

# Modelling Liturgical Texts for NLP

## Challenges in Cataloguing Liturgical Manuscripts Today. A Round Table Discussion

Krzysztof Nowak

July 1, 2024

# Outline

- modelling a liturgical book for linguistic analysis
- modelling the text with TEI XML
- linkin data with Semantic Web technologies

# Motivation (1)

Wrocław, Chapter Library  
Ms. 148 (Figurski 2022)

- 2nd half of the XII c.
- prayers known from earlier mss.
- prayers unknown from earlier mss.
  - *benedictio pedulum*
  - *benedictio sandalium*

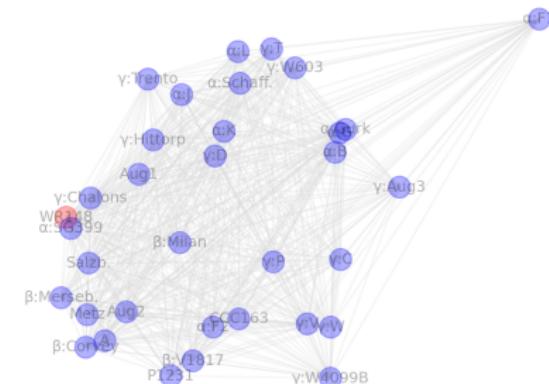
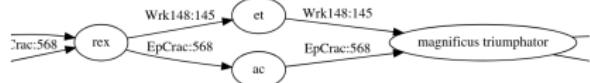
## Challenges

- linguistic annotation
  - tokens
  - lemmas
  - POS
- orthography
  - text normalization
  - abbreviation expansion
- textual variants
- sequential features
- metadata

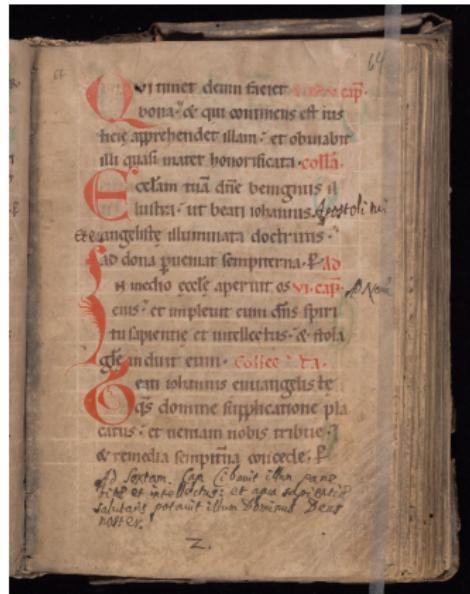
# Motivation (2)

- single-text analysis
  - number of tokens, lemmas
  - vocabulary richness
  - most frequent n-grams
- comparative analysis
  - automatic collation
  - clustering

	orth	count
4781	persecutor	1
4793	persuasionibus	1
4794	pertinax	1
4954	possessione	1
6683	triumphatorem	1



# Case Study: Beinecke MS 883



Beinecke Rare Book and  
Manuscript Library MS 883

## Collectarium Landense / Cisterciense (s. XII 3/4)

Paweł Figurski\*

Institytut Sztuki Polskiej Akademii Nauk w Warszawie

„Liturgica Polonica”:  
katalog polskich rękopisów liturgicznych  
(cz. 1: do około 1300 r.).

Założenia projektu i propozycja  
wzoru opisu katalogowego  
na przykładzie kolektarza z Ładu  
(New Haven (CT), Yale University, Beinecke Rare  
Book and Manuscript Library, MS 883)\*\*

Projekt finansowany przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach  
Narodowego Programu Rozwoju Humanistyki, realizowany w Instytucie Sztuki PAN  
w latach 2022–2027, ma na celu przygotować szczegółowych opisów katalogowych  
około 110 rękopisów liturgicznych powstających do ok. 1300 roku, które znajdują się  
obecnie w zbiorach polskich, oraz takich manuskryptów, o których wiele, ze stałym  
związek ze środowiskiem Polski, do dziś znajduje się w instytucjach zagranicznych.



