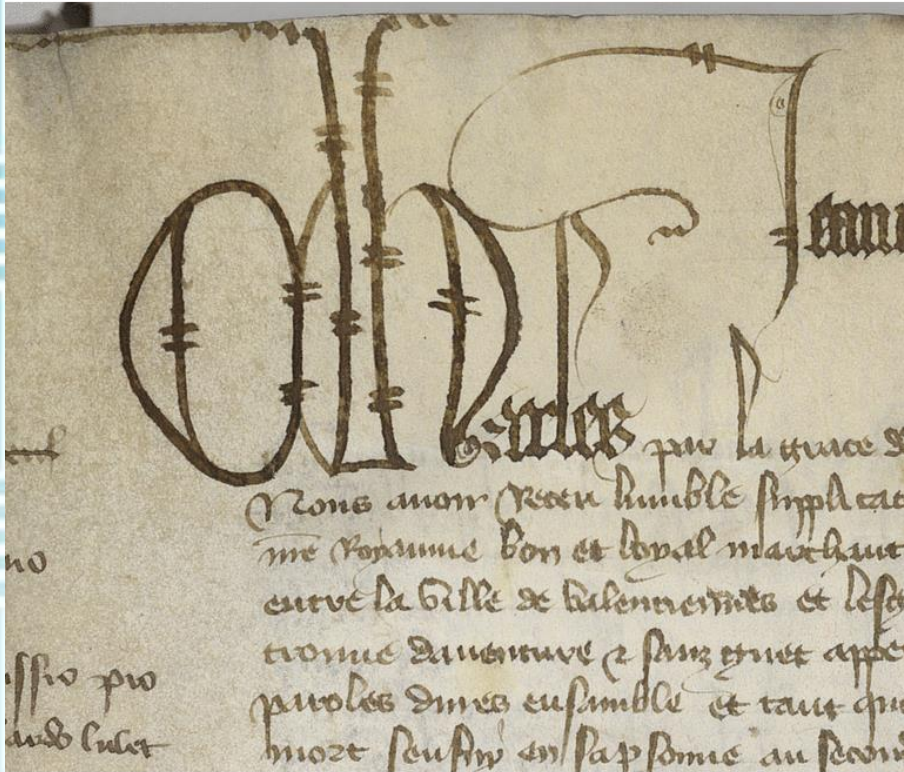


HIMANIS

Historical MANuscript Indexing for user- controlled Search



Dominique Stutzmann
Institut de Recherche et d'Histoire des Textes
@himanis6

<http://himanis.huma-num.fr/himanis/>

49e congrès de l'ADBU.

Tous bibl-IA-thécaires ? L'intelligence artificielle, vers un nouveau service public

Bordeaux, 18 sep. 2019

Introduction

- bit.ly/himanis
- Recherche plein texte sur docs manuscrits du Moyen Âge
- « Lecture distante », une pratique usuelle
 - Sources numériques natives (Twitter...)
 - OCR sur livres et journaux imprimés
 - « Distant reading », mais aussi « Distant viewing » et « Visual data analysis » en croissance

Plan

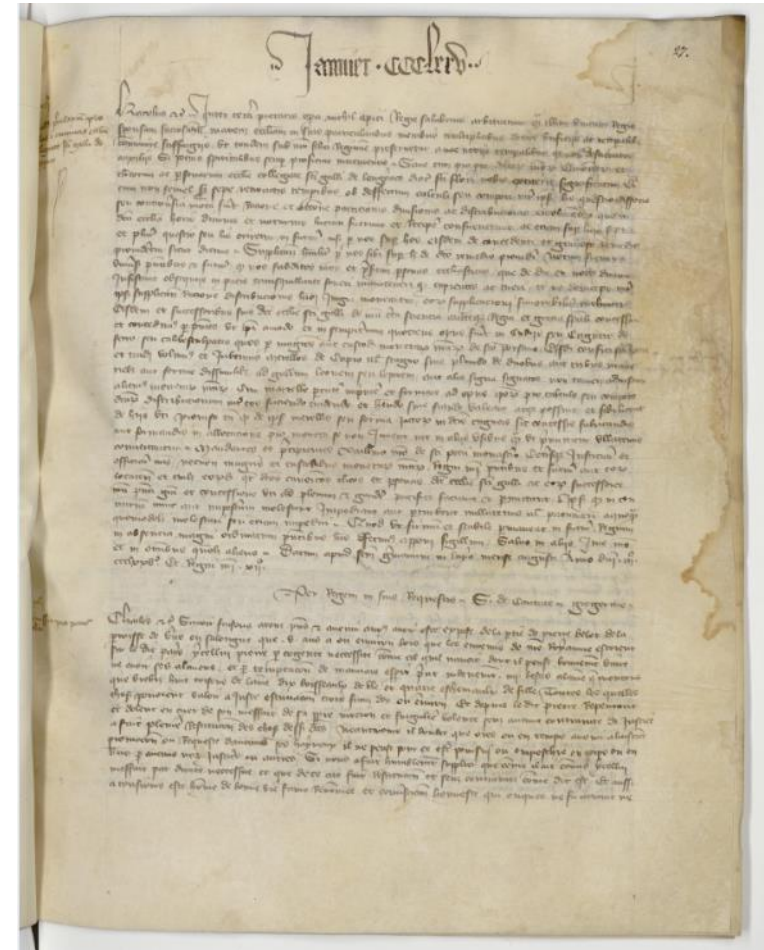
- **Projet HIMANIS**
 - Objectifs
 - Corpus
 - Partenariat transdisciplinaire
- **IA: mise en œuvre et résultats**
 - Métadonnées, HTR, « Indexation »
 - Accès plein texte à 80 000+ pages
 - Évaluation et crowdsourcing
- **Perspectives**
 - Environnements de recherche (utilisateur final)
 - Nouveaux projets de recherche (public/privé)
 - Nouveaux métiers, rôles, médiations

