

Identification and Localisation of Formulaic Language using Document-Term-Network- Visualizations

Sebastian Gensicke

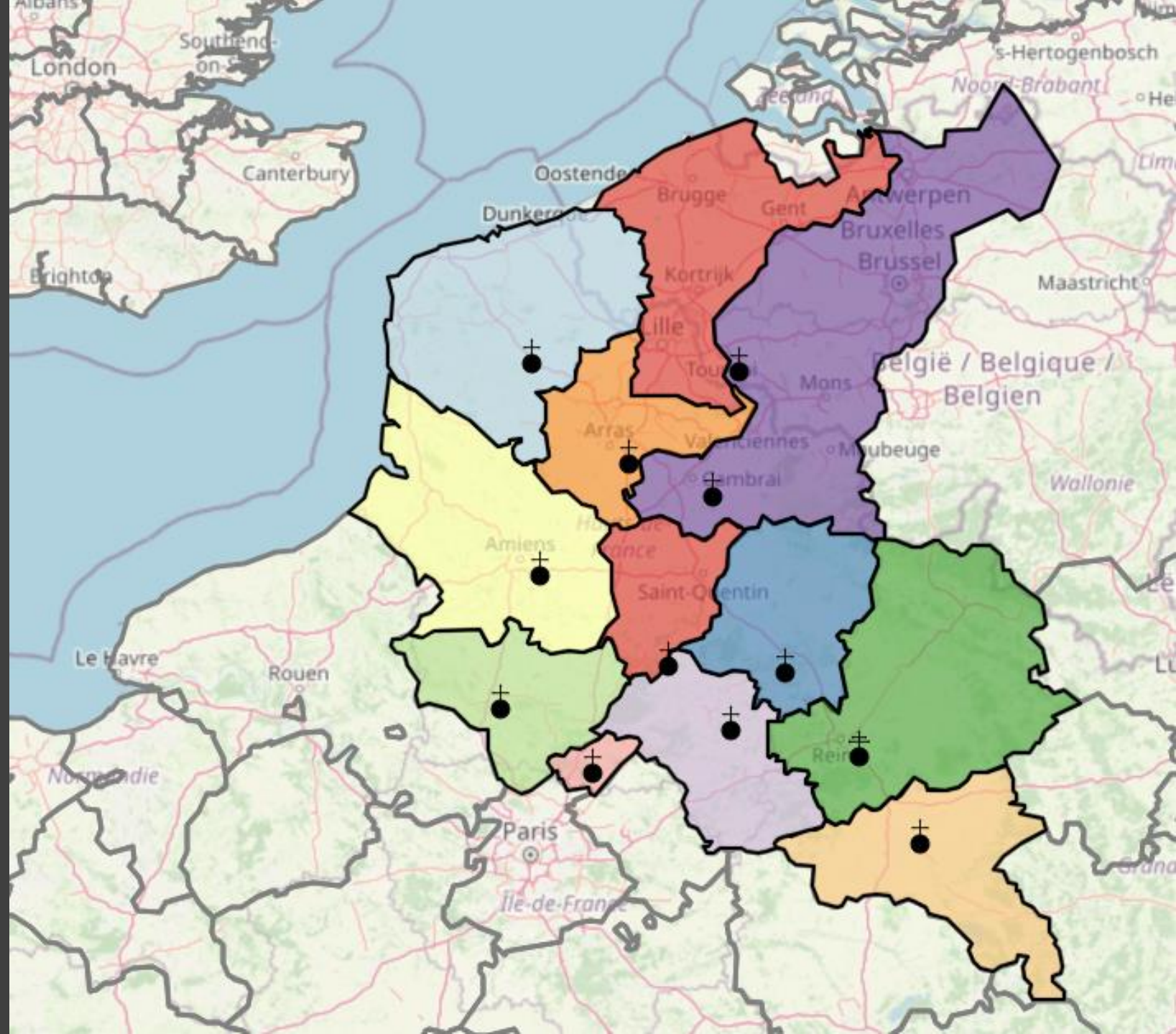


Die Formierung Europas
durch Überwindung der
Spaltung im 12. Jahrhundert

RWTHAACHEN
UNIVERSITY

HIBO
HISTORISCHES INSTITUT BOCHUM

Ecclesiastical province of Reims

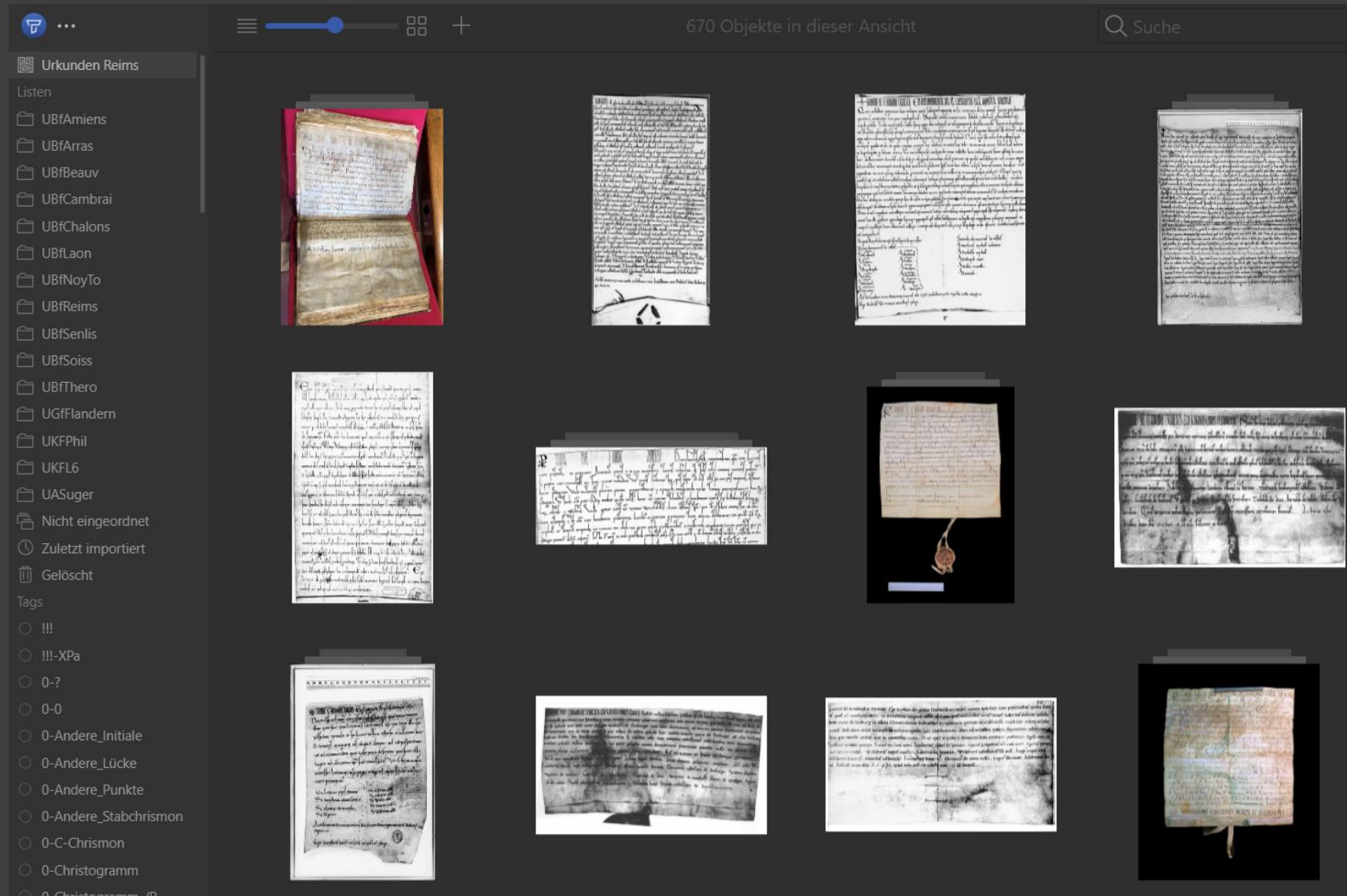


Database of tagged Images (Tropy)

