# Developing a Machine-Readable Dataset and Parsing Tools for Analyzing Apostolic Penitentiary Texts

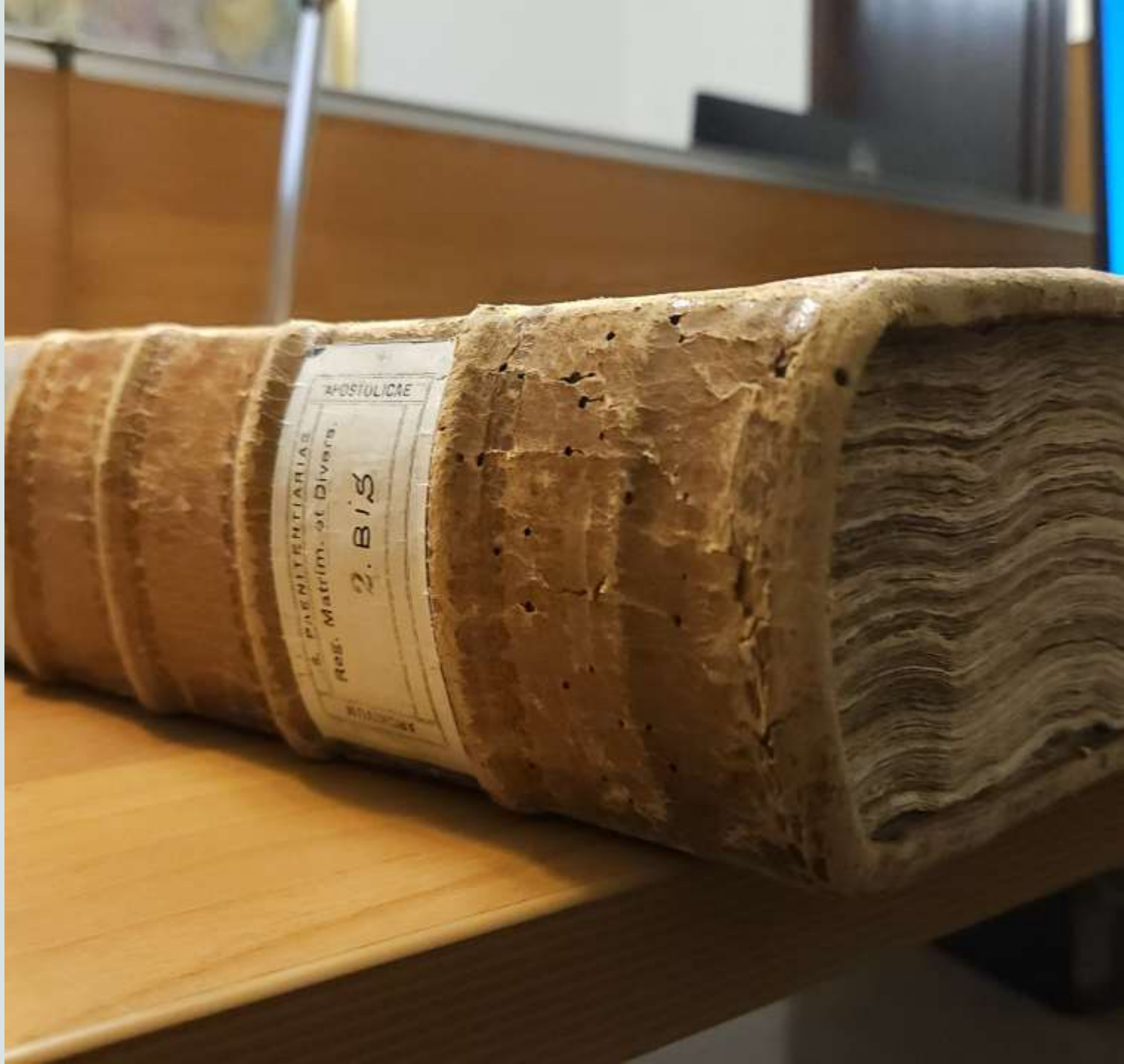
Hanna-Mari Kupari,

University of Turku

Circolo Gianicolense

American Academy in

Rome, 18th of February




# Contents:



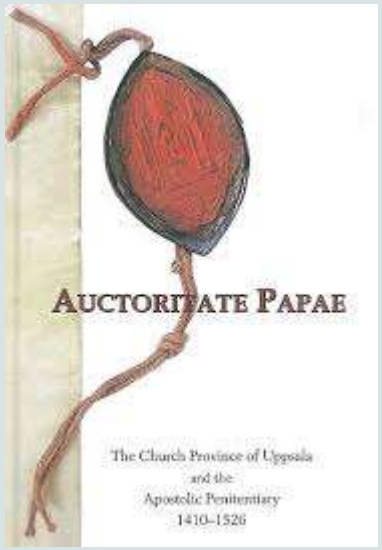
Slides as pdf to download

# Introduction

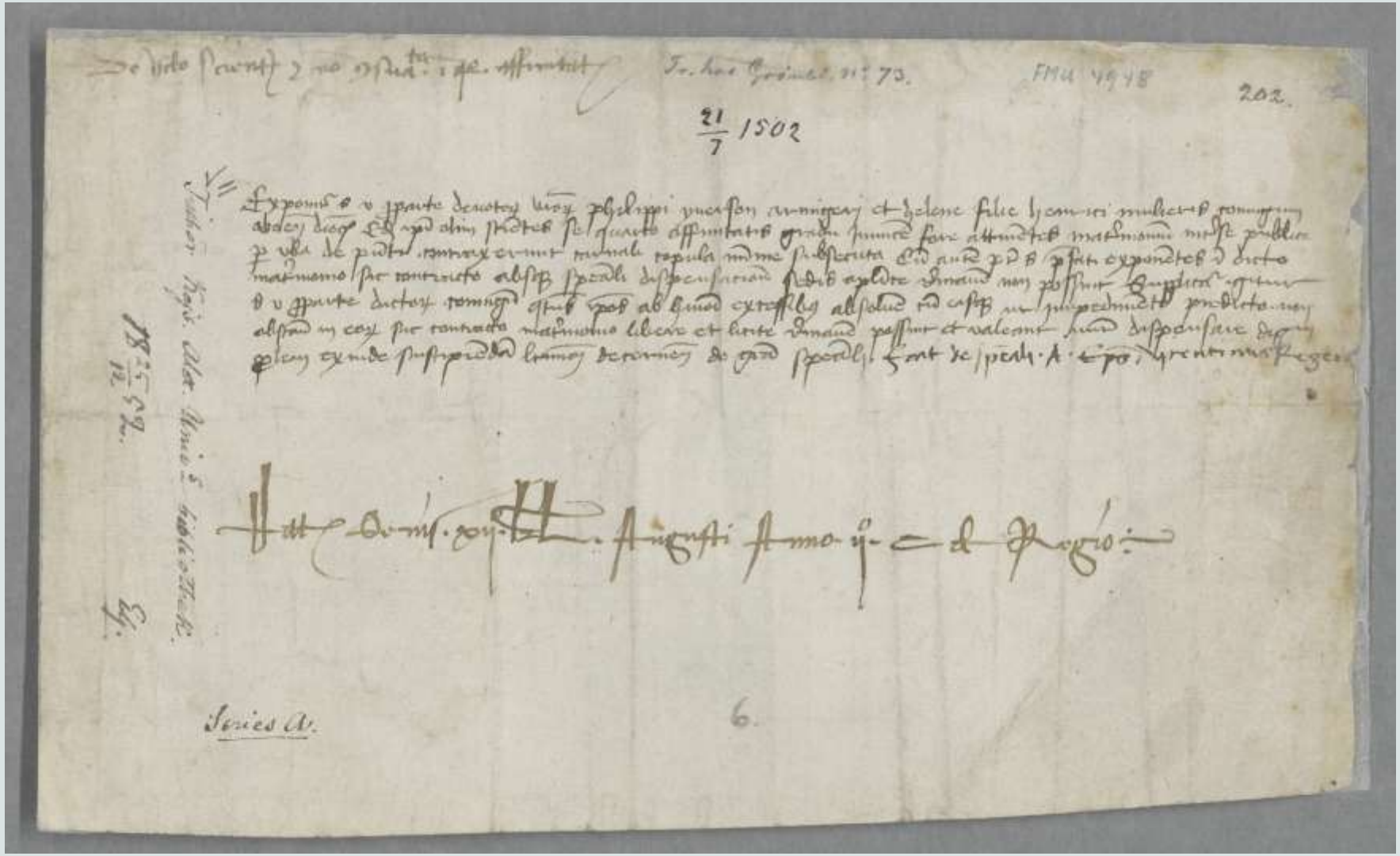
THE PENITENTIARY DOCUMENTS
OF THE VATICAN

# The Apostolic Penitentiary

- Cases of diverse petitions from the Apostolic See of the Vatican
- Late Middle Ages, c. 1390-1530
- A “tribunal of mercy”, responsible for issues relating to the forgiveness of sins in the Church
- Copybooks survive and edited according to modern principles for historians using quantitative methods