Projet HUMANIS



Objectifs

- **Indexation = mots-clefs**
 - Indexation “profonde” ou “deep indexing” plutôt que “transcription” des écritures médiévales
- **Moteur de recherche**
 - avec négociation de contenu
- **Reconnaissance de scripteurs**
- **Grande échelle (“big data” à l’échelle médiévale) :**
 - 200 volumes = 80 000 pages
 - Analyse de données massives



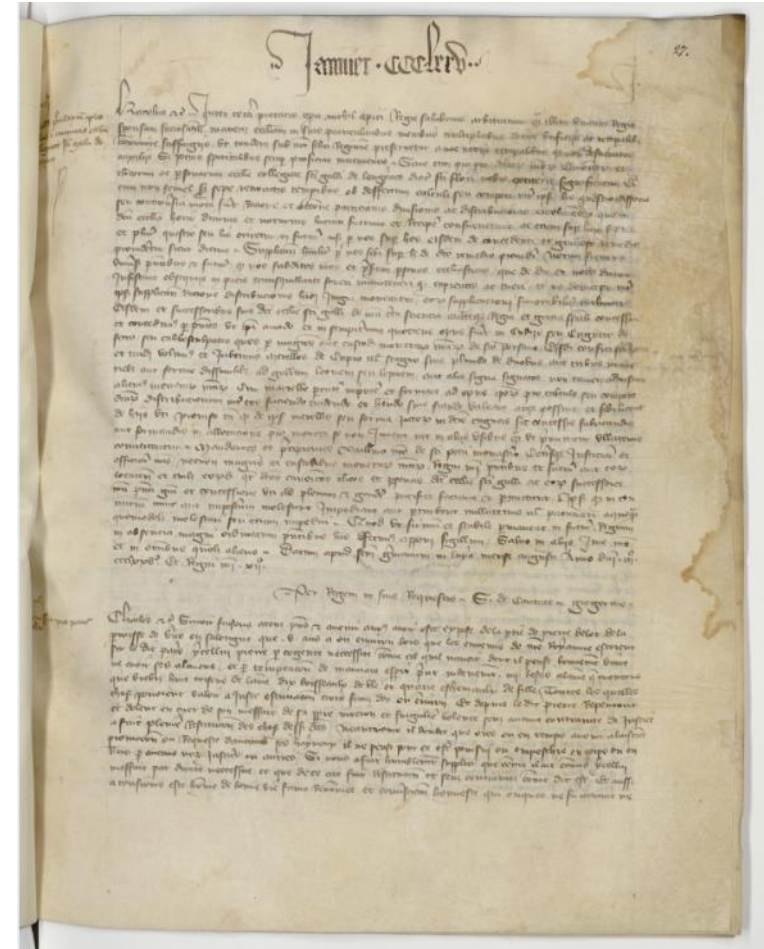
Objectifs

- **Corpus nodal**

- lié au pouvoir royal
- histoire européenne
- utilisé depuis le XVII^e siècle

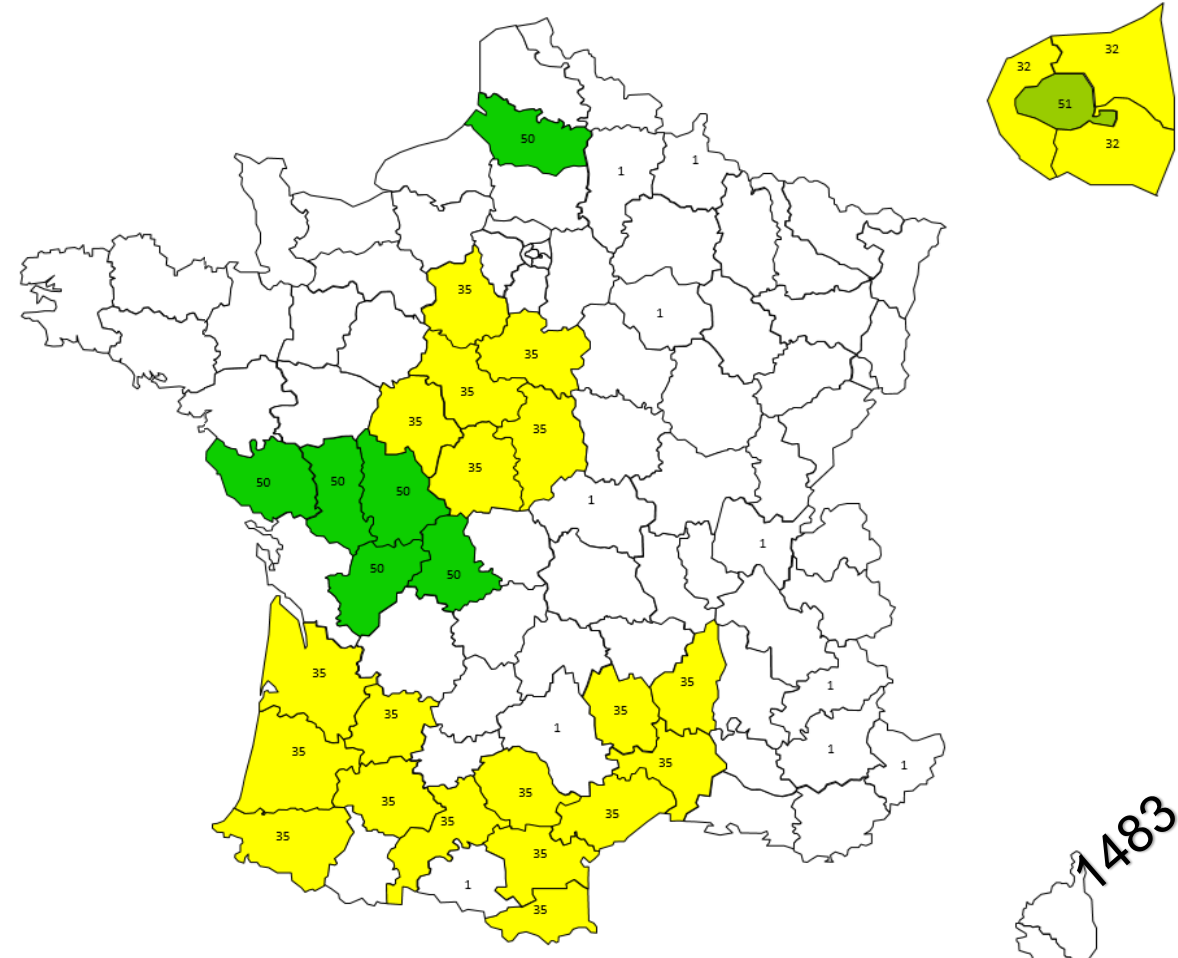
