



 **Nexus**  
Lingarum

4th Summer Datathon on Linguistic Linked Open Data  
Cercedilla, Spain  
30th May – 3June 2022

# LiLa – Linking Latin Introduction

Francesco Mambrini (Università Cattolica del Sacro Cuore, Milan, Italy)





UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore



# LiLa - Linking Latin

An introduction

Francesco Mambrini

4th Summer Datathon on Linguistic Linked Open Data  
Cercedilla, Spain  
30 May 2022



This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme - Grant Agreement No. 769994.

## Introduction and Fundamentals

LiLa: Mission and Architecture

## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Services and Tools

To sum up

## Introduction and Fundamentals

LiLa: Mission and Architecture

## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Services and Tools

To sum up

# Studying Latin (and other ancient languages)

how it was done





We have built and collected (for Latin and other languages):

- ▶ Textual Resources
- ▶ Lexical Resources
- ▶ NLP Tools

We have built and collected (for Latin and other languages):

- ▶ Textual Resources
- ▶ Lexical Resources
- ▶ NLP Tools

Scattered and unconnected



# No “One-Ring-to-Rule-Them-All” Approach

Pelagios



What Pelagios isn't | One ring to rule them all



- Data aggregation X
- Standard data representation X
- MEGA search portal X



Connectivity through *common references* rather than a *common schema*

3

# Sustaining Linked Ancient World Data

From the book [Digital Classical Philology](#)

Hugh A. Cayless

<https://doi.org/10.1515/9783110599572-004>

Cite this

Citations

1

Hugh A. Cayless

## Sustaining Linked Ancient World Data

**Abstract:** May 31st, 2018 marked the sixth anniversary of the Linked Ancient World Data Institute (LAWDI), a workshop funded by the US National Endowment For the Humanities. This makes it a good time to take stock of the Ancient World Linked Data initiatives that have been around for some time, as well as some that have foundered and some that are new. What makes for sustainable Linked Open Data? Why do some initiatives thrive while others fail? What resources do successful LOD sites need, and how may *they* be obtained? The promise of LOD is that it frees our information from the silos in

[Download chapter \(PDF\)](#) 

From the book



**Digital Classical Philology**

Search book



**Chapters in this book (24)**

[Frontmatter](#)

[Editor's Preface](#)

[Preface](#)

[Contents](#)

[Introduction](#)

[The Free First Thousand Years of Greek](#)

[The Digital Latin Library: Cataloging and Publishing Critical Editions of Latin Texts](#)

## ERC Consolidator Grant 2018-2023

A collection of multifarious, interoperable linguistic resources described with the same vocabulary for knowledge description (by using common data categories and ontologies)

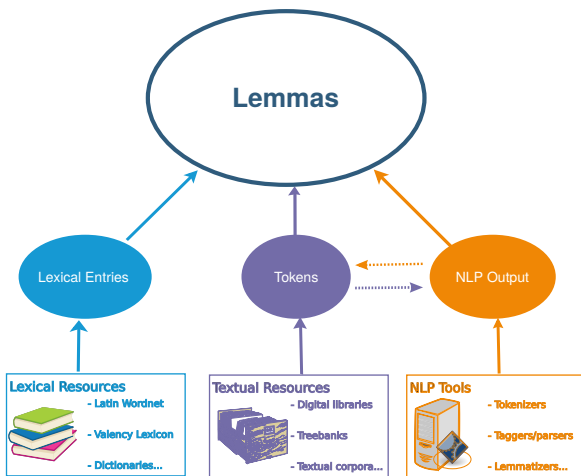
**Interlinking as a Form of Interaction**

# LR: What

“Language resources are collections of data which primarily document communicative acts of humans by some form of recording and/or descriptions, both directly as in corpora, or at higher levels of abstraction in lexicons and ontologies”

(International Standard for Language Engineering, ISLE)

- Pivotal components in NLP
- Machine-readable form
- Data: written, spoken, audio/video
- Tools for acquiring, building, maintaining, sharing, querying LR



## Introduction and Fundamentals

LiLa: Mission and Architecture

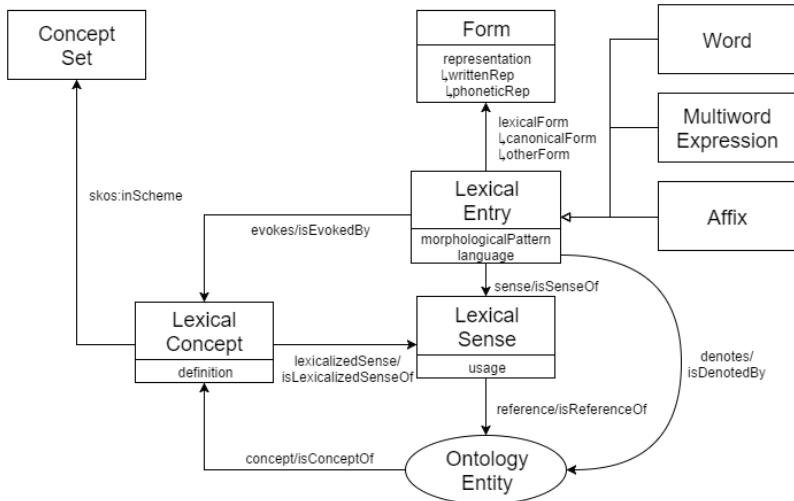
## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Services and Tools

