Lemmi, alberi e classi

Dall'*Index Thomisticus* ai Linked Data

Marco Passarotti

Digital Spritz - Università di Verona
24 Novembre 2021

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.

Overview

The Story So Far (through the lenses of the *Index Thomisticus*)

- Analogical and Isolated …but Findable
- Digital and (Partly) Accessible

Prefixes Matter. Infrastructures and Interoperability

- CLARIN. One place fits all
- Linguistic Linked Open Data. One place interlinks all
- It Works! The LiLa Knowledge Base

Conclusions …and Hopes for the Future

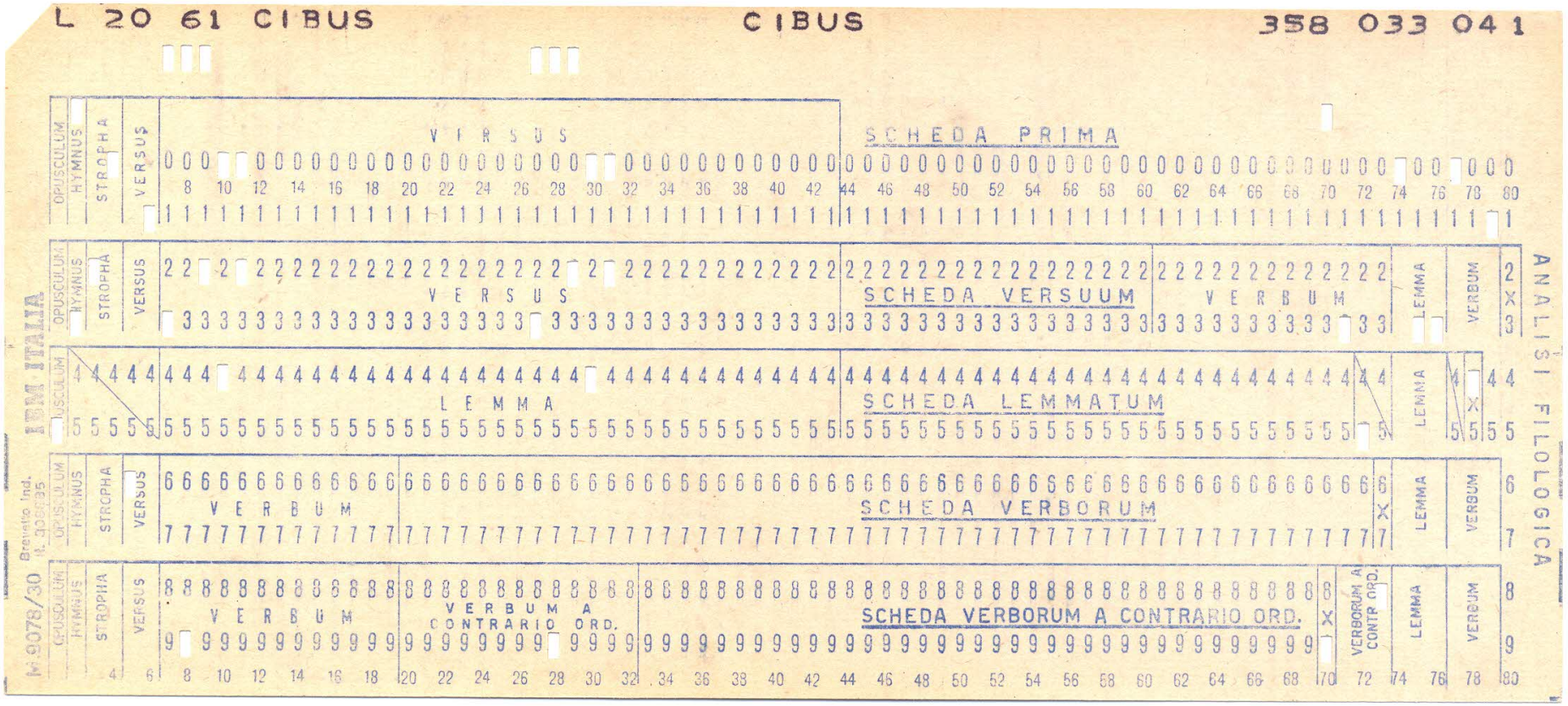
Table of Contents

Millions of Punched Cards

Analogical and isolated …but Findable

Millions of Punched Cards

Analogical and isolated …but Findable



Millions of Punched Cards

Analogical and isolated …but Findable

The Result ...on Paper

The *Index Thomisticus*

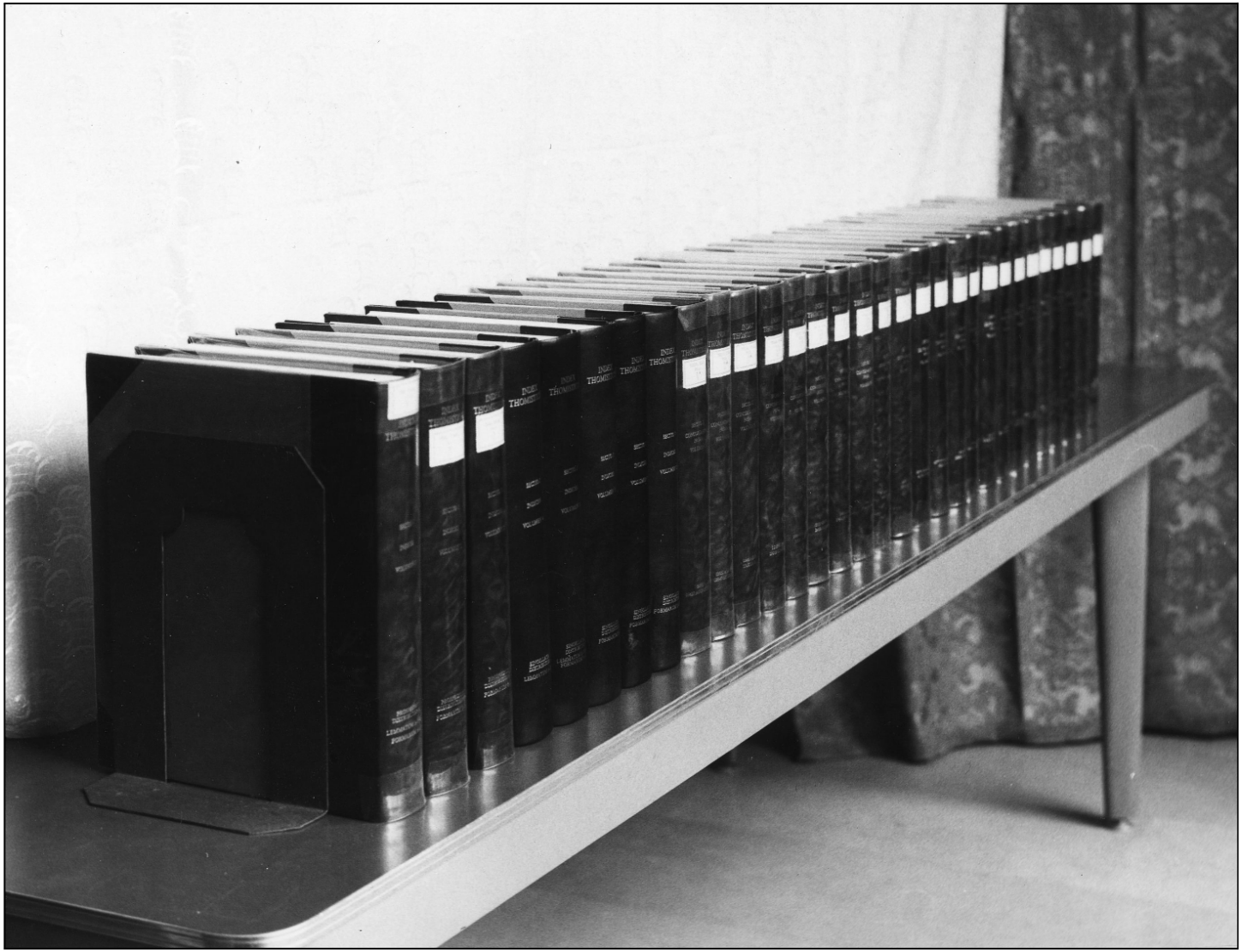
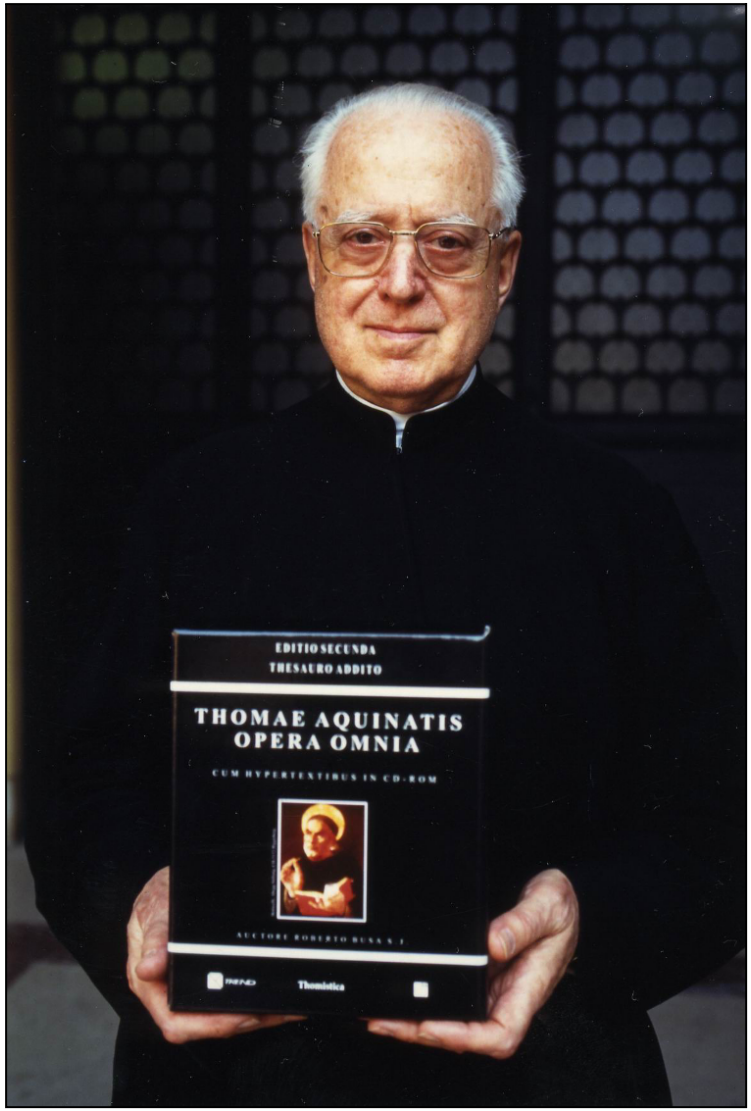


Table of Contents

The Digital Turn: (meta)data (partly) accessible

The CD-ROM of the *Index Thomisticus*



The Digital Turn: (meta)data (partly) accessible

Thanks God for the CD-ROM!