- Recherche historique

- source complexe
- fonctionnement de l'administration
- expansion du contrôle régalien
- modèles et influences



Corpus difficile d'accès

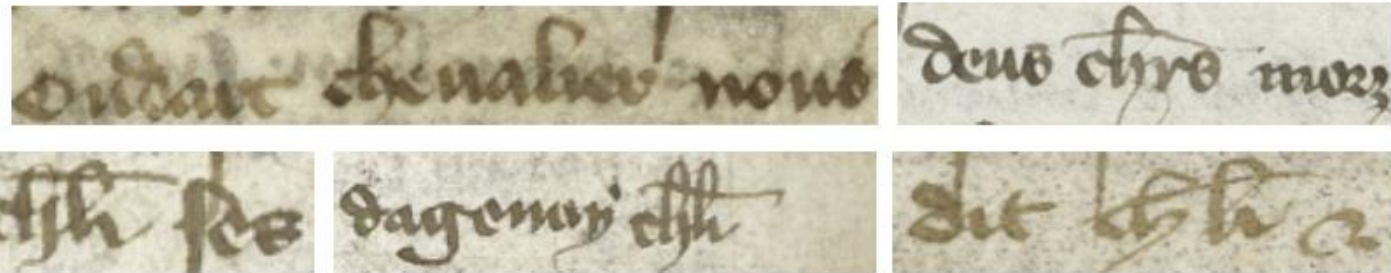
- **Taille**
 - 200 volumes, 80000 pages
- **Inventaires partiels**
 - qualité diverse
 - inaccessibles (manuscrits)
- **Métadonnées (résumés, index des noms et lieux)**
 - Chronologique (avant 1368)
 - Géographique
- **Éditions régionales**



Un accès difficile

- Abréviations (60 % des mots en latin, 30 % en français)
- Écritures cursives
- Nombreuses mains différentes