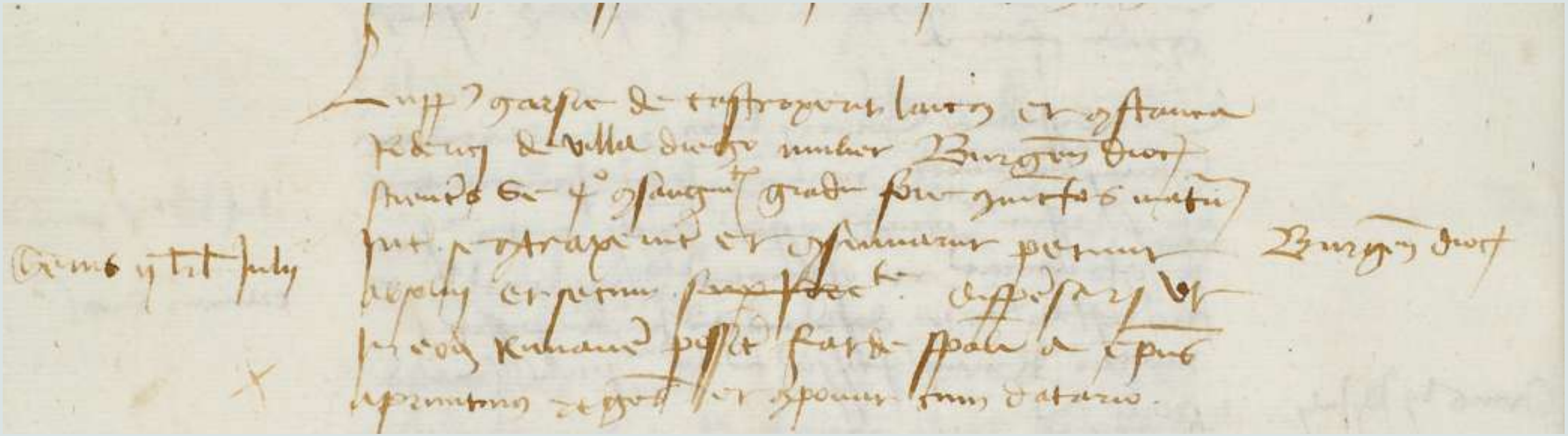


# An original petition from a Finnish local archive



https://df.narc.fi/document/4948#picture

# A register copy of the same document



APA Reg Matrim. et Divers. 8, 63 verso

# Background & Aims

DIGITAL HUMANITIES METHODS

# Background

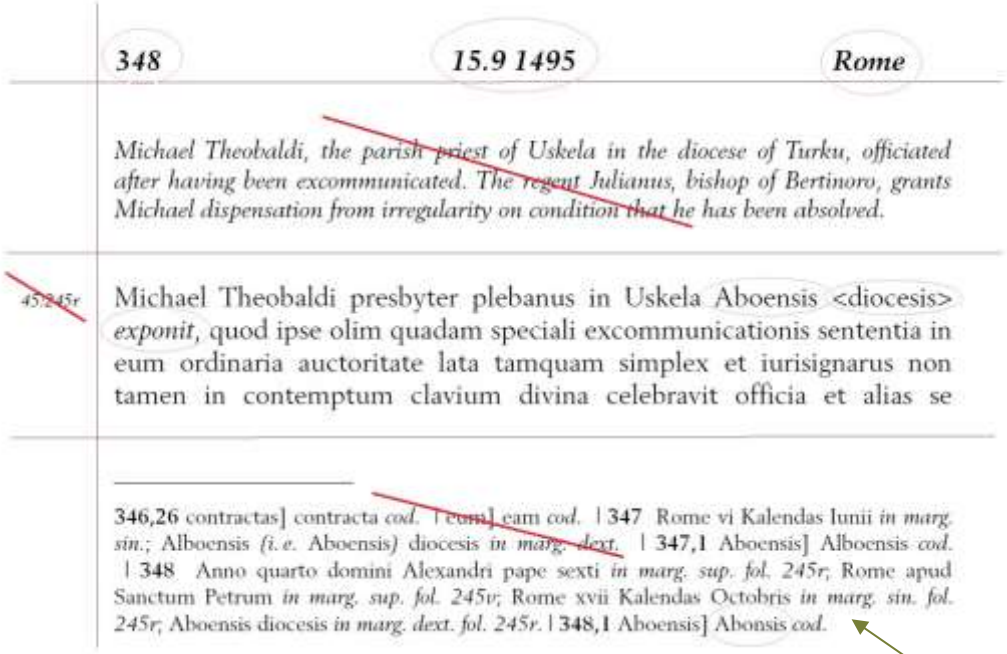
- The need to have more diverse sources:
  - Many of the current sources are:
  - Official documents or literary texts
  - Penitentiary documents offer a window into all social strata
- To understand history
- Better understanding of language use