The Digital Turn: (meta)data (partly) accessible

Thanks God for Internet!

CORPUS THOMISTICUM

INDEX THOMISTICUS

by Roberto Busa SJ and associates

web edition by Eduardo Bernot and Enrique Alarcón

la versione italiana non è ancora disponibile

Search: computare

concordances | terms | works | options | new search

FOUND 13 CASES IN 13 PLACES

1-10

CASE 1. PLACE 1. Super Sent., lib. 2 d. 14 q. 1 a. 4 arg. 4. [...]$^{-1}$ Sed ratione hujus convenientiae aer et ignis caelum dicuntur. Ergo videtur quod oportuisset similiter aqueum elementum inter caelos **computare.**

CASE 2. PLACE 2. Super Sent., lib. 4 d. 30 q. 2 a. 1 qc. 2 co. Ad secundam quaestionem dicendum, quod conveniens fuit matrem Christi matrimonio esse junctam tum propter causas in littera assignatas, tum etiam propter alias causas: quarum prima est, ut significaret Ecclesiam, quae est virgo et sponsa. Secunda, ut per Joseph genealogia Mariae texeretur: non enim erat consuetudo apud Hebraeos ex parte mulierum genealogiam **computare.** Tertia, ut virginibus excusatio tolleretur, si de fornicatione infamantur. [...]$^{-2}$

CASE 3. PLACE 3. Super Sent., lib. 4 d. 43 q. 1 a. 3 qc. 2 co. [...]$^{-7}$ Unde illi omnes qui tempus praedictum numerare voluerunt, hactenus falsiloqui sunt inventi. Quidam enim, ut Augustinus dixit ibidem, dixerunt ab ascensione domini usque ad ultimum ejus adventum quadringentos annos posse compleri, alii quingentos, alii mille: quorum falsitas patet; et similiter patebit eorum qui adhuc **computare** non cessant.

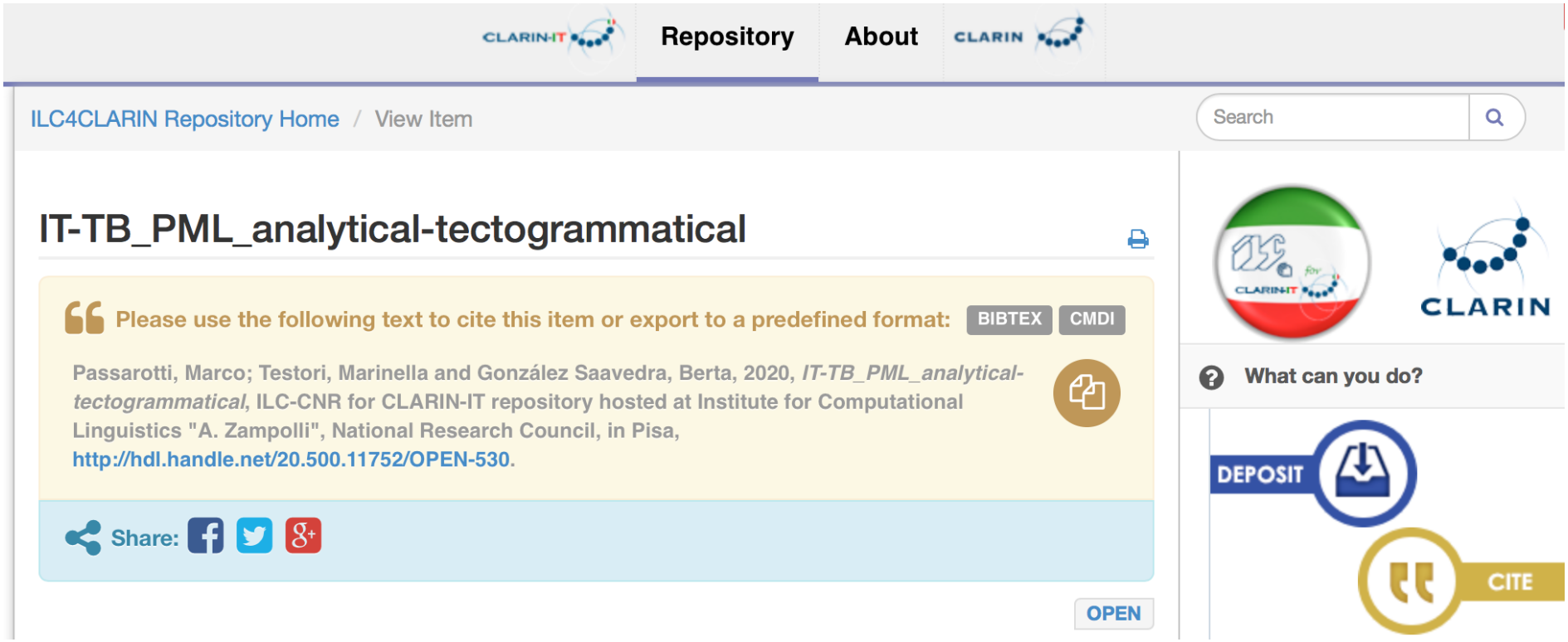
Table of Contents

The CLARIN Infrastructure

One place fits all: making resources Accessible, Reusable and (partly) Findable

The CLARIN Infrastructure

One place fits all: making Resources Accessible, Reusable and (partly) Findable



The CLARIN Infrastructure

One place fits all: but what about making resources (semantically) Interoperable?