Chevalier



Guillaume

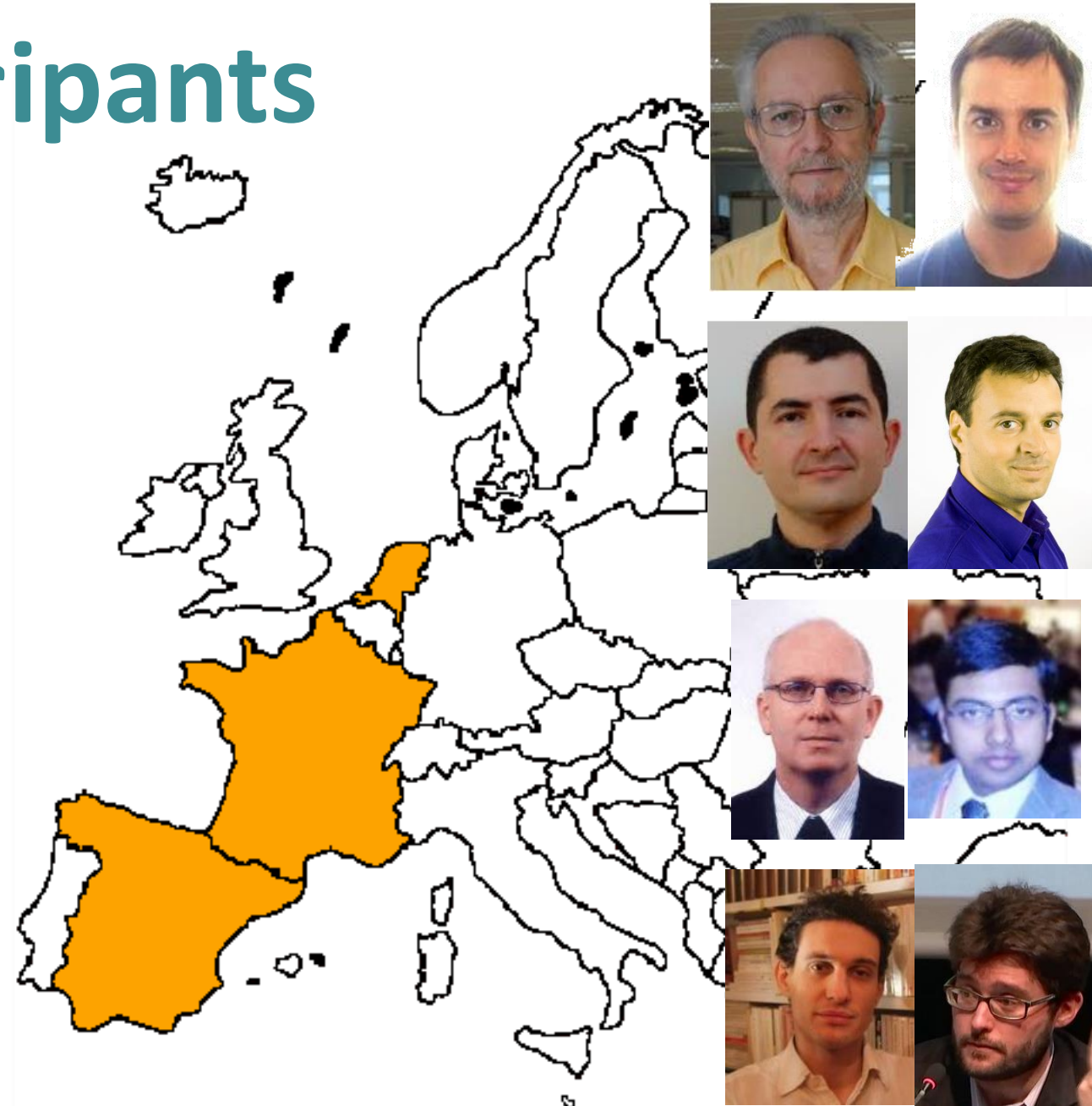


HIMANIS

- **JPI Cultural Heritage and Global Change “Cultural Heritage: a challenge for Europe”**
 - Initiative de programmation conjointe
 - HERITAGE PLUS Joint Call
- **HIMANIS = Historical MANuscript Indexing for user-controlled Search**
 - Appel en deux tours
 - Nov. 2015 – Avril 2018
 - Financement: 410 k€

Participants

- **Recherche transdisciplinaire**
 - Humanities | SHS
 - Computer Science | Informatique
 - Cultural Heritage | Patrimoine
- **Partenaires**
 - IRHT (CNRS) – France
 - Univ. Politècnica València – Spain
 - Univ. Groningen – Netherlands
 - A2iA – France
 - Teklia – France
 - Partenaires associés
 - Archives Nationales
 - Bibliothèque nationale de France



IA

mise en œuvre et résultats

Est-ce que ça marche ?



