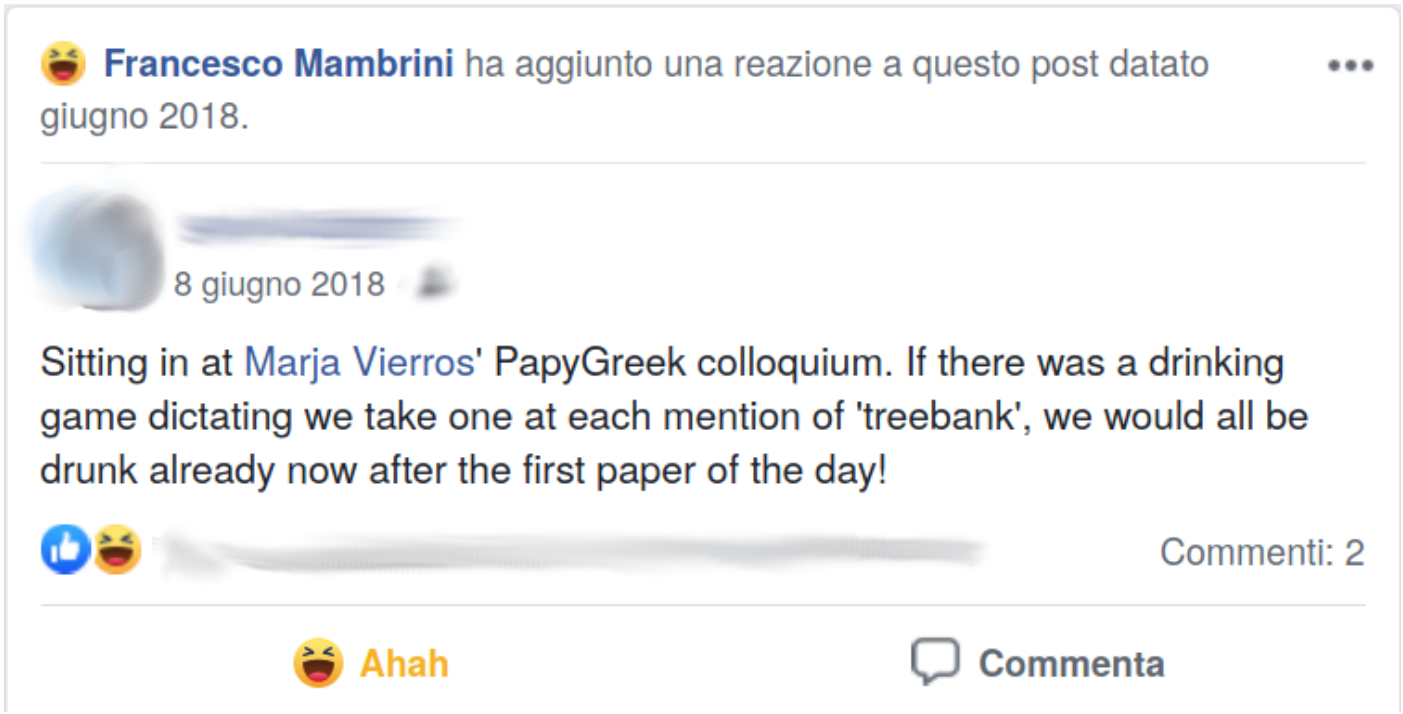


Figure: A comment posted on Facebook about the workshop of the PapyGreek Project, Helsinki 2018.

Open problems

sparseness: there is a multitude of projects involving linguistic annotation;

standardization: projects jealously hang on their guidelines and tagset and refusing to consider any form of standardization;

interoperability: no way to make morphosyntactic annotation interact with other levels of information (e.g. lexical resources);

usability: lack of general-purpose tools for annotating, manipulating and querying the data.