The CLARIN Infrastructure

One place fits all: but what about making resources (semantically) Interoperable?

- **Virtual Language Observatory**: aggregating metadata from various sources to find a set of resources

The CLARIN Infrastructure

One place fits all: but what about making resources (semantically) Interoperable?

- **Virtual Language Observatory**: aggregating metadata from various sources to find a set of resources
- **Language Resource Switchboard** (+ **WebLicht**, to combine the outputs of different NLP tools into custom processing chains): web-based NLP services grouped by task/function

The CLARIN Infrastructure

One place fits all: but what about making resources (semantically) Interoperable?

- **Virtual Language Observatory**: aggregating metadata from various sources to find a set of resources
- **Language Resource Switchboard** (+ **WebLicht**, to combine the outputs of different NLP tools into custom processing chains): web-based NLP services grouped by task/function
- **Component MetaData Infrastructure** (CMDI):
  - Components: groups of semantically coherent metadata elements
  - Profiles: sets of components describing specific resource types
  - Component Registry: a collection of concepts with persistent identifiers

The CLARIN Infrastructure

One place fits all: but what about making resources (semantically) Interoperable?

- **Virtual Language Observatory**: aggregating metadata from various sources to find a set of resources
- **Language Resource Switchboard** (+ **WebLicht**, to combine the outputs of different NLP tools into custom processing chains): web-based NLP services grouped by task/function
- **Component MetaData Infrastructure** (CMDI):
  - Components: groups of semantically coherent metadata elements
  - Profiles: sets of components describing specific resource types
  - Component Registry: a collection of concepts with persistent identifiers
- **Semantic interoperability**: linking metadata components to concepts stored in the CLARIN Concept Registry (CCR)

The CLARIN Infrastructure

Limitations to interoperability in CLARIN (P. Labropoulou: LLD CLARIN Café, 29/4/21)

The CLARIN Infrastructure

Limitations to interoperability in CLARIN (P. Labropoulou: LLD CLARIN Café, 29/4/21)

- CCR: a collection of concepts, identifiable by their persistent identifiers, relevant for the domain of language resources
  - no relations between internal concepts (hence, no semantic inference)
  - no relations to external concepts (no crosswalks to popular schemas)
  - lack of curation (too many similar concepts)

Table of Contents

The Linked Data Principles

…just to be FAIR

The Linked Data Principles
...just to be FAIR

- Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)

The Linked Data Principles

…just to be FAIR

- Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)
- Use HTTP URIs to allow people (and machines) to look up things

The Linked Data Principles

…just to be FAIR

- Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)
- Use HTTP URIs to allow people (and machines) to look up things
- Use web standards to represent/query (meta)data, such as RDF and SPARQL

The Linked Data Principles

…just to be FAIR

- Use URIs for things (e.g. an entry in a lexicon, a token in a corpus)
- Use HTTP URIs to allow people (and machines) to look up things
- Use web standards to represent/query (meta)data, such as RDF and SPARQL
- Include links to other URIs

Why To Apply LD to Linguistic Resources

J. Gracia: LLD CLARIN Café, 29/4/21

Why To Apply LD to Linguistic Resources

J. Gracia: LLD CLARIN Café, 29/4/21

- Resources disconnected from each other (silos of LRs)

Why To Apply LD to Linguistic Resources

J. Gracia: LLD CLARIN Café, 29/4/21

- Resources disconnected from each other (silos of LRs)
- Proprietary and heterogeneous formats

Why To Apply LD to Linguistic Resources

J. Gracia: LLD CLARIN Café, 29/4/21

- Resources disconnected from each other (silos of LRs)
- Proprietary and heterogeneous formats
- Different representation schemes, query languages, annotation criteria and tagsets

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

- Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

- Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- Structural Interoperability: common data model consisting of shared protocols and data formats (HTTP, URIs, RDF, SPARQL)

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

- Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- Structural Interoperability: common data model consisting of shared protocols and data formats (HTTP, URIs, RDF, SPARQL)
- Conceptual Interoperability: common ontologies and re-usable vocabularies to understand how to use the URIs: resources are explicitly linked

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

- Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- Structural Interoperability: common data model consisting of shared protocols and data formats (HTTP, URIs, RDF, SPARQL)
- Conceptual Interoperability: common ontologies and re-usable vocabularies to understand how to use the URIs: resources are explicitly linked
- Federation: to combine information from separated repositories

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

- Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- Structural Interoperability: common data model consisting of shared protocols and data formats (HTTP, URIs, RDF, SPARQL)
- Conceptual Interoperability: common ontologies and re-usable vocabularies to understand how to use the URIs: resources are explicitly linked
- Federation: to combine information from separated repositories
- Dynamicity: to provide access to the most recent version of a resource

Benefits of Applying LD to Linguistic Resources

Chiarcos et al. (2013)