(Figurski et al. 2024)

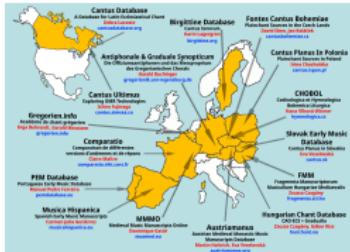
# Projects

## Cantus: A Database for Latin Ecclesiastical Chant - Inventories of Chant Sources



CANTUS: A Database for Latin Ecclesiastical Chant

## Cantus Index: Catalogue of Chant Texts and Melodies



Krzysztof Nowak

Modelling Liturgical Texts for NLP

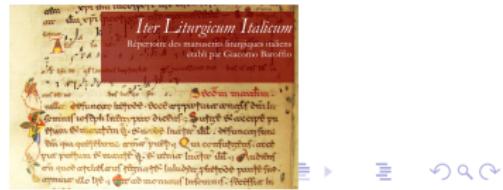
## Fragmentarium



## Usuarium



## Iter Liturgicum italicum



July 1, 2024

6 / 15

# XML modelling: schemas

- TEI
  - 10 Manuscript Description
  - 11 Representation of Primary Sources
  - 12 Critical Apparatus
- TEI P5 Customization and Encoding Guidelines for Manuscript Description
- Schémas pour les notices reliures (BNF)

# XML modelling: examples

manuscripta.se

```
<bibl>
    <title type="uniform">
        Oratio: Ecclesiam tuam quae sumus domine benignus illustra ut beati
        Iohannis
    </title>
    <textLang mainLang="la"/>
    <note xml:lang="la" type="incipit">
        Ecclesiam tuam quae sumus domine benignus illustra ut beati Iohannis
    </note>
    <bibl>Corpus orationum nr 2416 c. Bruylants nr 276.</bibl>
</bibl>
```

## CURSUS. An Online Resource of Medieval Liturgical Texts

```
<prayer previd="o1662" nextid="o1664" id="o1663">
    <header>
        <usage ms="Cdm" code="02122600">[26r] commemoration</usage>
        <usage ms="Alb2" code="02122600">[33v] commemoration</usage>
        <usage ms="Evm" code="02122600">[42r] vesp2</usage>
        <usage ms="Hyd" code="02122600"/>
        <usage ms="Wcb" code="02122700">[60v] vespl</usage>
        <notesStmt>
            <note resp="DC" place="unspecified" anchored="yes"> Deshusses no.67</note>
        </notesStmt>
    </header>
    <pBody wit="Hyd Alb2 Evm Wcb Cdm">
        Ecclesiam tuam quae sumus domine benignus illustra ut beati Iohannis evangeliste illuminata doctrinis ad dona perveniat sempiterna
    </pBody>
</prayer>
```

# XML modelling: codicological information

## granularity in text-oriented data

### data normalization

```
<dimensions unit="mm" type="leaf" xml:id="leaf.2">
  <height min="65" max="70">65–70</height>
  <width quantity="155">155 mm</width>
</dimensions>
```

## managing identifiers

```
<msDesc xml:lang="en" xml:id="liturgica_113">
  <msIdentifier>
    <country>United States of America</country>
    <settlement>New Haven (CT)</settlement>
    <institution>Yale University</institution>
    <repository>Beinecke Rare Book and Manuscript
      Library</repository>
    <idno type="shelfmark">Ms. 883</idno>
    <altIdentifier type="external">
      <idno type="MSPL">EXPL 599</idno>
    </altIdentifier>
    <altIdentifier type="external">
      <idno type="RISM">US-NHub Ms. 883</idno>
    </altIdentifier>
  </msIdentifier>
</msDesc>
```

## managing terminology

```
<dimensions unit="mm" type="leaf" xml:id="leaf.2">
  <height min="65" max="70">65–70</height>
  <width quantity="155">155 mm</width>
</dimensions>
```