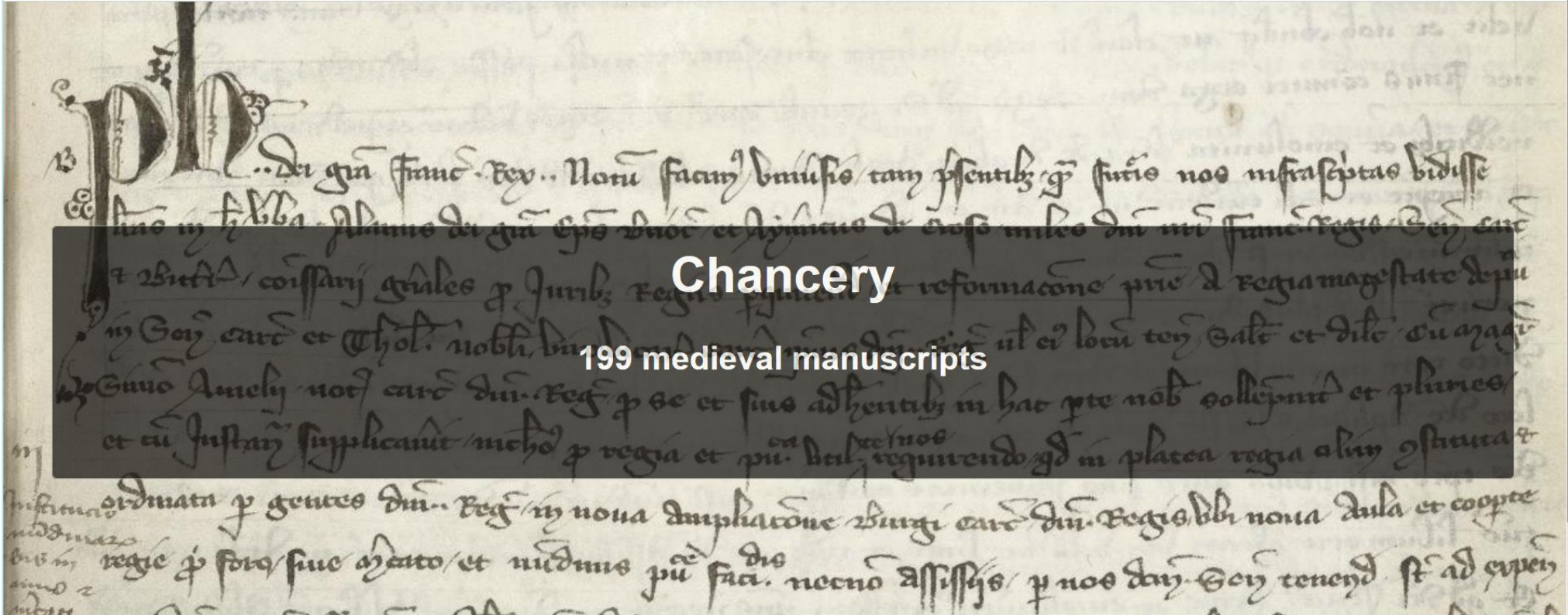
bit.ly/himanis

HIMANIS PRHLT search engine **BETA**

[Search options](#) [Indexing details](#)

Confidence: Max. results:

You are here: [home](#)