# Aims

- Several digital resources:
  - A database in TEI XML
  - Three new parsers
  - A sample treebank of the PeDoCo
- Exploring variation in language use by computational methods

# TEI XML structure

A DATABASE



**348** **15.9 1495** **Rome**

*Michael Theobaldi, the parish priest of Uskela in the diocese of Turku, officiated after having been excommunicated. The regent Julianus, bishop of Bertinoro, grants Michael dispensation from irregularity on condition that he has been absolved.*

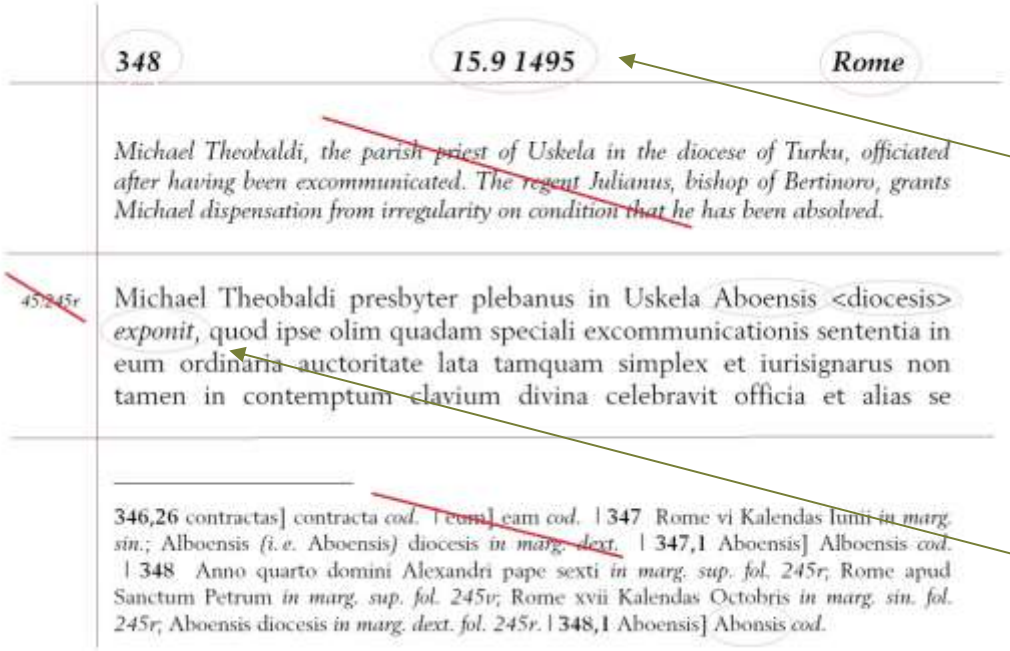
45:245r Michael Theobaldi presbyter plebanus in Uskela Aboensis <diocesis> *exponit*, quod ipse olim quadam speciali excommunicationis sententia in eum ordinaria auctoritate lata tamquam simplex et iurisignarus non tamen in contemptum clavium divina celebravit officia et alias se

**346,26** contractas] contracta *cod.* | eum] eam *cod.* | **347** Rome vi Kalendas Iunii *in marg. sin.*; Alboensis *(i.e.* Aboensis*)* diocesis *in marg. dext.* | **347,1** Aboensis] Alboensis *cod.* | **348** Anno quarto domini Alexandri pape sexti *in marg. sup. fol. 245v*; Rome apud Sanctum Petrum *in marg. sup. fol. 245v*; Rome xvii Kalendas Octobris *in marg. sin. fol. 245r*; Aboensis diocesis *in marg. dext. fol. 245r.* | **348,1** Aboensis] Abonsis *cod.*

Screenshot from *Auctoritate Papae* by Risberg and Salonen (2008)

# The XML framework

- A human and machine-readable text-based annotation to differentiate between different parts of text
- A basic structure of <tag> text </tag>
- Tags can contain information on words, editorial choices and background information
- Removing from the database linguistically irrelevant data



**348** **15.9 1495** ***Rome***

*Michael Theobaldi, the parish priest of Uskela in the diocese of Turku, officiated after having been excommunicated. The regent Julianus, bishop of Bertinoro, grants Michael dispensation from irregularity on condition that he has been absolved.*