670 Objekte in dieser Ansicht

Suche

- Urkunden Reims
- Listen
 - UBfAmiens
 - UBfArras
 - UBfBeauv
 - UBfCambrai
 - UBfChalons
 - UBfLaon
 - UBfNoyTo
 - UBfReims
 - UBfSenlis
 - UBfSoiss
 - UBfThero
 - UGfFlandern
 - UKFPhil
 - UKFL6
 - UASuger
 - Nicht eingeordnet
 - Zuletzt importiert
 - Gelöscht
- Tags
 - !!!
 - !!!-XPa
 - 0-?
 - 0-0
 - 0-Andere_Initiale
 - 0-Andere_Lücke
 - 0-Andere_Punkte
 - 0-Andere_Stabchrismon
 - 0-C-Christmon
 - 0-Christogramm
 - 0-Christogramm-/P



XML-Database

Sources:

- Chartes originales
- Chartae Galliae
- Diplomata Belgica
- Print-Editions
(OCR and PDF)
- Transcriptions

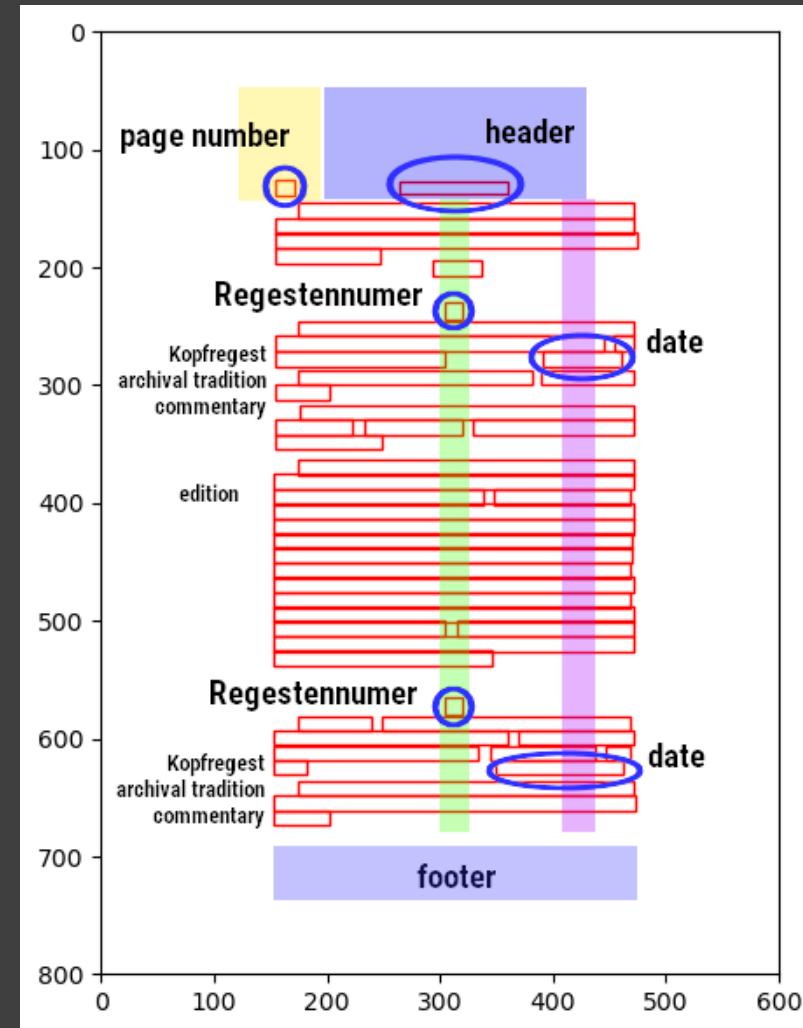
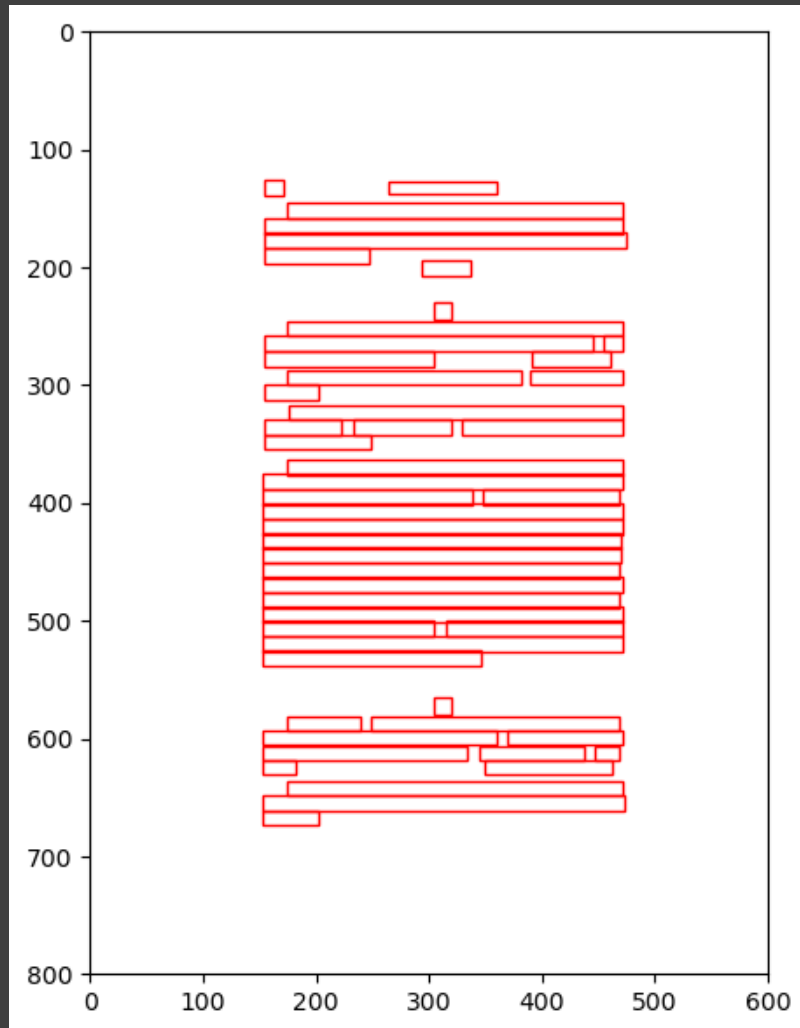
„Urkteile“:

- Invocatio
- Intitulatio
- Inscriptio (Adress + Salutation)
- Arenga
- Publicatio
- Narratio
- Dispositio
- Corroboratio
- Sanctio
- Aprecatio
- Subscriptio (Issuer)
- Subscriptiones (Witnesses)
- Datatio
- notariusSub (Chancellor)
- Devise (Chirographs)

```
<charter>
  <idno>UBfReims-204</idno><authentic
  <form>Kopie</form><text_source>http
  <tenor>
    <invocatio>In nomine sancte [et
    <publicatio>Notum sit omnibus p
    <dispositio>quid et quomodo de
    <intitulatio>Ego Rainaldus eccl
    <dispositio>beneficium quod Rem
    <corroboratio>Quia ergo sicut s
    <subscriptiones>Signum Odonis a
    <datatio>Actum Remis anno incar
    <notariusSub>Fulcradus cancella
  </tenor></charter>
```

```
<charter>
  <idno>UBfReims-205</idno><authentic
  <form>Kopie</form><text_source>Edit
  <tenor>
    <intitulatio>Ego Rainaldus Reme
    <publicatio>seriem cause excomm
    <dispositio>Excommunicaveramus
    <corroboratio>Hanc igitur aucto
    <subscriptiones>Signum Odonis a
    <datatio>Actum Remis anno incar
```