Chancery

199 medieval manuscripts

intelligere universaliter

Search

Search options

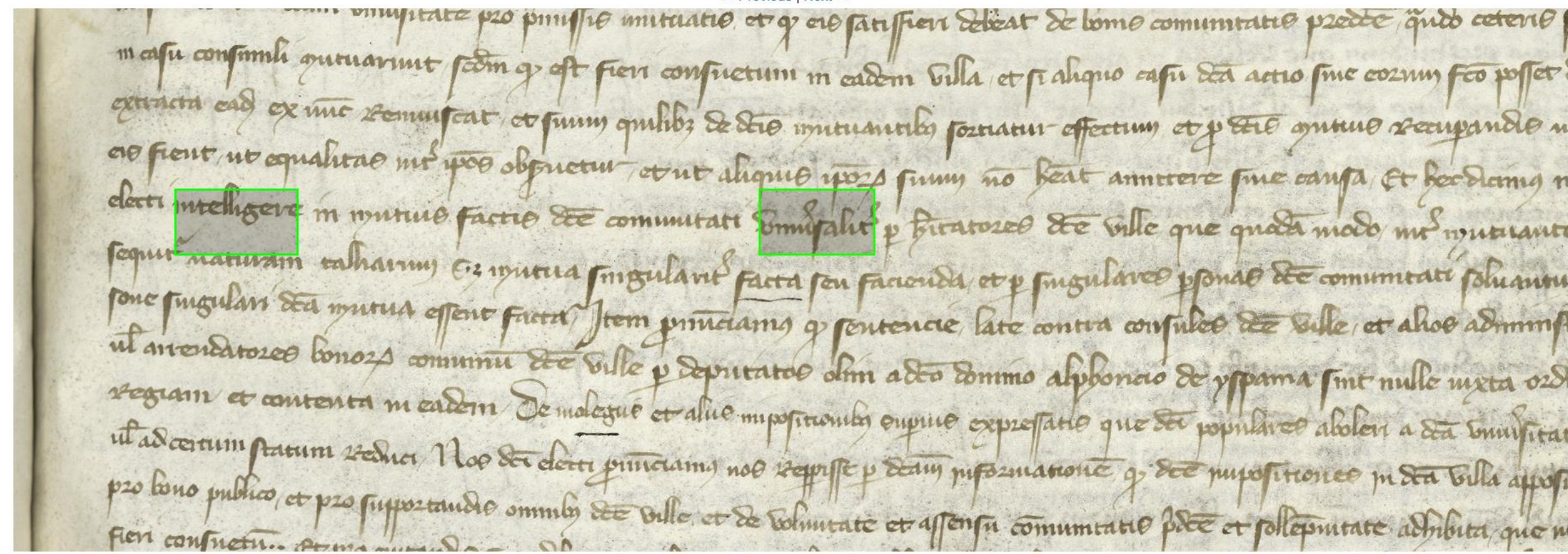
Indexing details

Confidence: Max. results:

You are here: [home](#) » [chancery](#) » [JJ066](#) » [page 441](#)

2 matches found for "intelligere universaliter" with an average confidence of 96.3 !

← Previous | Next →





allemaigne

Search

Search options

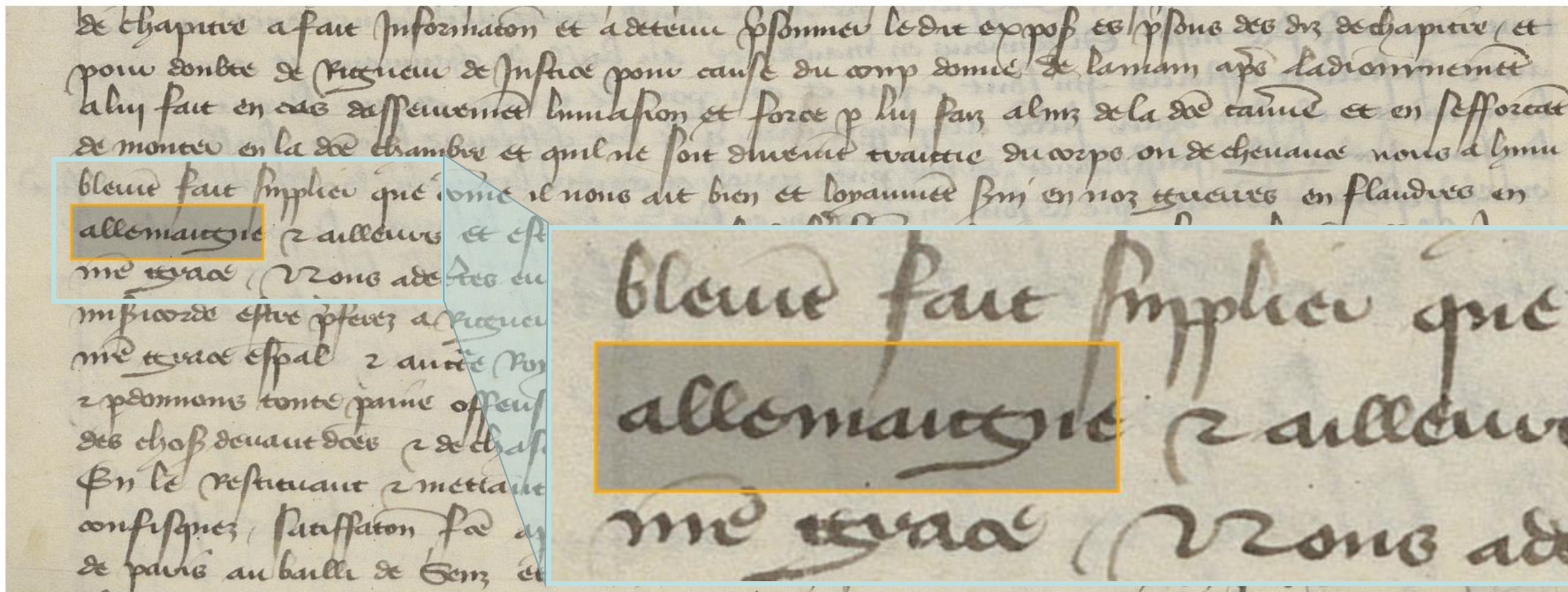
Indexing details

Confidence: Max. results:

You are here: [home](#) » [chancery](#) » [JJ138](#) » [page 91](#)

1 match found for "allemaigne" with an average confidence of 33.3 !