- Representation and Modelling: RDF is a very versatile data model to represent stand-off annotations, dependency parses etc.
- Structural Interoperability: common data model consisting of shared protocols and data formats (HTTP, URIs, RDF, SPARQL)
- Conceptual Interoperability: common ontologies and re-usable vocabularies to understand how to use the URIs: resources are explicitly linked
- Federation: to combine information from separated repositories
- Dynamicity: to provide access to the most recent version of a resource
- Ecosystem: a large and active community with common tools and practices. Initiatives: (1) COST Action *Nexus Linguarum* (COST Action 2019-2023): European network for Web-centred linguistic data science; (2) *Prêt-à-LLOD* (RIA 2019-2022): Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors; (3) LD4LT (*Linked Data for Language Technology Community Group*): to create a consolidated LOD vocabulary for web (linguistic) annotation

Table of Contents

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

CLARIN
Common Language Resources and Technology Infrastructure

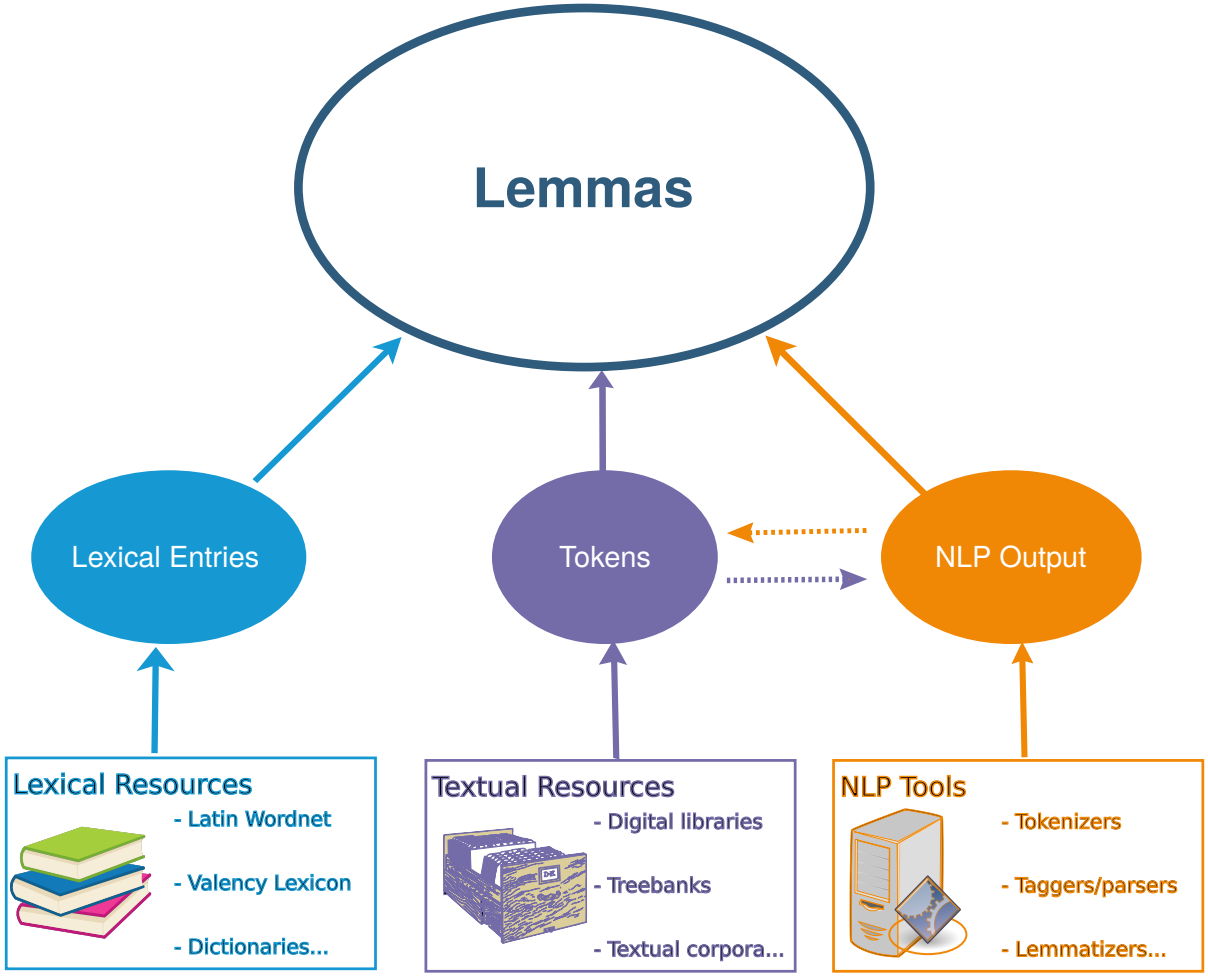
Infrastructure

LiLa
Linking Latin

Interoperability

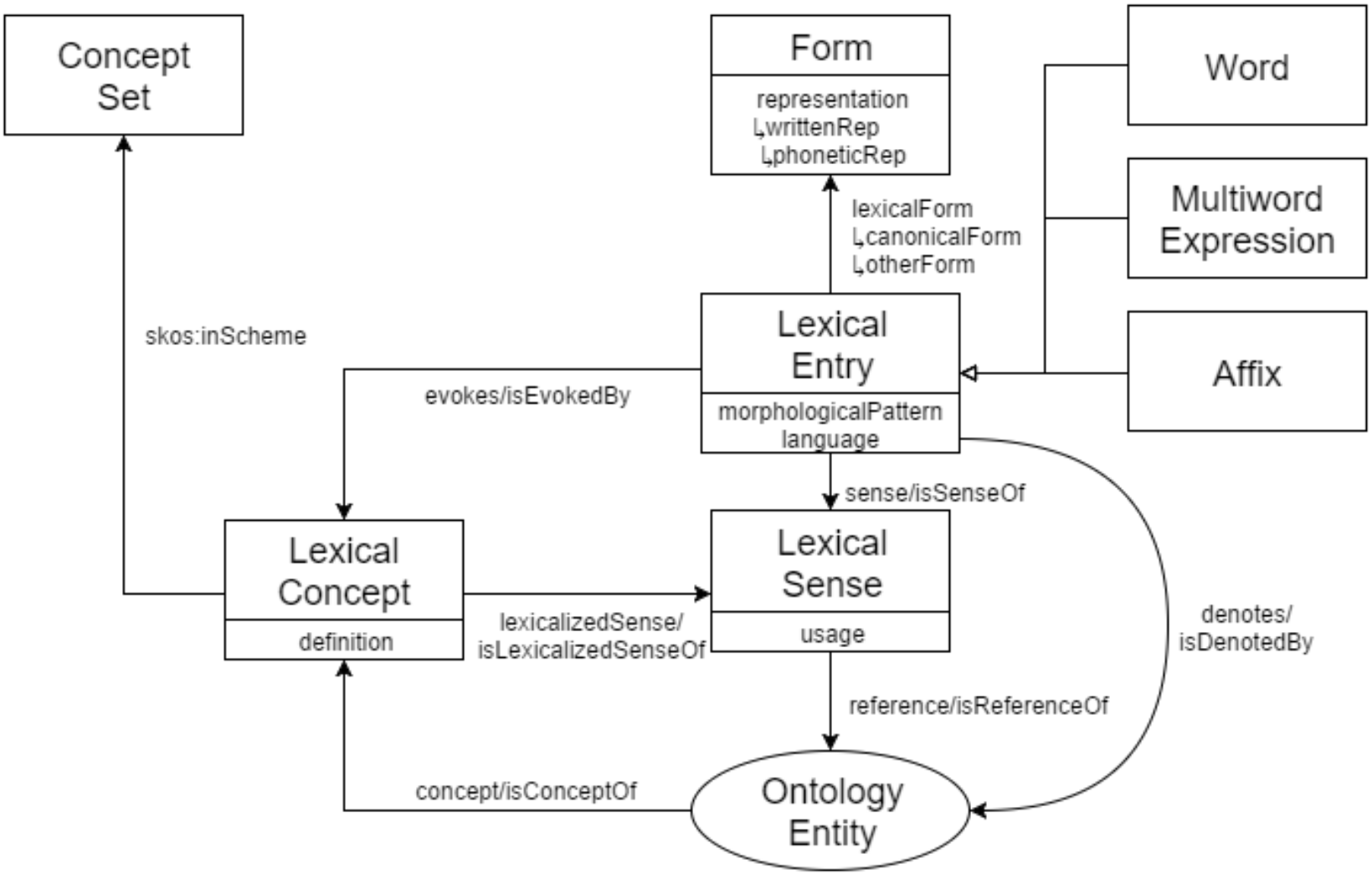
LiLa Knowledge Base

Lexically-based architecture and (meta)data sources



LiLa and Ontolex Lemon

A *de facto* W3C standard for publishing lexical data as LLOD



Lemma *admiror* 'to admire, to respect'

`http://lila-erc.eu/data/id/lemma/87541`

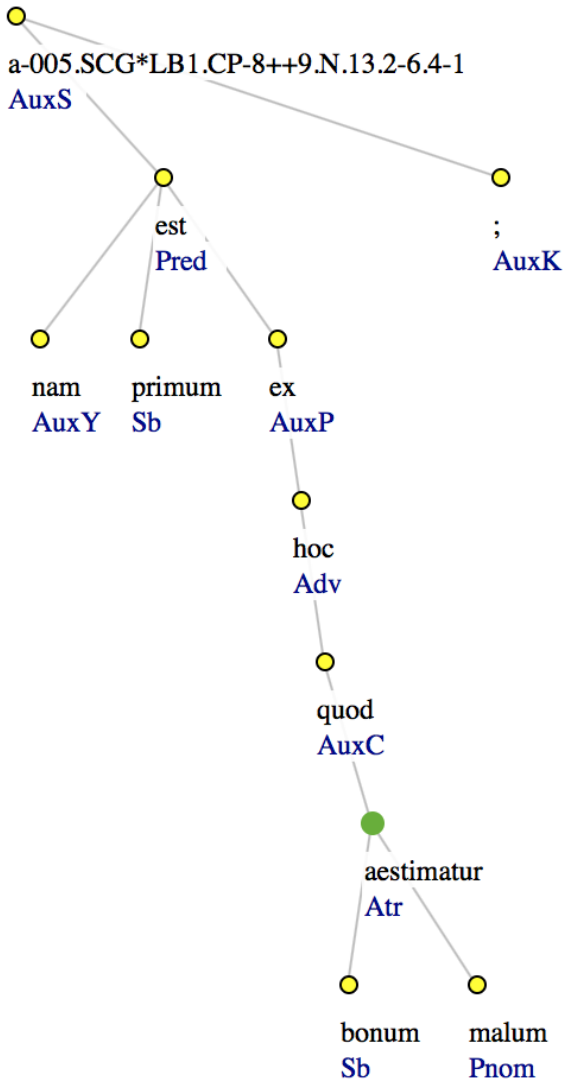
- Lemma Bank
- A bilingual dictionary (Lewis & Short)
- A derivational lexicon (Word Formation Latin)
- A polarity lexicon (LatinAffectus)
- An etymological dictionary (De Vaan)
- A Valency Lexicon (Latin Vallex)
- A manually checked subset of the Latin WordNet