# XML modelling: textual content

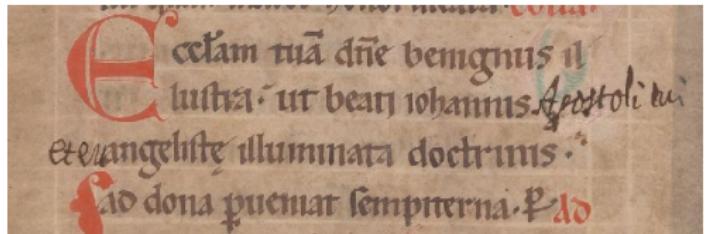
## textual content in the catalogue

```
<msItem xml:id="liturgica_113_msItem_1_10">
<rubric>
  In sabbato adventus Domini
  <locus from="7v">(f. 7v)</locus>;
</rubric>
<rubric>
  Dominica
  <locus from="8r">(f. 8r)</locus>;
</rubric>
<rubric>
  Sabbato ante secundam dominicam
  <locus from="8v">(f. 8v)</locus>;
</rubric>
<rubric>
  Dominica secunda
  <locus from="9r">(f. 9r)</locus>;
</rubric>
<rubric>
  Sabbato ante <supplied>III</supplied> dominicam
  <locus from="9r">(f. 9r)</locus>;
</rubric>
</msItem>
```

## textual variants

```
<p>
<expan>Eccl<ex>esi</ex>am</expan> <expan>tua<ex>m</ex></expan>
<expan>d<ex>omi</ex>ne</expan> benignus il<lb break="no"/>
lustra ut beati Iohannis <add hand="#hand-suppl-1">Apostoli tui <lb/>
et eu</add>angeliste illuminata doctrinis
ad dona <expan>p<ex>er</ex>ueniat</expan> sempiterna.
<abbr type="word" subtype="formula">P<ex>er</ex></abbr>.
```

## standalone linguistic analysis



# Semantic Modelling: goals and models

## Goals

- unique identifiers
- explicit interpretation
- managing terminology
- data accessibility
- harmonizing datasets

## Previous Research

- bibliographic information
  - SPAR Ontologies
- circulation of ancient and medieval works
  - Sharing Ancient Wisdoms
  - Biblissima
- manuscript description
  - The Medieval Manuscript Ontology (MeMO)
- controlled vocabularies and reconciliation
  - Simple Knowledge Organization System (SKOS)

# Semantic Modelling: identification and description

```
@prefix ex: <https://w3id.org/irnerio/data/memo/> .  
@prefix memo: <https://w3id.org/irnerio/ontology/memo/> .  
  
# composition  
ex:liturgica_cod_113 a memo:Codex ;  
    dcterms:identifier "liturgica_cod_113" ;  
    frbr:part ex:liturgica_cod_113_binding_1 ,  
    ex:liturgica_cod_113_folio_64,  
    ex:liturgica_cod_113_folio_151 .  
  
# single folio  
ex:liturgica_cod_113_folio_64 a memo:Folio ;  
    dcterms:medium memo:parchment ;  
    dcterms:extent ex:liturgica_cod_113_ext_1 ;  
    frbr:part ex:recto_liturgica_cod_113_folio_64,  
              ex:verso_liturgica_cod_113_folio_64 .  
  
# dimensions  
ex:liturgica_cod_113_ext_1 a memo:SizeMeasurement ;  
    memo:hasSizeMeasurementUnit memo:millimeters ;  
    memo:hasLength "154"^^xsd:positiveInteger ;  
    memo:hasWidth "102"^^xsd:positiveInteger .
```

# Semantic Modelling: textual content

```
@prefix ex: <https://w3id.org/irnerio/data/memo/> .  
@prefix memo: <https://w3id.org/irnerio/ontology/memo/> .  
  
# intellectual content  
ex:liturgica_ms_113_1 a memo:Manuscript ;  
    frbr:part ex:liturgica_113_text_118;  
    memo:hasTextualMetadata ex:liturgica_113_rubric_1 ,  
                            ex:liturgica_113_text_118 .  
  
# textual entity  
ex:liturgica_113_text_118 a memo:Text ;  
    # text string  
    literal:hasLiteral [  
        a literal:Literal;  
        literal:hasLiteralValue "Ecclesiam tuam domine benignus illustra  
                                ut beati Iohannis euangeliste illuminata  
                                doctrinis ad dona perueniat sempiterna"  
    ] ;  
    # locus  
    frbr:embodiment ex:liturgica_cod_113_folio_64 .
```