45/245r Michael Theobaldi presbyter plebanus in Uskela Aboensis <diocesis> *exponit*, quod ipse olim quadam speciali excommunicationis sententia in eum ordinaria auctoritate lata tamquam simplex et iurisignarus non tamen in contemptum clavium divina celebravit officia et alias se

**346,26** contractas] contracta *cod.* | eam] eam *cod.* | **347** Rome vi Kalendas Iunii *in marg. sin.*; Alboensis *(i.e.* Aboensis*)* diocesis *in marg. dext.* | **347,1** Aboensis] Alboensis *cod.* | **348** Anno quarto domini Alexandri pape sexti *in marg. sup. fol. 245r*; Rome apud Sanctum Petrum *in marg. sup. fol. 245v*; Rome xvii Kalendas Octobris *in marg. sin. fol. 245r*; Aboensis diocesis *in marg. dext. fol. 245r.* | **348,1** Aboensis] Abonsis *cod.*

# Example

```
<text n="300" source="AP" onum="348" type="diversis" bundle="n" several_witnesses="n"><front>

<docDate><date when="1495-09-15"/></docDate>
<placeName type="place-issue">Rome</placeName>

</front><body><p>. . . <choice><corr>Aboensis</corr>

<sic>Abonsis</sic></choice> <supplied reason="omitted-unintentional">diocesis</supplied>

<supplied reason="omitted-intentional">exponit</supplied>. . . </p>

</body></text>
```

# TEI XML

- Since the formulation of different tags in free
- A framework has been set up to propose some shared guidelines
- Text Encoding Initiative for diverse types of data
- The TEI Consortium is a nonprofit membership organization
- https://tei-c.org/
- The editions of three editorial regions of the PeDoCo are structured in this format