Import: Making information implied in layout digitally explicit

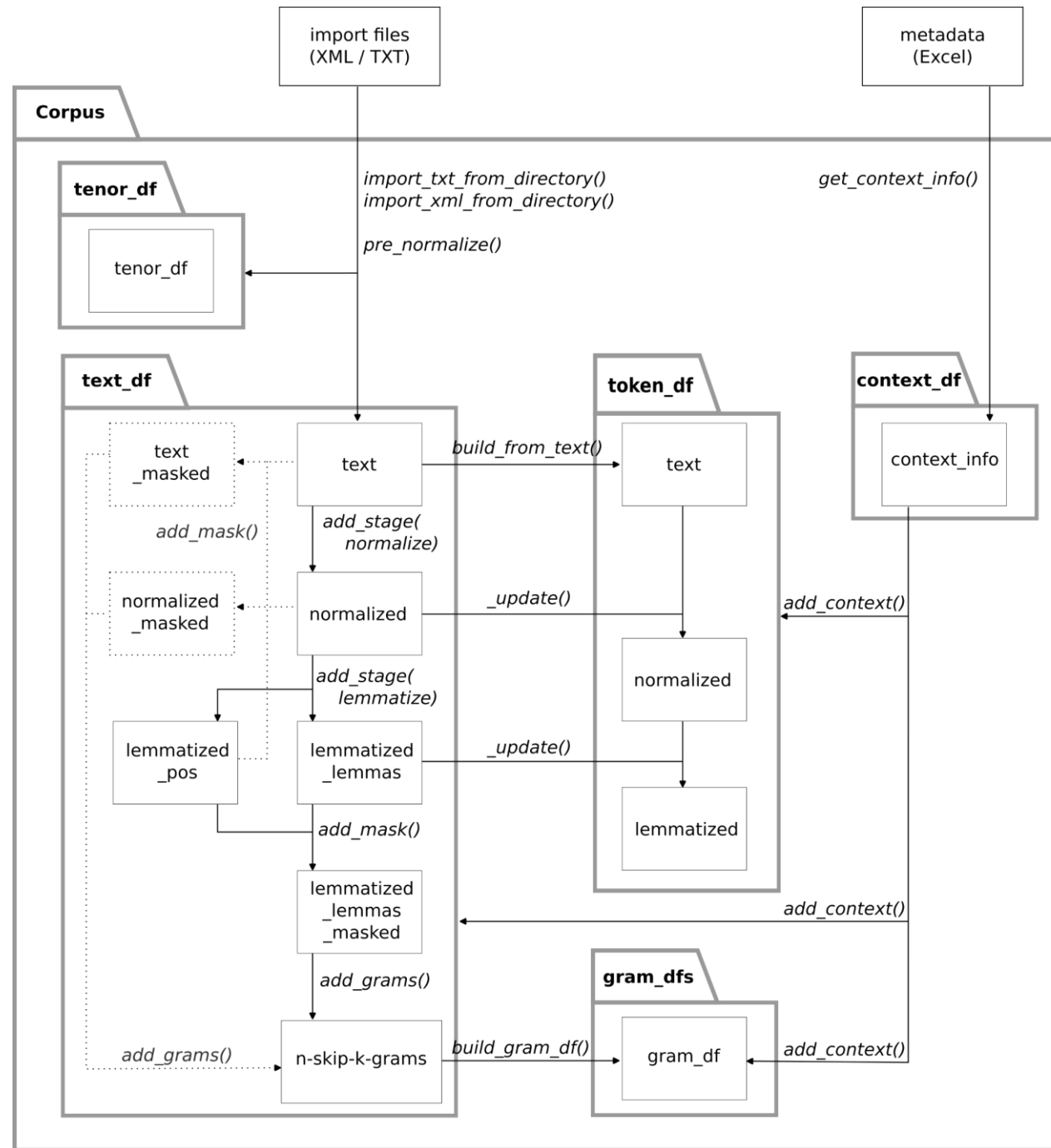


https://github.com/SGensicke/PDFtoCSV_extract_regesta

Spreadsheet with Metadata

| | UrkID | Status | AusstellerID | Ausstelle | Ort | OrtErschlo | OrtKomm | DatumVerl | DatumA | DatumE | EmpfängerI | EmpfängerInst |
|------|---------------------|------------|------------------|--|-----|------------|---------|-----------------------|-------------------|-------------------|-----------------|------------------------------------|
| 1719 | UBfReims-130 | | PERS-0005 | Manassès II de Châtillon, Erzbischof v. Reims (1096-11 | | | | 1102 | 1102-01-01 | 1102-12-31 | INST-139 | Douai, collégiale Saint-Amé |
| 1720 | UBfReims-131 | | PERS-0005 | Manassès II de Châtillon, Erzbischof v. Reims (1096-11[1102] | | | | 1102 | 1102-01-01 | 1102-12-31 | | |
| 1721 | UBfReims-132 | | PERS-0005 | Manassès II Reims | | | | 1103, après le | 1103-03-30 | 1103-12-31 | INST-360 | Saint-Thierry-lès-Reims, abbaye |
| 1722 | UBfReims-133 | | PERS-0005 | Manassès II Reims | | | | 1103, après le | 1103-03-30 | 1103-12-31 | INST-215 | Reims, abbaye Saint-Nicaise |
| 1723 | UBfReims-134 | | PERS-0005 | Manassès II Reims | | | | 1103, après le | 1103-03-30 | 1103-12-31 | INST-197 | Reims, abbaye Saint-Remi |
| 1724 | UBfReims-135 | | PERS-0005 | Manassès II Reims | | | | 1103, après le | 1103-03-30 | 1103-12-31 | INST-214 | Reims, abbaye Saint-Denis |
| 1725 | UBfReims-136 | | PERS-0005 | Manassès II Reims | | | | 1104, n. st., 30 | 1104-03-30 | 1104-04-17 | INST-131 | Cluny, abbaye Saint-Pierre et Sain |
| 1726 | UBfReims-137 | | PERS-0005 | Manassès II Reims | | | | 1104, après le | 1104-03-30 | 1104-12-31 | INST-197 | Reims, abbaye Saint-Remi |
| 1727 | UBfReims-138 | | PERS-0005 | Manassès II de Châtillon, Erzbischof v. Reims (1096-11[1104- 5 octol | | | | 1104 | 1104-10-05 | 1104-12-30 | | |
| 1728 | UBfReims-139 | | PERS-0005 | Manassès II Reims | | | | 1104, ler sept | 1104-09-01 | 1105-03-30 | INST-168 | Liessies, abbaye Saint Lambert |
| 1729 | UBfReims-140 | | PERS-0005 | Manassès II de Châtillon, Erzbischof v. Reims (1096-11 | | | | 1105, 30 mars | 1105-03-30 | 1105-08-03 | INST-162 | Laon, abbaye Saint-Vincent |
| 1730 | UBfReims-141 | Deperditum | PERS-0005 | Manassès II de Châtillon, Erzbischof v. Reims (1096-11[1105] | | | | 1105 | 1105-01-01 | 1105-12-31 | INST-024 | en Ardenne, abbaye Saint-Huber |
| 1731 | UBfReims-142 | | PERS-0005 | Manassès II Reims | | | | 1106, avant le | 1106-01-01 | 1106-03-30 | INST-214 | Reims, abbaye Saint-Denis |
| 1732 | UBfReims-143 | | PERS-0005 | Manassès II Reims | | | | 1106, avant le | 1106-01-01 | 1106-03-30 | INST-214 | Reims, abbaye Saint-Denis |
| 1733 | UBfReims-144 | | PERS-0005 | Manassès II Reims | | | | 1106, avant le | 1106-01-01 | 1106-03-30 | INST-197 | Reims, abbaye Saint-Remi |
| 1734 | UBfReims-145 | | PERS-0005 | Manassès II Reims | | | | 1106, 30 mars | 1106-03-30 | 1106-08-03 | INST-215 | Reims, abbaye Saint-Nicaise |
| 1735 | UBfReims-146 | | PERS-0005 | Manassès II Reims | | | | 1106, 30 mars | 1106-03-30 | 1106-09-18 | INST-197 | Reims, abbaye Saint-Remi |

Processing in Python



Normalization

```
def normalize(text):
    # normalize common forms
    normalization_patterns = {'v': 'u',
                              'j': 'i',
                              'y': 'i',
                              'ae': 'e',
                              'ë': 'e',
                              'æ': 'e',
                              'œ': 'oe',
                              'ę': 'e',
                              'ę': 'e',
                              'o': 'o',
                              'o': 'o'}
    for pattern, replacement in normalization_patterns.items():
        text = re.sub(pattern, replacement, text)
    return text
```

- delete all punctuation
- „aggressive“ towards
 $ae \rightarrow e$

Lemmatization (and Part-of-Speech-Tagging)

| Token | Lemma | POS |
|--------------|--------------|------------|
| in | in | PRE |
| nomine | nomen | SUB |
| sancte | sanctus1 | QLF |
| et | et | CON |
| indiuidue | indiuiduus | QLF |
| trinitatis | trinitas | SUB |

TreeTagger + parameter by *Omnia/Glossaria*

See: <https://www.glossaria.eu/treetagger/>

Mask Named Entities

Thanks to the POS-Tagging by *TreeTagger*

| Token | Lemma | POS | Maske |
|--------------|---------------|------------|--------------|
| ego | ego | PRO | ego |
| robertus | §robertus | NAM | NAME |
| dei | deus | SUB | deus |
| gratia | gratia2 | SUB | gratia2 |
| atrebatensis | §atrebatensis | NAM | NAME |
| episcopus | episcopus | SUB | episcopus |

Tokenization

1. Create *n*-grams or *n-skip-k*-grams

Sentence: ... *excommunicaueramus predictum comitem pro injustis...*

3-gram 1: [excommunicaueramus predictum comitem]

3-gram 2: [predictum comitem pro]

3-gram 3: [comitem pro injustis]

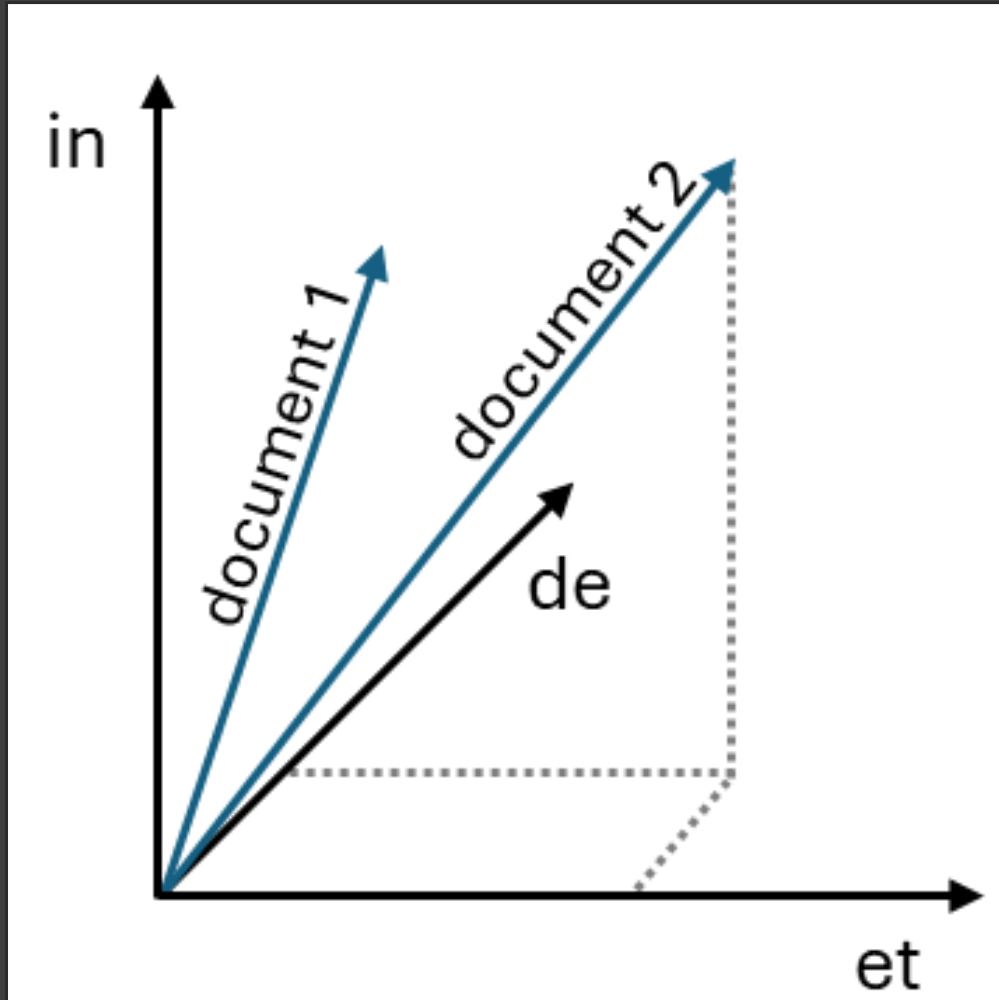
2. Sort words alphabetically

3-gram 1: [comitem, excommunicaueramus, predictum]

3-gram 2: [comitem, predictem, pro]