Textual Resources

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

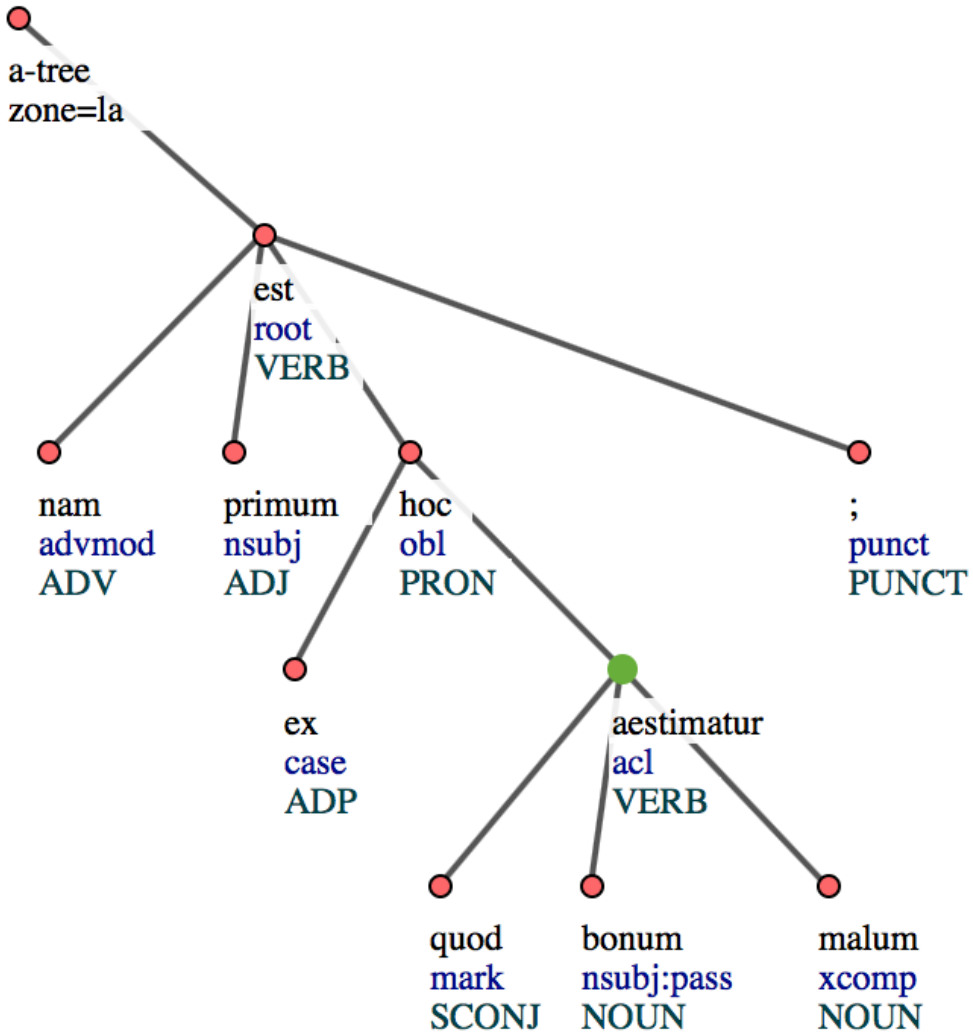
nam primum est ex hoc quod bonum ***aestimatur*** *malum*; (IT-TB: SCG, lib. 1, cap. 89, n. 13)

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



Textual Resources

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



Textual Resources

A token from the *Index Thomisticus* Treebank in LiLa

Token *aestimatur*

http://lila-erc.eu/lodview/data/corpora/ITTB/id/token/005.SCG*LB1.CP-8++9.N.13.2-6.4-1W8

LiLa: Overview

Resources connected and upcoming connections

- **Corpora**
  - [x] Index Thomisticus Treebank (*Summa contra Gentiles*): ca. 450,000 nodes
  - [x] Dante Search (700th death anniversary): ca. 46,000 tokens
  - [x] *Querolus sive Aulularia*: ca. 17,000 tokens
  - [ ] PROIEL and LLCT treebanks
  - [ ] Computational Historical Semantics, LASLA and CroALa Corpora
- **Lexica**
  - [x] Word Formation Latin: ca. 46,000 lemmas (Classical Latin)
  - [x] Etymological dictionary of Latin & the other Italic Langs.: ca. 1,400 entries
  - [x] LatinAffectus: ca. 2,300 entries
  - [x] Index Graecorum Vocabulorum in Linguam Latinam: ca. 1,800 entries
  - [x] Latin WordNet: ca. 1,000 manually checked entries
  - [x] Latin Vallex 2.0: Valency Lexicon
  - [x] Lewis & Short Dictionary
- **NLP tools**
  - [x] LEMLAT (lemma bank): ca. 150,000 lemmas
- **TOTAL: approximately 16 million triples** + 10 further millions