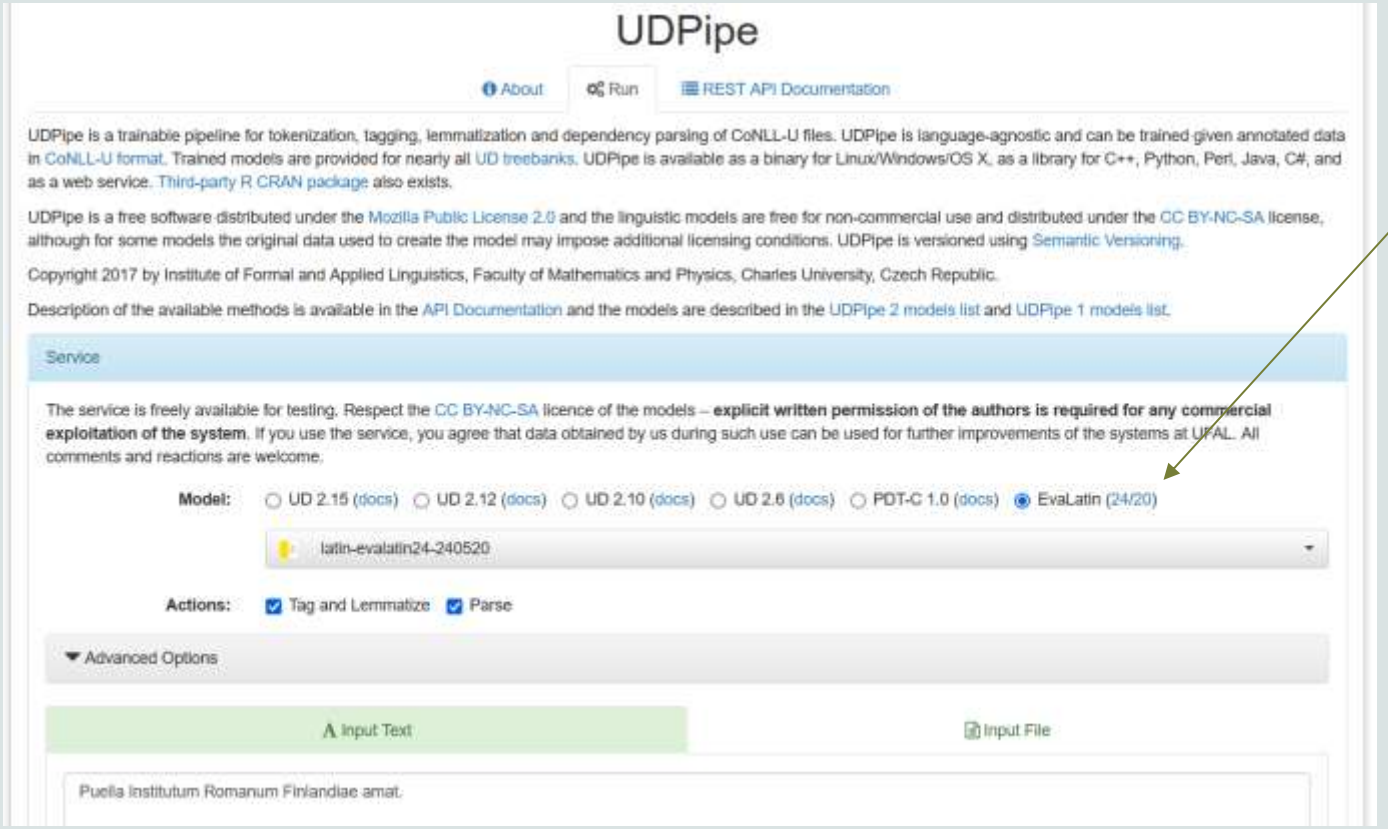
# Parsing

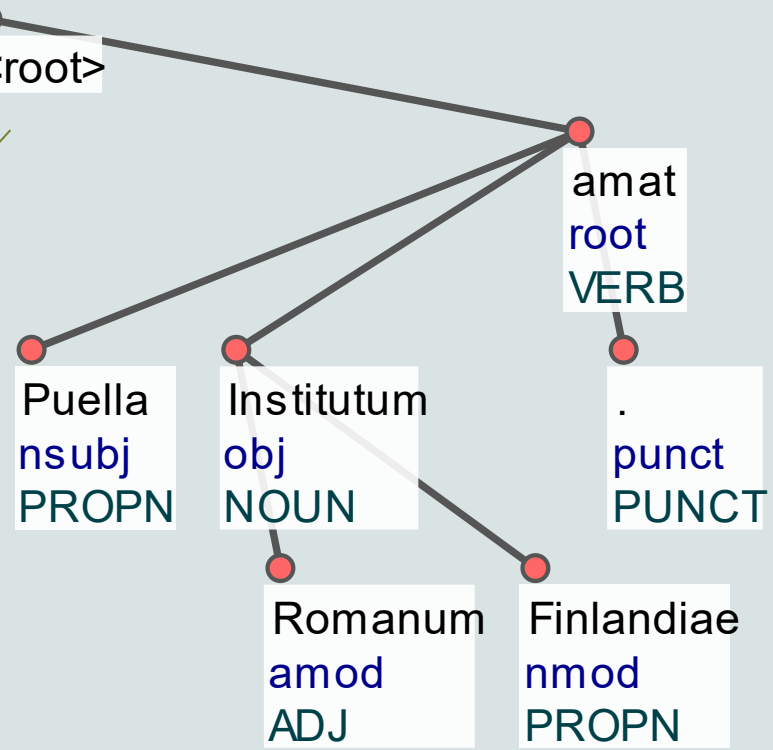
DEVELOPING NATURAL LANGUAGE PROCESSING TOOLS

# Automatic morpho-syntactic annotation tools i.e. parsers

- Tools to analyze a large amount of text
- Based on machine learning
- Algorithms learning from (large quantities of) human annotated data
- Very high accuracy for example English part-of-speech-tagging
- Produces as output tabular data that can be also visualized to trees in dependency grammar format