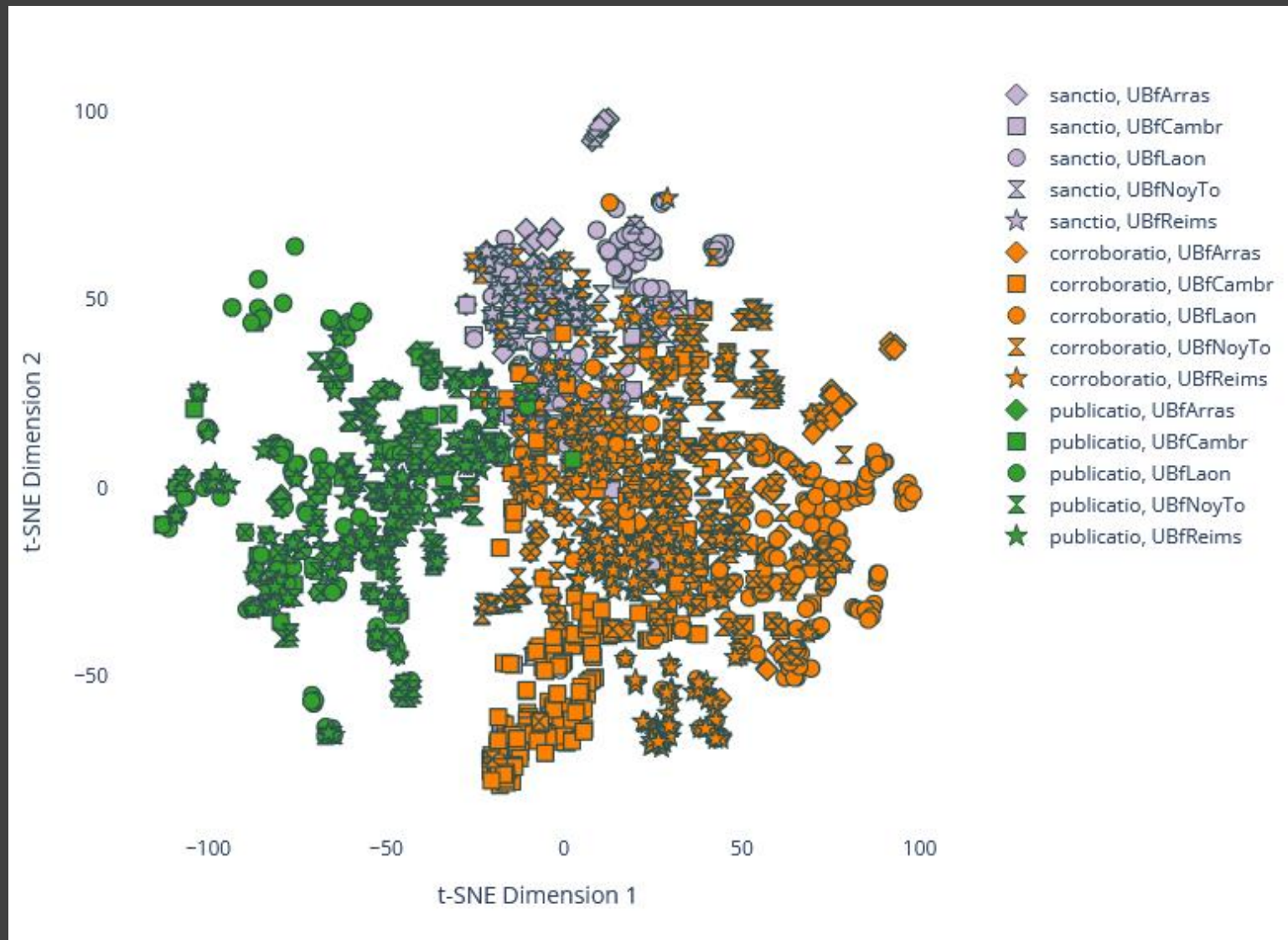
3-gram 3: [comitem, injustis, pro]

Text vectorization

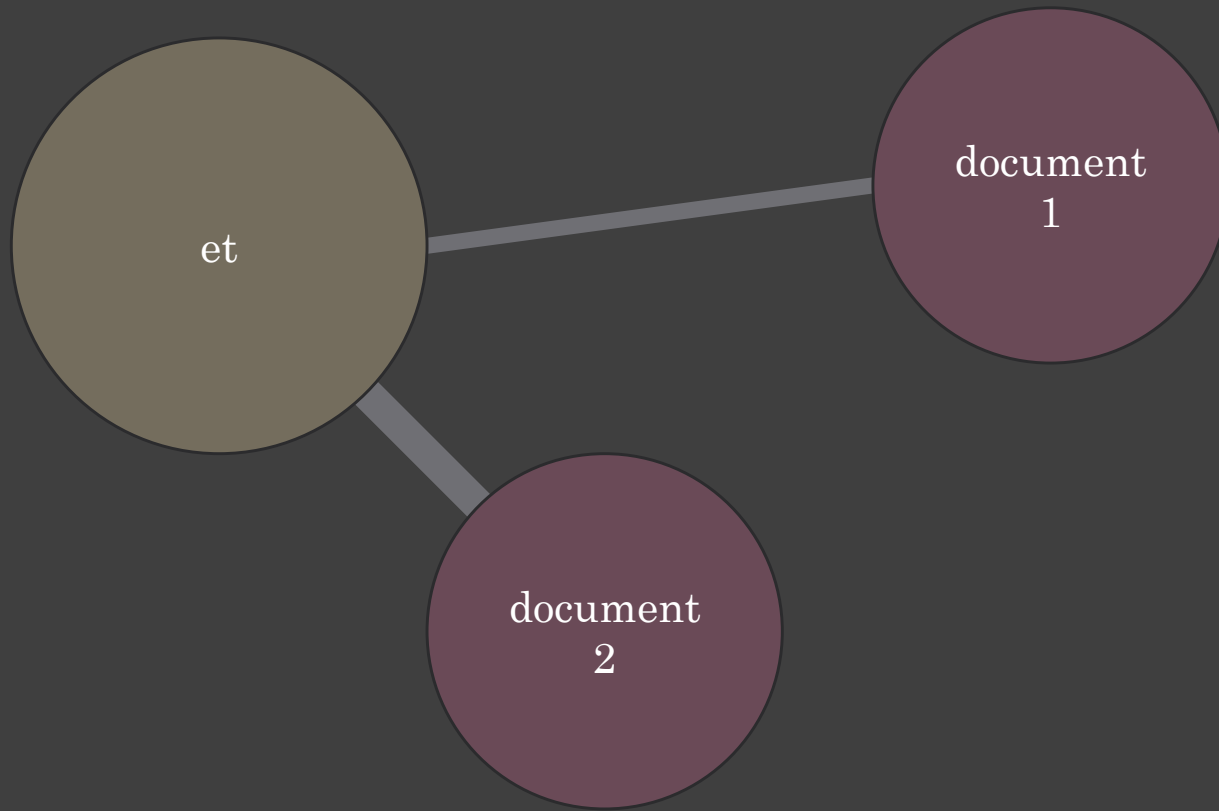


| term | document 1 | document 2 | ... |
|----------|------------|------------|-----|
| et | 15 | 20 | ... |
| in | 12 | 15 | ... |
| de | 14 | 5 | ... |
| ecclesie | 4 | 3 | ... |
| ad | 2 | 5 | ... |
| ut | 0 | 1 | ... |
| sancti | 4 | 2 | ... |
| quod | 3 | 6 | ... |