# Semantic Modelling: linking

## texts

```
ex:liturgica_113_text_118 a memo:Text ;
  # text string
  literal:hasLiteral [
    a literal:Literal;
    literal:hasLiteralValue "Ecclesiam tuam domine benignus illustra
      ut beati Iohannis euangeliste illuminata
      doctrinis ad dona perueniat sempiterna"
  ];
  # locus
  frbr:embodiment ex:liturgica_cod_113_folio_64 .
```



A screenshot of a digital manuscript page. The page contains Latin text and some marginalia. Below the page, there is a detailed list of metadata:

- Title: Oratio: Ecclesiam tuam quæsumus Domine benignus illustra ut beati Iohannis
- Incipit: Ecclesiam tuam quæsumus domine benignus illustra ut beati Iohannis
- Manuscript ID: 102188
- Stable URL: <https://www.manuscripta.net/eor/102188>
- Size: 390 x 248 mm
- Licence: CC0 1.0 Universal
- Last revision: 2019-06-10

Below the metadata, there is a section titled "ASSOCIATED MANUSCRIPTS" with a table:

Repository	Shelfmark	Heading
Kungliga Biblioteket	A 36	Honorum med biljet på svenska

## vocabularies

```
ex:liturgica_cod_113_folio_64 a memo:Folio ;
  dcterms:medium memo:parchment ;
  dcterms:extent ex:liturgica_cod_113_ext_1 ;
  frbr:part ex:recto_liturgica_cod_113_folio_64,
  ex:verso_liturgica_cod_113_folio_64 .
```

A screenshot of a detailed vocabulary entry for "Parchment". The entry includes:

- Ambient:** Parchment
- Type:** Internal (Resource) / Comment / Shell (Basic)
- Values:** Tong (Latinized animal)
- Definition:** Parchment is a writing material made from specially prepared parchment skin of an animal, especially a sheep or goat, which has been treated so that it is used for writing or drawing. It is a writing material made from specially prepared parchment skin of an animal, especially a sheep or goat, which has been treated so that it is used for writing or drawing. It is a writing material made from specially prepared parchment skin of an animal, especially a sheep or goat, which has been treated so that it is used for writing or drawing.
- Relationships:**
  - skos:closeMatch: Tong
  - skos:closeMatch: Animal
  - skos:closeMatch: Parchment
  - skos:closeMatch: Paper
  - skos:closeMatch: Skin

## identifiers

```
ex:liturgica_cod_113 a memo:Codex ;
  dcterms:identifier "liturgica_cod_113" ;
  frbr:part ex:liturgica_cod_113_binding_1 ,
  ex:liturgica_cod_113_folio_64,
  ex:liturgica_cod_113_folio_151 .
```

A screenshot of the Orbis library catalog. It shows a book record for "Oratio: Ecclesiam tuam quæsumus Domine benignus illustra ut beati Iohannis". The record includes fields such as Title, Author, Publication Information, and Call Number. Below the catalog, there is a "Table Holdings" section and a "Location" section.

# Bibliography

-  Figurski, Paweł (2022). "Beyond the Monarchical Church Model: Liturgy, Manuscripts, and Bishops in the Kingdom of Poland. Case Studies from the Eleventh and Twelfth Centuries". In: *Die 'Episkopalisierung Der Kirche' Im Europäischen Vergleich*. Ed. by Andreas Bihrer and Hedwig Röcklein. 1st ed. Studien Zur Germania Sacra 13. Boston: De Gruyter Akademie Forschung, pp. 361–404. ISBN: 978-3-11-077665-2.
-  Figurski, Paweł et al. (2024). "*Liturgica Poloniae*: Katalog Polskich Rękopisów Liturgicznych (Cz. 1: Do Około 1300 r.) : Założenia Projektu i Propozycja Wzoru Opisu Katalogowego Na Przykładzie Kolektarza z Lądu (New Haven (CT), Yale University, Beinecke Rare Book and Manuscript Library, MS 883)". In: *Codices et Fragmenta: Średniowieczny Kodeks Rękopiśmienny w Aspekcie Historycznego Trwania*. Ed. by Monika Jakubek-Raczkowska and Marta Czyżak. Studia Nad Skryptorium i Spuścizną Rękopiśmienną Średniowiecza. Toruń: Wydawnictwo Naukowe UMK, pp. 199–222. 