# Try it yourself: online graphical interface demo UD Pipe

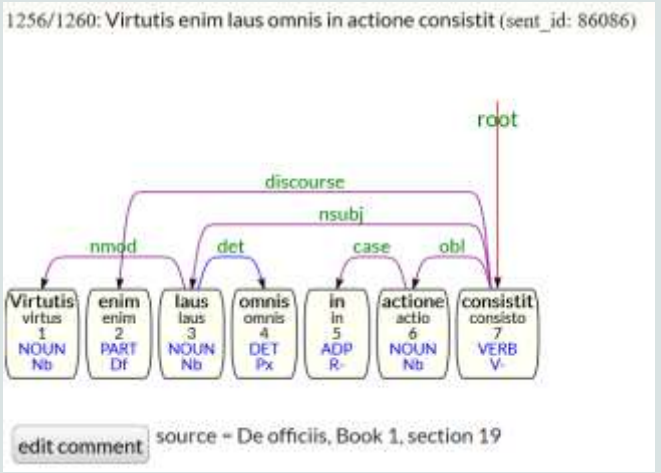



https://lindat.mff.cuni.cz/services/udpipe/

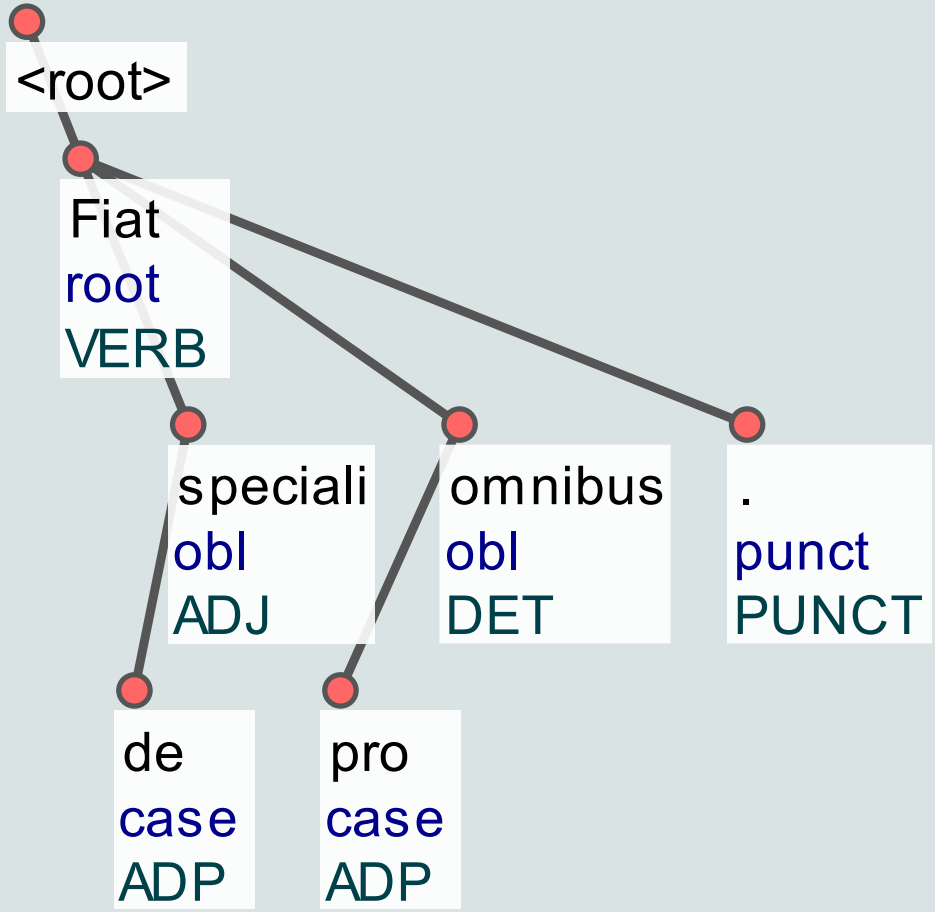
# Treebanks and CoNNL-U format

- Treebanks represent large language corpora in dependency grammar structure
- Treebank by analogy from seedbank
- A need to represent machine and human readable linguistic annotation
- CoNNL-U files that are tabulator separated text files with columns for different information types




UD for Latin
https://universaldependencies.org/la/

A sample tree made with UD Pipe 2 Online demo




‘Let it be granted to all [the supplicants] with special conditions.’

# A sample CoNNL-U table made with CoNLL-U editor

1256/1260: Virtutis enim laus omnis in actione consistit (sent_id: 86086)

| ID | FORM + − | LEMMA + − | UPOS + − | XPOS + − | FEATS + − | HEAD | DEPREL | DEPS + − | MISC + − |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Virtutis | virtus | NOUN | Nb | Case=Gen\|Gender=Fem\|Number= | 3 | nmod | _ | Ref=1.19 |
| 2 | enim | enim | PART | Df | _ | 7 | discourse | _ | Ref=1.19 |
| 3 | laus | laus | NOUN | Nb | Case=Nom\|Gender=Fem\|Number | 7 | nsubj | _ | Ref=1.19 |
| 4 | omnis | omnis | DET | Px | Case=Nom\|Gender=Fem\|Number | 3 | det | _ | Ref=1.19 |
| 5 | in | in | ADP | R- | AdpType=Prep | 6 | case | _ | Ref=1.19 |
| 6 | actione | actio | NOUN | Nb | Case=Abl\|Gender=Fem\|Number= | 7 | obl | _ | Ref=1.19 |
| 7 | consistit | consisto | VERB | V- | Aspect=Imp\|Mood=Ind\|Number= | 0 | root | _ | Ref=1.19\|Tradi |

edit comment source = De officiis, Book 1, section 19

"For all the praise of virtue lies in action."
Cicero's work *De Officiis*, Book 1, Chapter 6.