1000 MFW, TFIDF, tSNE

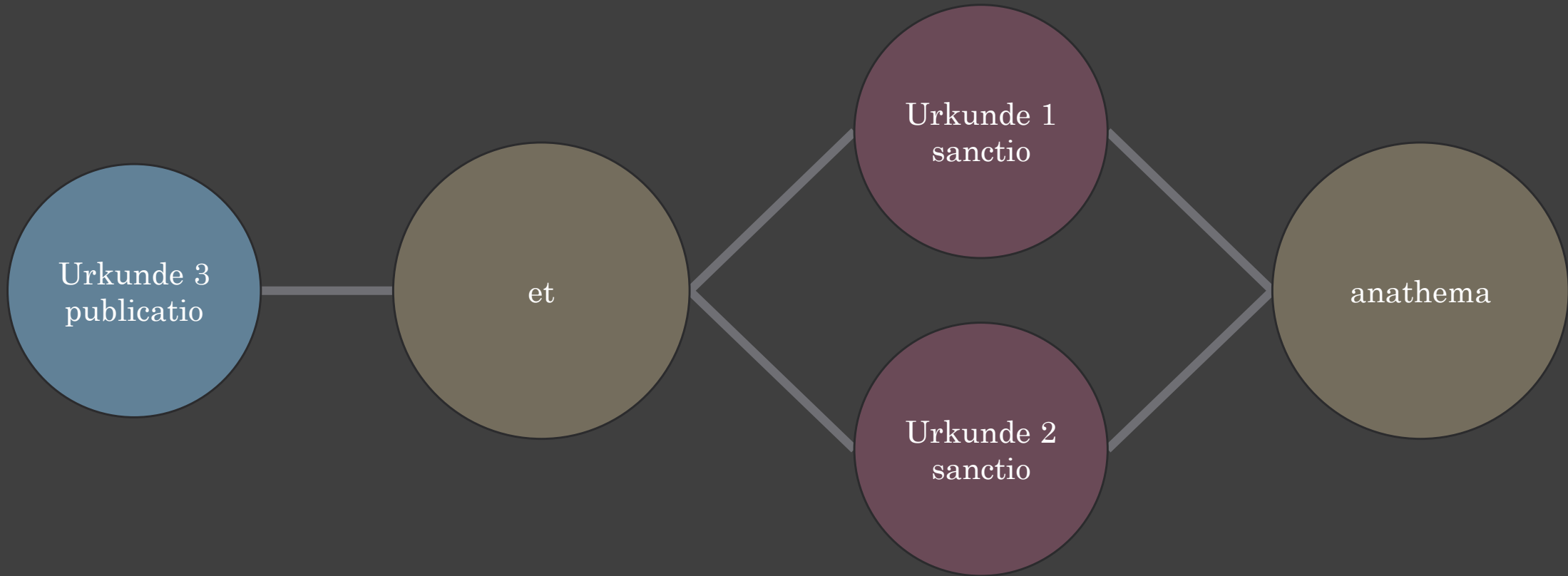


Document-Term- Network visualizations (Text as a graph?)