To sum up



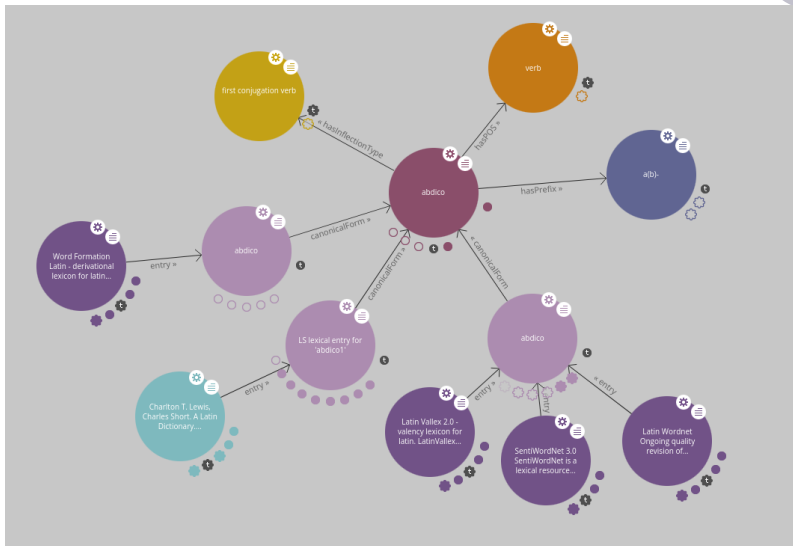
Lemma *abdico* 'to deny, disown'

<http://lila-erc.eu/data/id/lemma/86860>



# Lemmas and Lexical Entries

*abdico*, 'deny, disown'



- ▶ Lemma Bank: OntoLex, custom
- ▶ A Lat.-Eng. dictionary (Lewis & Short): OntoLex + lexicog
- ▶ A derivational lexicon (WFL): Morph
- ▶ A polarity lexicon (LatinAffectus): OntoLex + Marl
- ▶ An etymological dictionary (De Vaan): LemonEty
- ▶ A dictionary of Greek loanwords: LemonEty
- ▶ A Valency Lexicon (Latin Vallex): PreMON
- ▶ A manually checked subset of the Latin WordNet: OntoLex

## Introduction and Fundamentals

LiLa: Mission and Architecture

## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Services and Tools

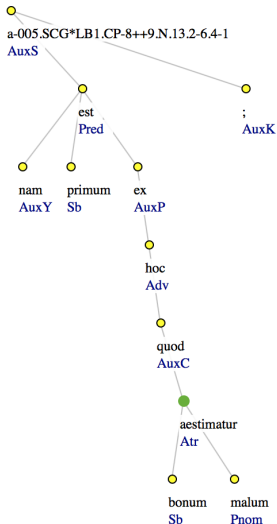
To sum up

# (Annotated) Corpora in LiLa

Source: The *Index Thomisticus* Treebank (original scheme)

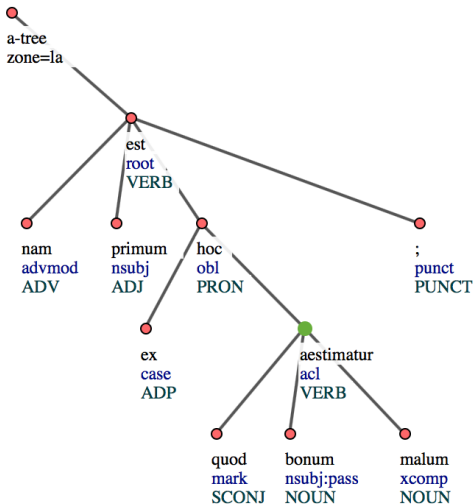
*nam primum est ex hoc  
quod bonum **aestimatur**  
malum;* (IT-TB: SCG, lib. 1,  
cap. 89, n. 13)

*for the first arises because  
the good **is judged** to be  
evil;* (Trans. Anton C. Pegis)



# (Annotated) Corpora in LiLa

Source: The *Index Thomisticus* Treebank (UD scheme)