Query Interface, Triplestore and Linker

- Query interface; Triplestore
- Linker

Linguistic Resources. Corpora

- Index Thomisticus Treebank
- Dante Search
- *Querolus sive Aulularia*

Linguistic Resources. Lexica

- Word Formation Latin
- Etymological Dictionary of Latin & the Other Italic Languages
- LatinAffectus + Latin WordNet
- Index Graecorum Vocabulorum in Linguam Latinam
- Latin Vallex 2.0
- Lewis & Short Latin-English dictionary

Open Challenges

It is not so easy ...but we are doing our best

To fully exploit digital texts

texts "in a form that both humans and machines can use,
preferably in a way that leverages the unique capabilities of both"
(S. Huskey, SunoikisisDC SS 2021 - Session 12)

Open Challenges

It is not so easy …but we are doing our best

To fully exploit digital texts

texts "in a form that both humans and machines can use,
preferably in a way that leverages the unique capabilities of both"
(S. Huskey, SunoikisisDC SS 2021 - Session 12)

- Making LOD more accessible and usable: facilitating wider participation in LOD, e.g. by automating the processing of (meta)data (workflows for creating LOD)

Open Challenges

It is not so easy …but we are doing our best

To fully exploit digital texts

texts "in a form that both humans and machines can use,
preferably in a way that leverages the unique capabilities of both"
(S. Huskey, SunoikisisDC SS 2021 - Session 12)

- Making LOD more accessible and usable: facilitating wider participation in LOD, e.g. by automating the processing of (meta)data (workflows for creating LOD)
- LOD is Open and Accessible, but producing it takes money, time, expertise: funding for data entry, modeling etc.

Open Challenges

It is not so easy ...but we are doing our best

To fully exploit digital texts

texts "in a form that both humans and machines can use,
preferably in a way that leverages the unique capabilities of both"
(S. Huskey, SunoikisisDC SS 2021 - Session 12)

- Making LOD more accessible and usable: facilitating wider participation in LOD, e.g. by automating the processing of (meta)data (workflows for creating LOD)
- LOD is Open and Accessible, but producing it takes money, time, expertise: funding for data entry, modeling etc.
- Models still missing for several types of (meta)data: e.g. for critical editions

Open Challenges

It is not so easy …but we are doing our best

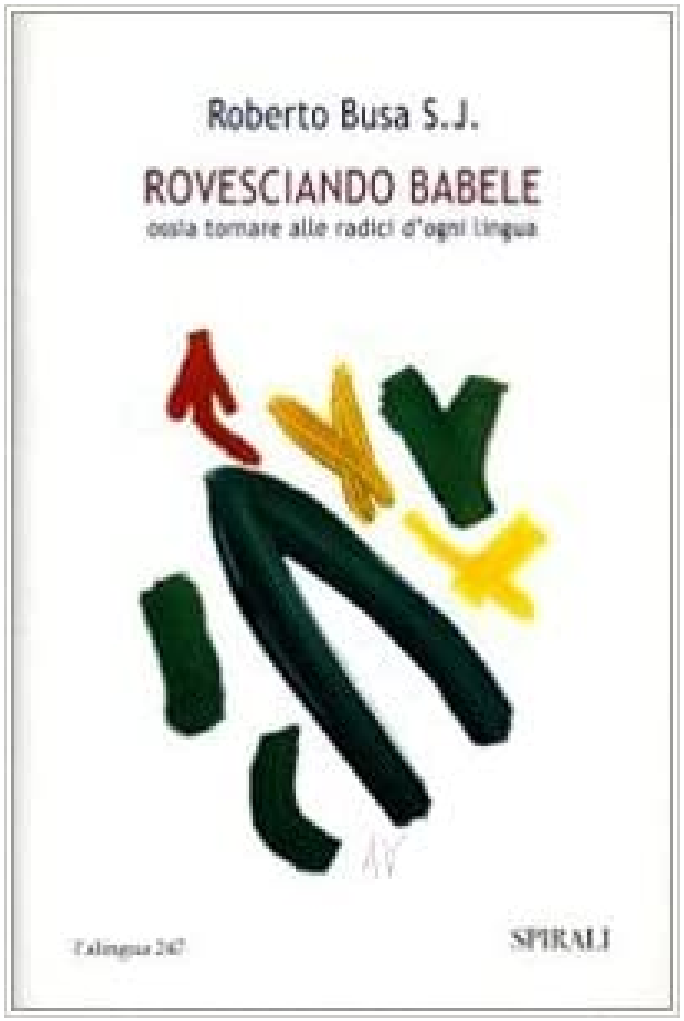
To fully exploit digital texts

texts "in a form that both humans and machines can use,
preferably in a way that leverages the unique capabilities of both"
(S. Huskey, SunoikisisDC SS 2021 - Session 12)

- Making LOD more accessible and usable: facilitating wider participation in LOD, e.g. by automating the processing of (meta)data (workflows for creating LOD)
- LOD is Open and Accessible, but producing it takes money, time, expertise: funding for data entry, modeling etc.
- Models still missing for several types of (meta)data: e.g. for critical editions
- Community-based effort: persuading resource developers to adopt LOD practices and reaching consensus around shared vocabularies, ontologies, data categories etc.

A Shared, Formal Representation of What Exists

A common *language* to turn Babel upside down



Thanks!
Get in touch

LiLa: Linking Latin
Università Cattolica del Sacro Cuore
CIRCSE Research Centre

info@lila-erc.eu

https://github.com/CIRCSE

https://lila-erc.eu

@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.