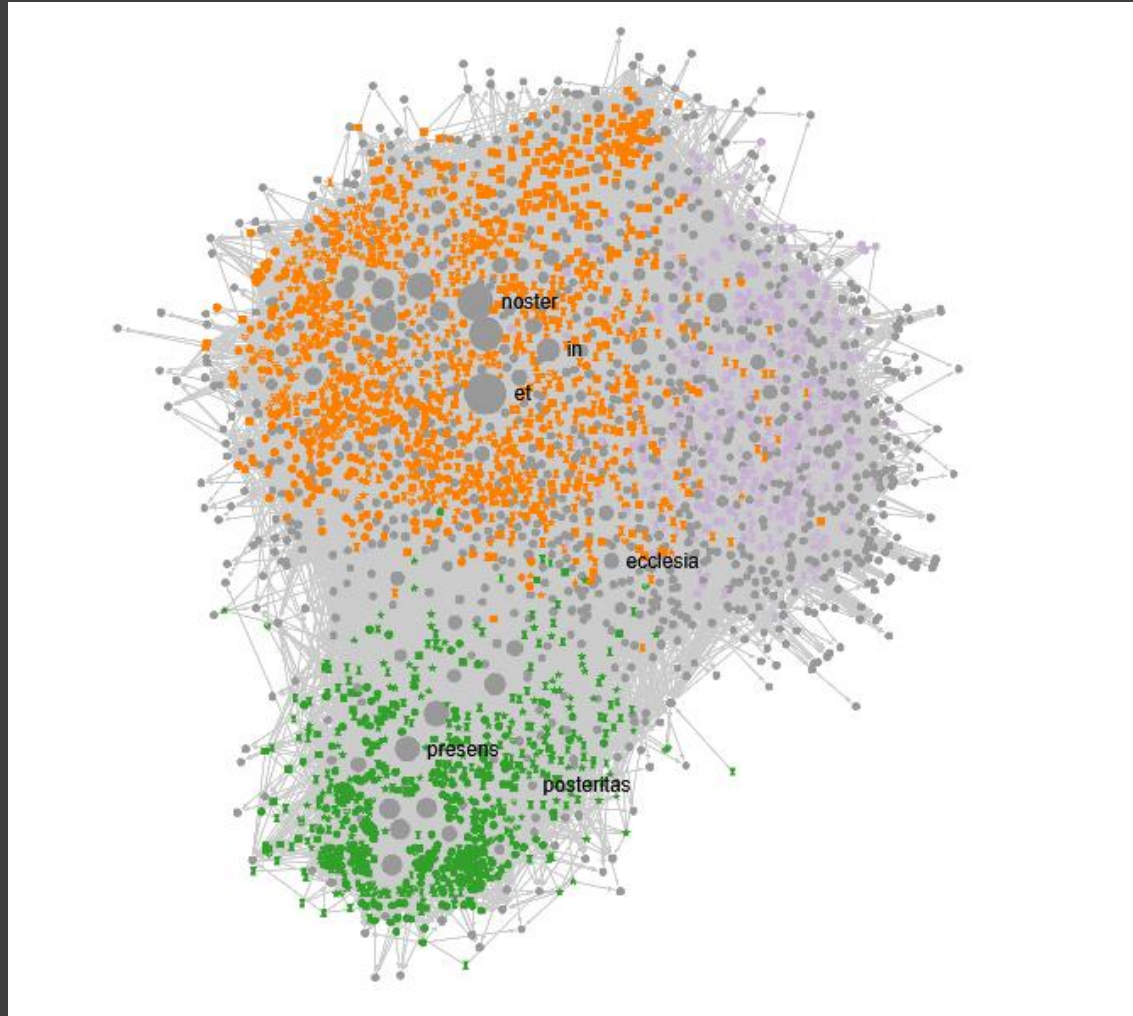


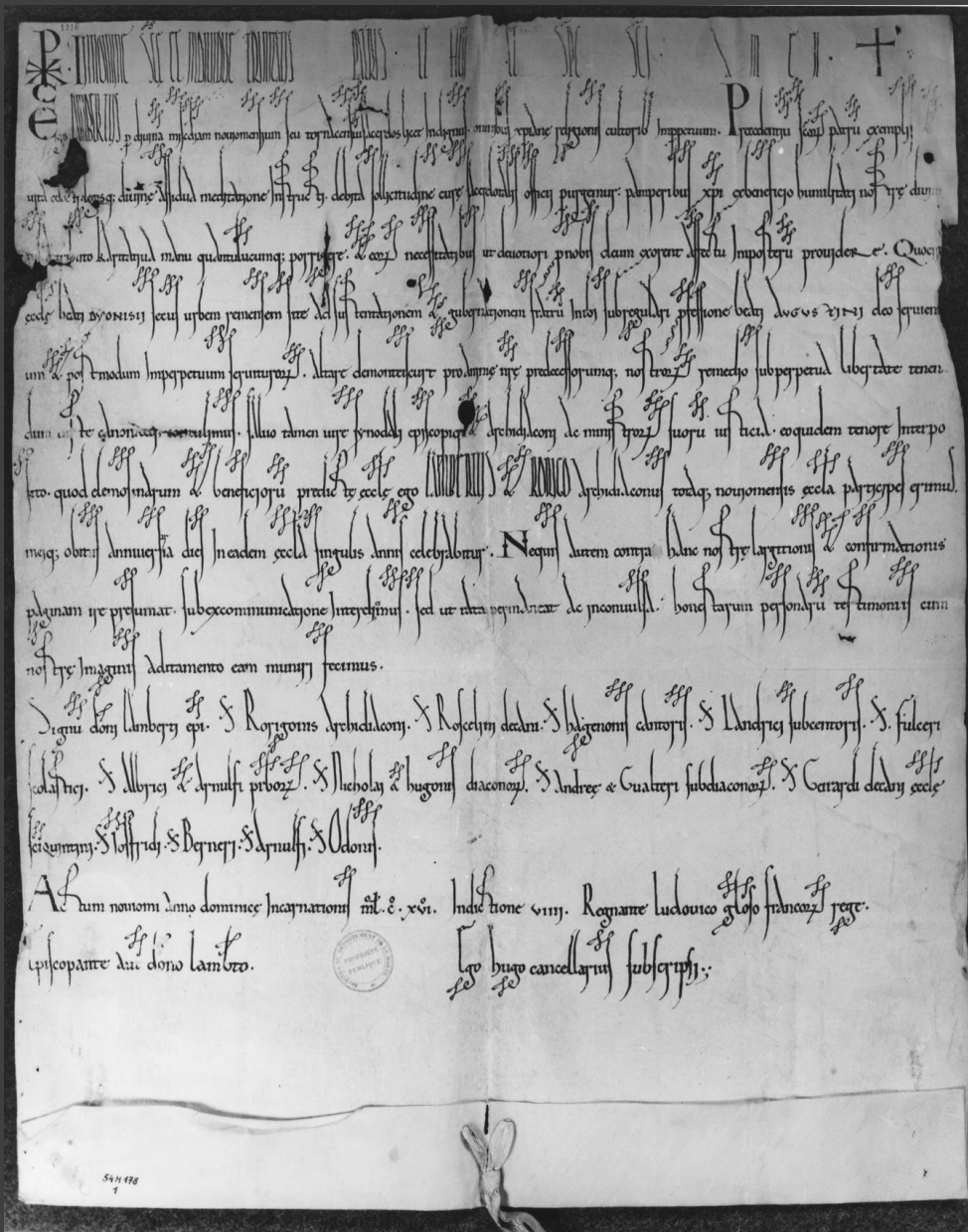
| term | document 1 | document 2 | ... |
|-------------|-------------------|-------------------|------------|
| et | 15 | 20 | ... |
| in | 12 | 15 | ... |
| de | 14 | 5 | ... |
| ecclesie | 4 | 3 | ... |
| ad | 2 | 5 | ... |
| ut | 0 | 1 | ... |
| sancti | 4 | 2 | ... |
| quod | 3 | 6 | ... |