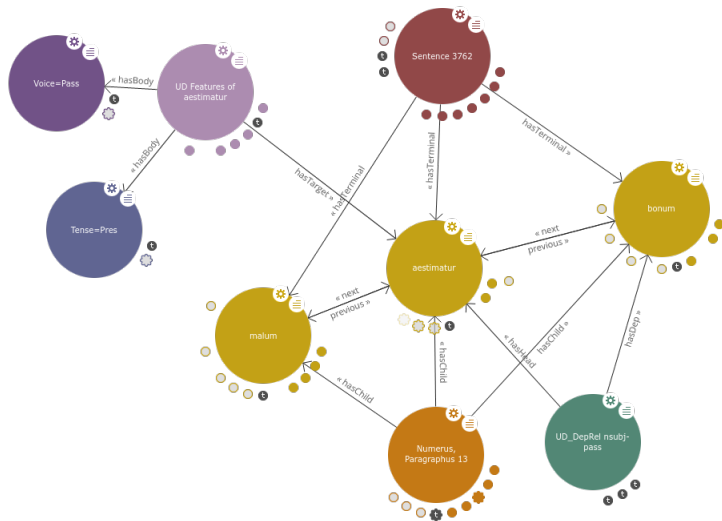


Token *aestimatur*

`http://lila-erc.eu/lodview/data/corpora/  
ITTB/id/token/005.SCG*LB1.CP-8++9.N.13.  
2-6.4-1W8`

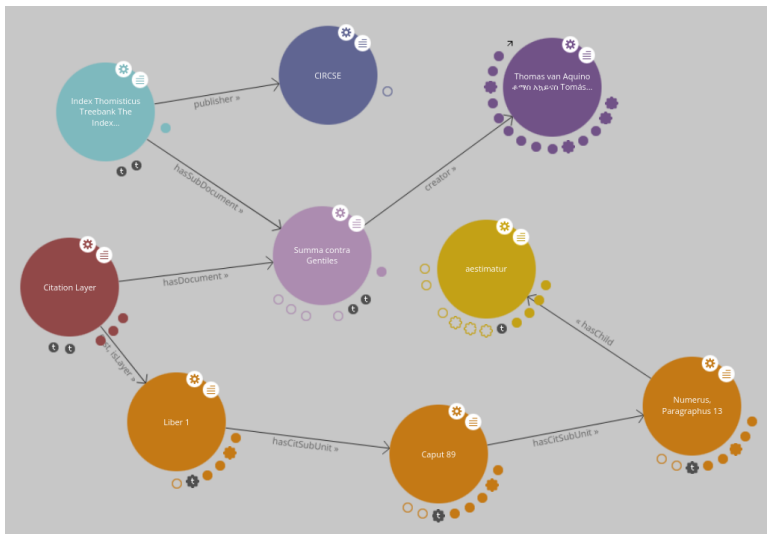
# The token *aestimatur*

Morpho-syntactic information



# The token *aestimatur*

Other metadata





## Introduction and Fundamentals

LiLa: Mission and Architecture

## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Services and Tools

To sum up

## Lemma Bank Query Interface

<https://lila-erc.eu/query/>

## SPARQL Access Point

<https://lila-erc.eu/sparql/>

## TextLinker

<http://lila-erc.eu:8080/LiLaTextLinker/>

## Introduction and Fundamentals

LiLa: Mission and Architecture

## LiLa now!

Lemma Bank and Lexical Resources

Textual Resources

Services and Tools

To sum up

## ▶ Textual Resources

- ✓ Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 400,000 nodes
- ✓ UDante Treebank: ca. 46,000 tokens
- ✓ *Querolus sive Aulularia*: ca. 17,000 tokens
- ✓ *Liber Abbaci* (ch. VIII) by Leonardo Fibonacci: ca. 30,000 tokens
- ✓ LASLA Corpus: ca. 1.7 million tokens
- PROIEL and LLCT treebanks, CompHistSem, CroALa, Musisque DeoQue

## ▶ Lexical Resources

- ✓ Lemma Bank: ca. 200,000 canonical forms
- ✓ Word Formation Latin: ca. 36,000 lemmas (Classical Latin)
- ✓ Etymological Dict. of Latin & the Other Italic Langs.: ca. 1,500 entries
- ✓ LatinAffectus: ca. 3,300 entries
- ✓ Index Graecorum Vocabulorum in L. Latinam Transl.: ca. 1,800 entries
- ✓ Latin WordNet: ca. 2,500 manually checked entries
- ✓ Latin Vallex 2.0: ca. 2,000 entries
- ✓ Lewis & Short Dictionary: ca. 50,000 entries
- Lexikon der Indogermanischen Verben (LIV). Wiktionary, BabelNet

**TOTAL: approximately 33 million triples**

# Keep in touch!!

Thank you



## LiLa: Linking Latin

Università Cattolica del Sacro Cuore  
CIRCSE Research Centre



[info@lila-erc.eu](mailto:info@lila-erc.eu)



<https://github.com/CIRCSE>



<https://lila-erc.eu>



[@ERC\\_LiLa](https://twitter.com/ERC_LiLa)



Largo Gemelli 1, 20123 Milan, Italy



This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme - Grant Agreement No. 769994.