Use-evaluation-correction

A virtuous circle

Logeion Greek-Latin
@LogeionGkLat

Replying to @hashtagoras

@FrancMambr reminds me of first version of Prometheus Bound treebank, lots of wheat🤣

8:03 PM · Nov 14, 2019 · Twitter Web App

Francesco Mambrini @FrancMambr · Nov 15
Replying to @LogeionGkLat and @hashtagoras
yeah! and there were vipers everywhere... 😚😂

1 1

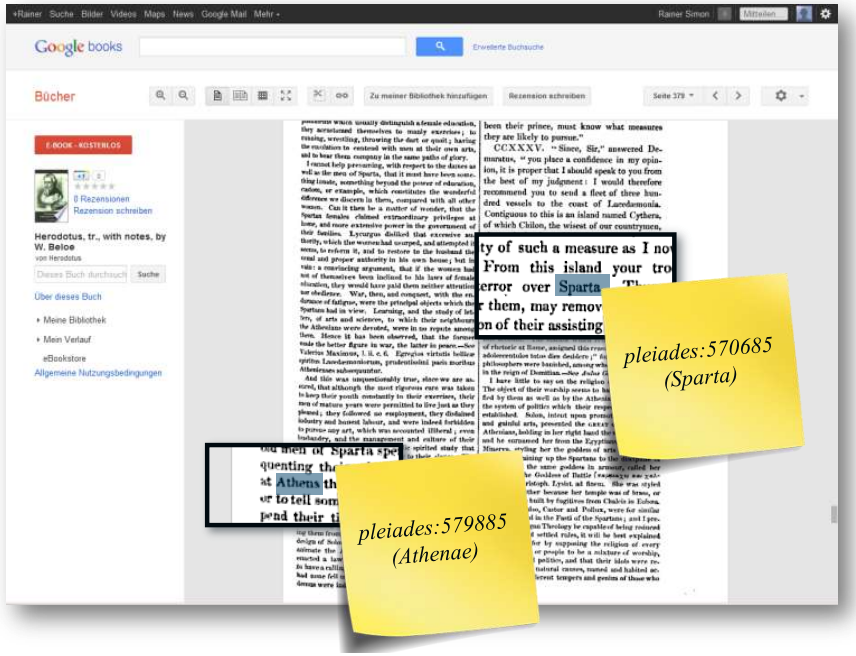
Logeion Greek-Latin @LogeionGkLat · Nov 15
Bonus points to the first person who is not Francesco or me who figures this out:-)

LOD and Semantic Web in the Classics

Pelagios
a LOD network of annotations

How? | Don't Unify the Model – Annotate!

- annotate place reference using gazetteer URIs from Pleiades
- publish annotation using the OAC vocabulary