Document-Term-Netzwerkgraphs



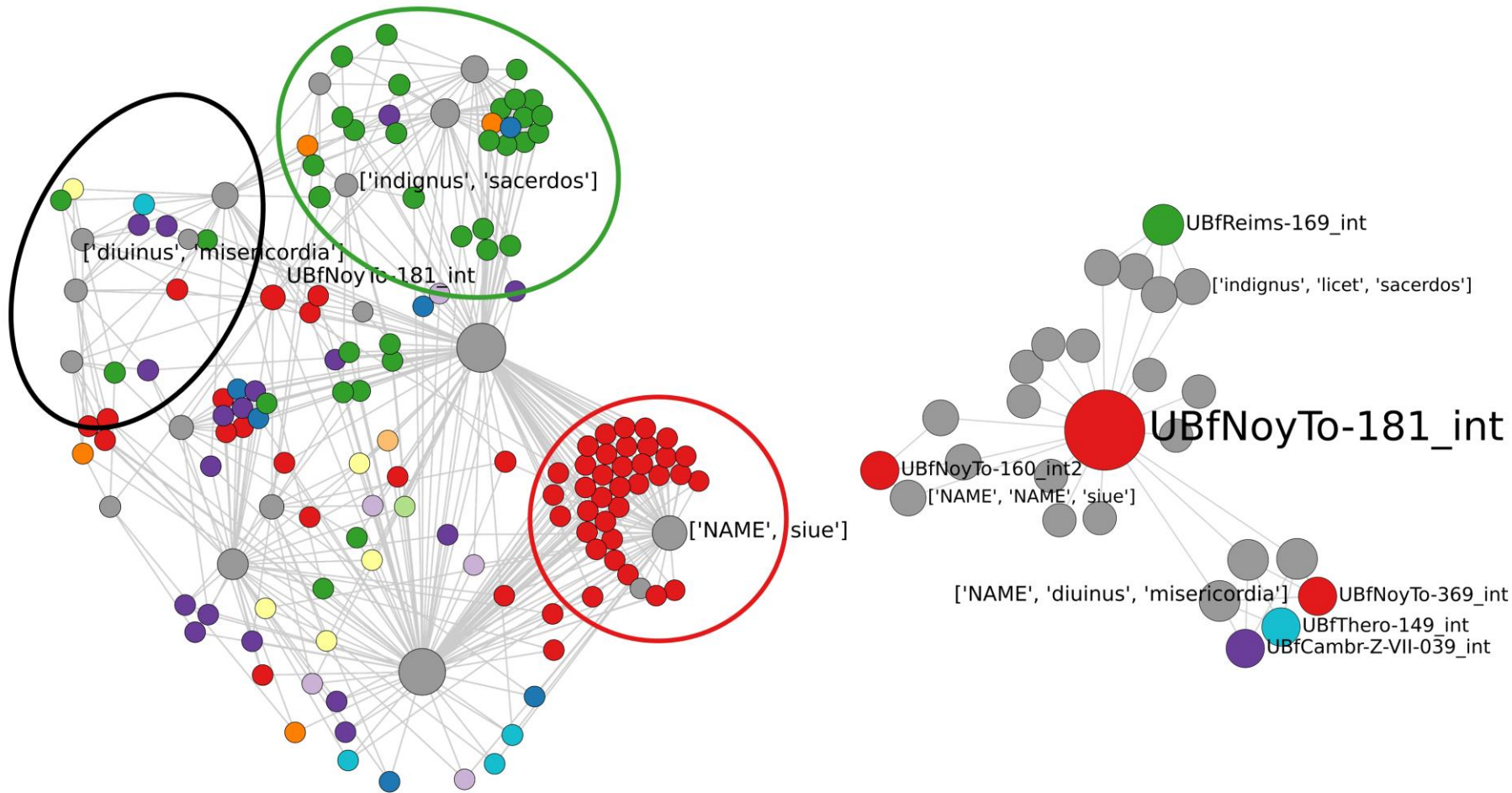
1000 MFW, TFIDF, ForceAtlas





A charter from Noyon looking like a charter from Reims

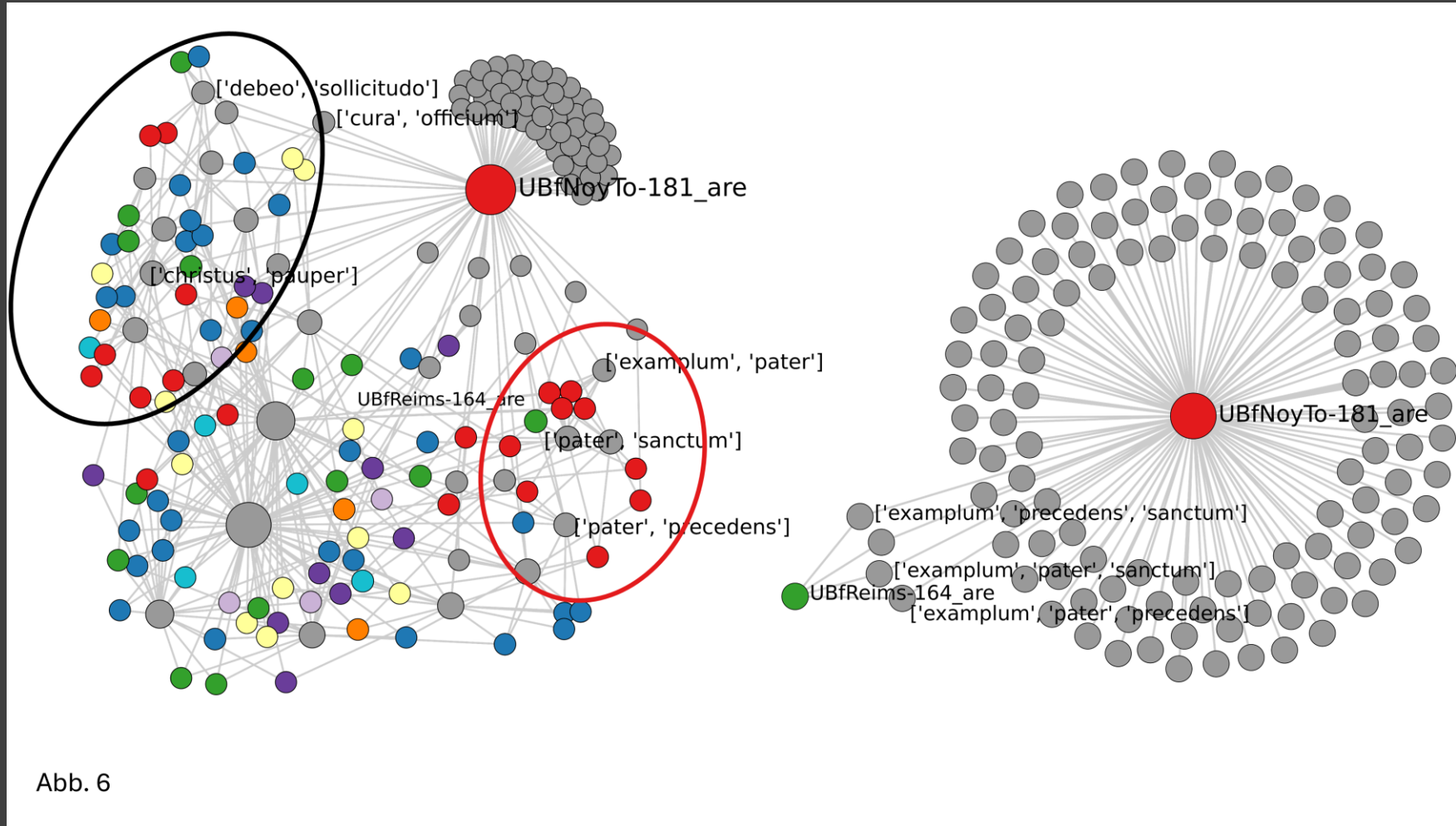
Ego networks of the *intitulatio* of Noyon 181: 2-skip-1-grams (left) and 3-skip-1-grams (right)



The cluster of green nodes indicates that *sacerdos licet indignus* is used mostly in Reims. *Name siue Name* is special to the double see of Noyon-Tournai. *Diuina misericordia* can not be located using this graphs.

Abb. 3

Ego networks of the *arenga* of Noyon 181: 2-skip-1-grams (left) and 3-skip-1-grams (right)



The cluster of red nodes indicates that *precedentium sanctorum Patrum exemplis* is used in Noyon-Tournai – and Reims 164. Expressions like *pauperes Christi* can be found in other charters as well.

Abb. 6

The Text of Noyon 181: Probable origin of individual text elements based on the interpretation of the network graphs.

Green: Reims, Red: Noyon-Tournai; Blue: more common formulas; Yellow: Papal charters.

(XPC) In nomine sanctę et individue Trinitatis, Patris et Filii et Spiritus Sancti. Amen (croix)

Ego LAMBERTUS, per divinam misericordiam Noviomensium seu Tornacensium sacerdos licet indignus, omnibus christianę religionis cultoribus imperpetuum. Precedentium sanctorum Patrum exemplis et vita edocti legisque divine assidua meditatione instructi debita sollicitudine curę sacerdotalis officii perurgemur pauperibus Christi ex beneficio humilitatis nostrę divinitus distributo karitativa manu quantumcumque porrigere et eorum necessitatibus ut devotiori pro nobis Deum exorent affectu imposterum providere.

Quocirca ęcclesię Beati Dyonisii secus urbem Remensem site ad sustentationem et gubernationem fratrum inibi sub regulari professione beati Augustini Deo servientium et postmodum imperpetuum servitutorum altare de Monteiscurt pro animę nostrę predecessorumque nostrorum remedio sub perpetua libertate tenendum iuste canoniceque contulimus, salvo tamen iure synodali episcopique et archidiaconi ac ministrorum suorum iusticia, eo quidem tenore interposito quod elemosinarum et beneficiorum predictę ęcclesię ego LAMBERTUS et RORIGO archidiaconus totaque Noviomensis ęcclesia participes erimus meique obitus anniversaria dies in eadem ęcclesia singulis annis celebrabitur.

Nequis autem contra hanc nostrę largitionis et confirmationis paginam ire presumat sub excommunicatione interdicimus.

Sed ut rata permaneat ac inconvulsa honestarum personarum testimoniis cum nostrę imaginis additamento eam muniri fecimus.

Signum domni Lamberti episcopi. [...]

Actum Noviomii, anno dominicę incarnationis millesimo C° XVI°, indictione VIII, regnante Ludovico glorioso Francorum rege, episcopante autem domno Lamberto. Ego Hugo cancellarius subscripsi.