# The PeDoCo treebank

THE PENITENTIARY DOCUMENTS INTO A LARGE LANGUAGE RESOURCE

# A small test set

- Creating a new treebank
- Using the Universal Dependencies (UD) framework
- First a test set (1,200 words) is used to evaluate the performance of parsers, like Stanza or Trankit, to grammatically analyze the text
- The best parsers (Kupari et al., 2024) are then used to make the analysis of the 200,000 PeDoCo corpus

# Some observations of the PeDoCo treebank

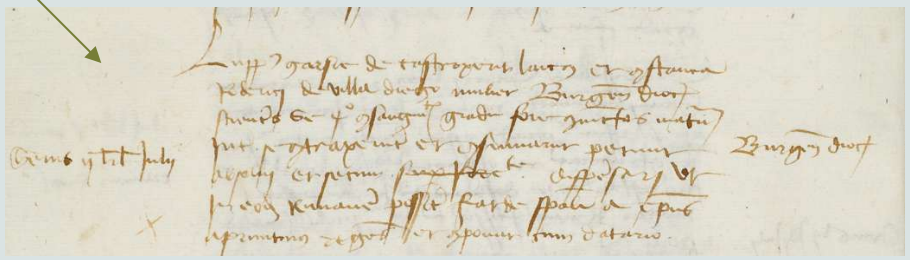
- In GS data e.g. `NOUN` 239, `VERB` 203 and `ADP` (adposition) 101 times
- In Latin participle forms are difficult to annotate (see e.g. Bamman et Burns, 2020)
- In GS all *dictus* and *predictus* annotated `DET` (determiner)
- These participle common forms have become substitutes for demonstrative and anaphoric pronouns

# The trouble of fitting everything in predetermined silos

- Illegible handwriting and corrupt passages
- E.g. For phrase *parvum cultrum ad modum †scindi pennium† vim vi repellendo et se defendendo* (AP doc. 252) lemma: *pennium*
- Diphthongs: e.g. *egritudine* lemma: *aegritudo*
- Vast occurrences of non-Latin proper names which do not follow the Latin inflection, e.g. *Courtenay, Lille* - Tag: Foreign=Yes
- https://github.com/HannaKoo/GoldStandard-penitentiary

# Conclusions

AIMING TOWARDS HIGH QUALITY COMPUTATIONAL MODELS



# Distant reading

```
<text>
<h2 num="88" bundle="n" several_wittnesses="y">88
<date when="1460-07-21">21.7 1460</date> Siena</h2>
<h3 num="88" version="a">88a</h3>
<version>
Exponitur sanctitati vestre pro parte devotorum
vestrorum
Philippi Yverson armigeri et Helene filie Henrici
mulieris coniugum
Aboensis diocesis, - - - </version>

<h3 num="88" version="b">88b</h3>
<version>
Philippus Yverson armiger et Helena Henrici mulier
Abuensis
diocesis quarto affinitatis gradu coniuncti petunt
similem gratiam sibi
fieri. Fiat de speciali. Antonius episcopus
Apruntinus regens.
</version>
</text>
```

# The benefits of the computational and treebank approach

- From original application
- To register copies
- A modern-day edition
- Using TEI XML to make a structured text to organize large amounts of text into machine readable format
- Parsing tools to add linguistic annotation
- A better understanding of medieval Latin language use that can be motivated with computational evidence from large datasets

# References

*Edition:*

*Auctoritate Papae. The Church Province of Uppsala and the Apostolic Penitentiary 1410–1526*. Sara Risberg (edition), Kirsi Salonen (introduction). Stockholm: National Archives of Sweden 2008.

*Papers:*

David Bamman, Patrick J. Burns, "Latin BERT: A Contextual Language Model for Classical Philology" https://arxiv.org/abs/2009.10053

Hanna-Mari Kristiina Kupari, Erik Henriksson, Veronika Laippala, and Jenna Kanerva. 2024. Improving Latin Dependency Parsing by Combining Treebanks and Predictions. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 216–228, Miami, USA. Association for Computational Linguistics.

# Thanks!

Institutum Romanum Finlandiae

Friends of Villa Lante Society

Keep in touch!

LinkedIn
hmknie@utu.fi

OTTO A. MALM lahjoitusrahasto

TURUN YLIOPISTOSÄÄTIÖ

EMIL AALTOSEN SÄÄTIÖ

https://github.com/HannaKoo