The Pelagios model
Strengths and weaknesses

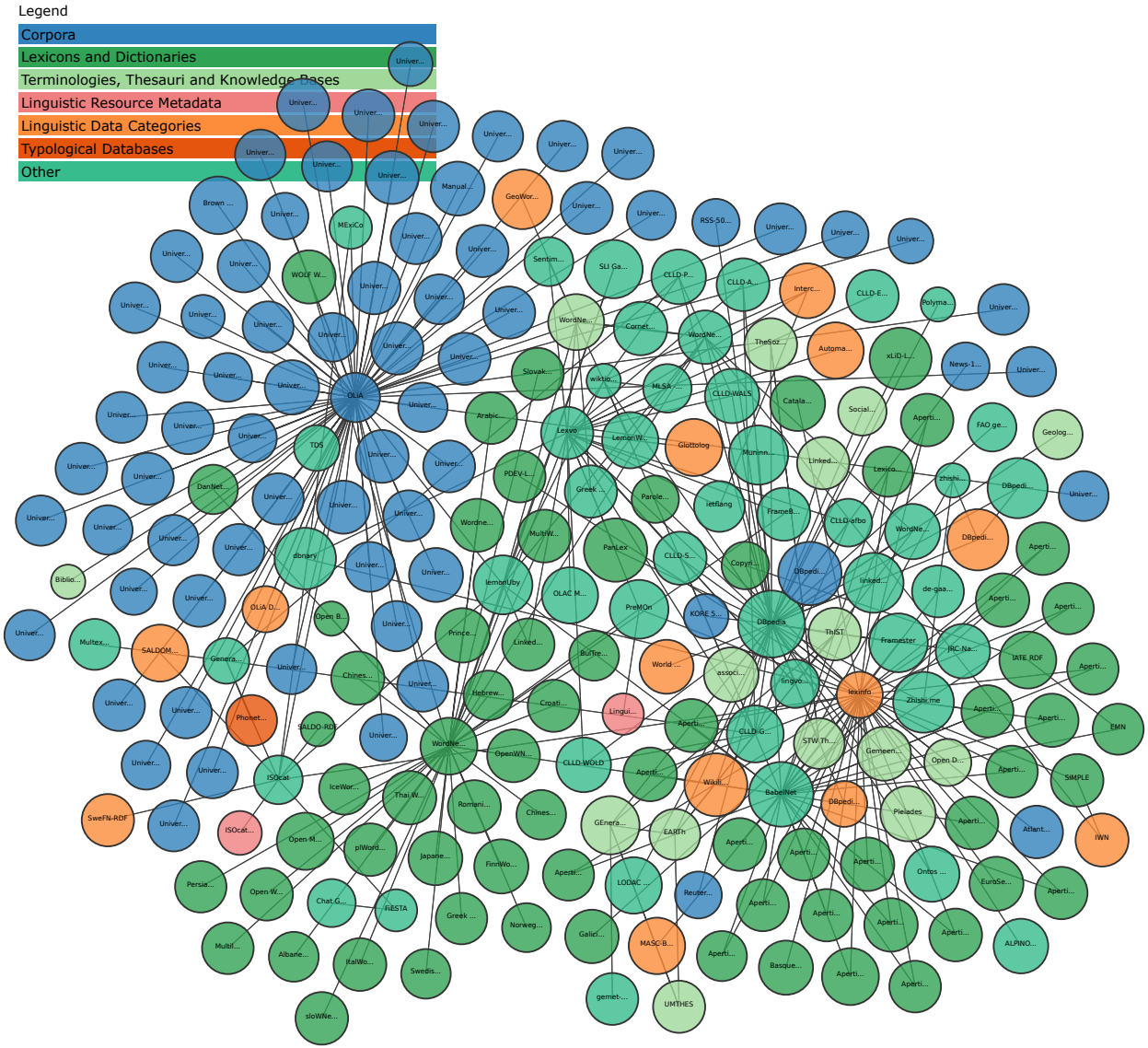
decentralization: as Pelagios only links data from many different project;

a simple model: based on one minimal vocabulary (no effort of conversion/mapping);

community effect: Pelagios is nowadays more than a successful platform; it is a well connected and motivated community of people

de facto standard: Pelagios has achived the critical mass to be a de facto standard.

The L-LOD network



Why LOD?

sparseness: all independent and self-standing projects can live and prosper across the web;

small size/marginality: newcomers can be adequately represented along with the “big players”;

lack of interoperability: as many layers of annotations can be added to enccode information about any level of linguistic analysis (syntax, morphology, semantics, pragmatics. . .);

lack of standardization: the adoption of common vocabularies is crucial for any LOD enterprise.

usability issues: interoperable and standardized data are ready to be reused; data integrated in a LOD network are easier discover and thus reuse.