← Previous | Next →



et dunt a grant a la terre pmet de fanceur et au preu de
u chapure me dame de par et ensmant en la terre sur anche
me pmet de par et me de censeme prise sur deux
e parie deuant se ppose ten dme et p dme me
a ala maison en pend le nseigne ducenal blanc
me de l'eglise me dame de par a la maison qui fu
En chasite pouse pour en par et posside perpetuel
seman d'auy et mouenme l'eglise sur une lude
adde egle me dame de par pour et pour te
all. etc. etc. etc. etc. etc. etc. etc. etc. etc. etc. etc.

me dame de par

me dame de par

me dame de par

You are here: [home](#) » [chancery](#) » [JJ053](#) » [page 47](#)

19 matches found for "nobilium nostrum fecimus apponi sigillum actum parisius" with an average confidence of 71.6 !

[← Previous](#) | [Next →](#)

sua pñedi. Aut pñendi finacia. pñidem. Saluo malis ure nro. r mñbz alieno. Preterea domnu ipam. r pñona
 pñas. cu familia. homa. r belus caru mñia. pñedone suscepim. r gaudia spali. Quod ut ratu. r stabile pñder
 pñentibz. nñ fecim. appom sigillu. Actum par. Anno dñj. millo. ccc. sexto decimo. mense febr.
 p. d. Regem ad rect. dñor cardinal. r epì macdouen
 nobilitudo dñca. p. Band. (n. d. q. d. mñell. fuer.)
 Arnaldo de portali. p. Arnaldo de portali.
 Joh. Si gra pñac. r nauam rex. Dñco Arnaldo de portali. pñuenti mñ armor. eate. r dñt. pñona tuam ipius anq
 exigentibz honorare volentes. r pñsegu spali gra r fauore tibi r pñdicati tue recipiendi a quocunq. qñlice angulu
 qñlice qñ voluñs. r feoda seu alia queque nobilia. acquisita. iam. r adhuc. etiam. agñenda. tenendi. r possidea
 do pñcio more fruendi. nobilitum. r honore. pñent. r dñam. de ma. dñt. r regni. facultatem. qñdñm. pñ pñentes h
 as. quibz. ut hoc firmu. r stabile. pñder. nñ fecim. appom sigillu. Actum par. Anno dñj. q. ccc. sexto decimo
 mense febr.
 p. d. Regem ad rect. dñj. p. d. Arnald. cardinal.
 2 Joy

Lemmatisation

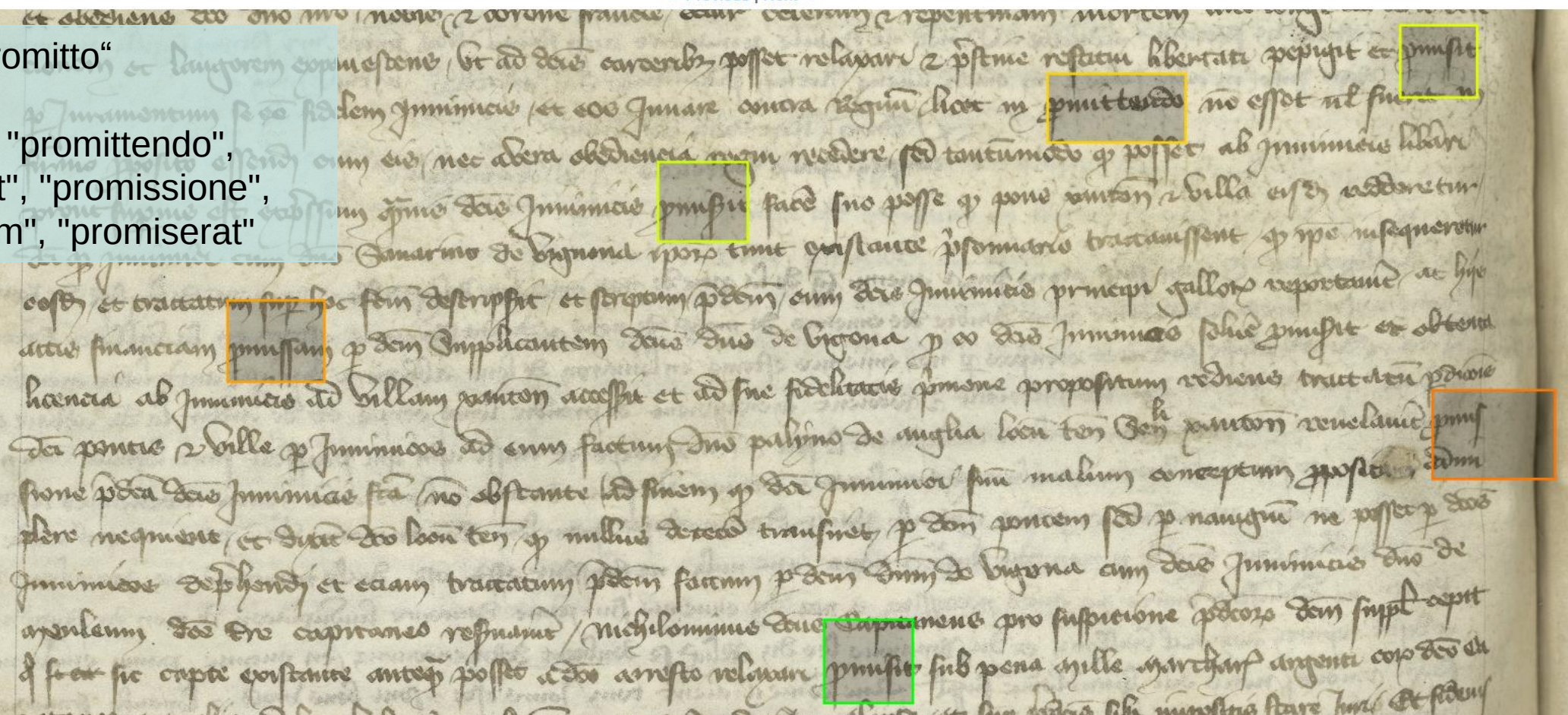
You are here: [home](#) » [guerin.lemmas](#) » [guerin.lemmas.chap1](#) » page 435
7 matches found for "promitto" with an average confidence of 48.6 !