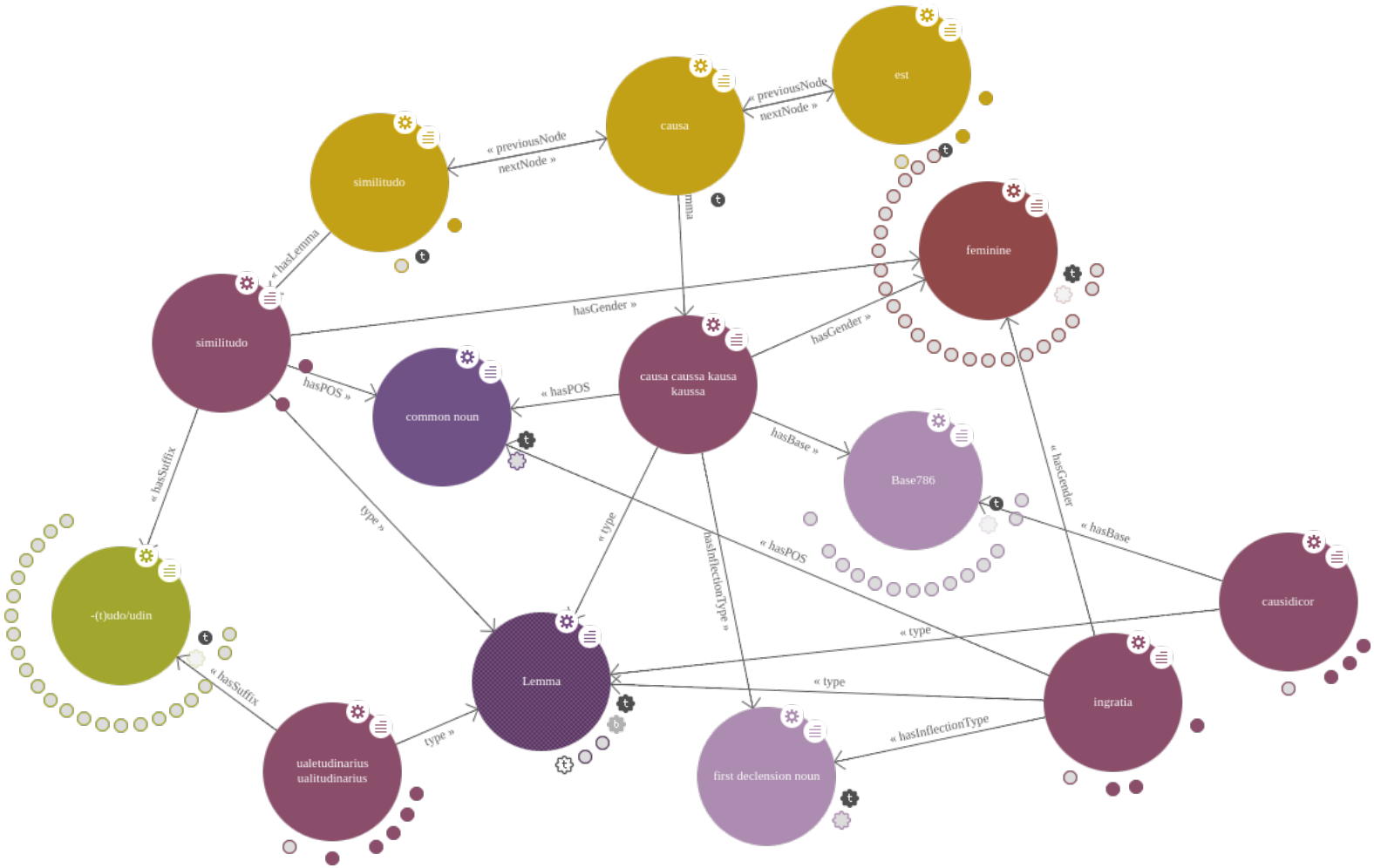
The LiLa project

https://lila-erc.eu/

- funded under the ERC program (principal investigator: Marco Passarotti)
- aims to connect linguistic resources (lexica, corpora, NLP tools) of Latin
- uses the lemma has the linking element (pretty much as Pelagious uses the gazetteer ID)
- provides URIs for latin lemmas, using an ontology based on Ontolex
- the collection of lemmas (and the first resources linked) can be:
  - visualized at: https://lila-erc.eu/lodlive/
  - queried at: https://lila-erc.eu/sparql/

LiLa: link via the lemma

"causa" in Thomas Aq. *SCG* 1.2.1



Summing up

With LOD we can produce data that are:

1. more connected
2. more discoverable
3. more standardized
4. easier to reuse