← Previous | Next →

Suche: „promitto“

Treffer:

"promisit", "promittendo",
"promiserit", "promissione",
"promissam", "promiserat"



Processus de reconnaissance

- **La reconnaissance utilise des réseaux de neurones**
 - Entraînement au niveau ligne
 - Dépasser la problématique des abréviations
- **Entraînement à plusieurs étapes**
 - Formaliser un terrain :
 - utiliser les éditions existantes
 - TEI pour les inventaires et pour les éditions
 - Segmenter la page
 - Aligner texte et image
 - Entraîner le « segmenteur » et le « reconnaisseur »
 - Production de « latices »

Indexation profonde (Deep Indexing)

- CNN/RNN (Convolutional/Recurrent Neural Networks)

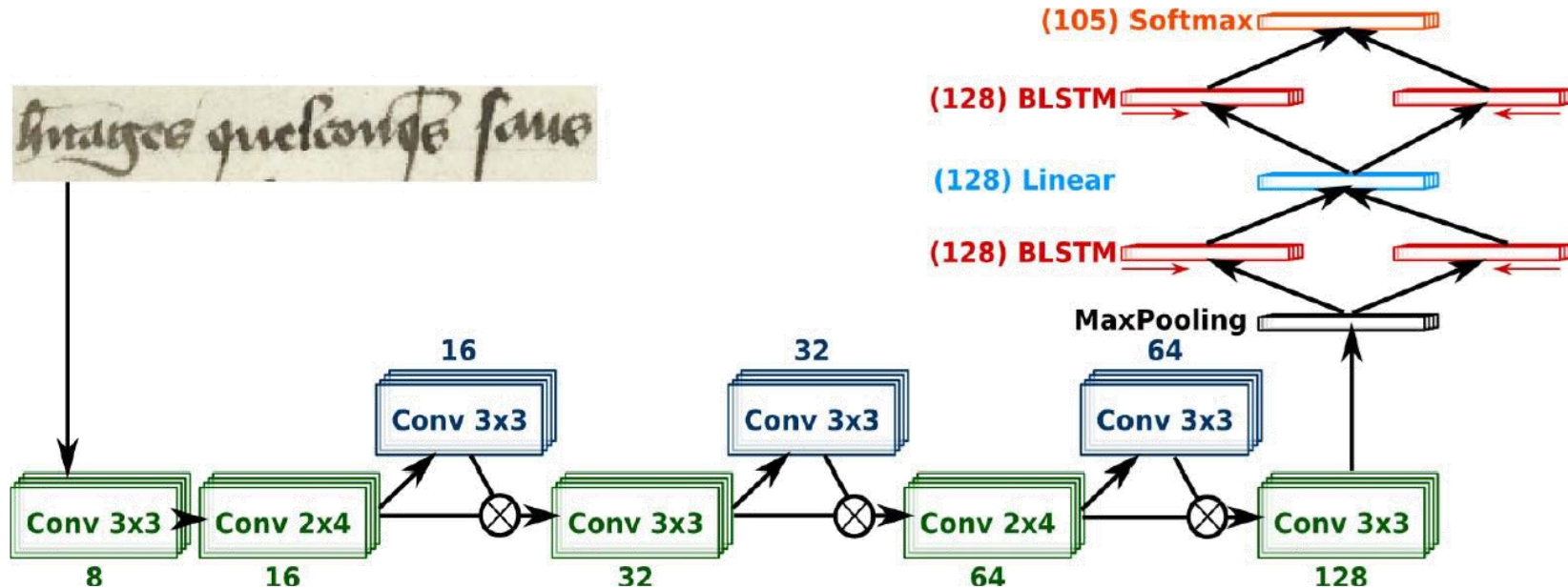
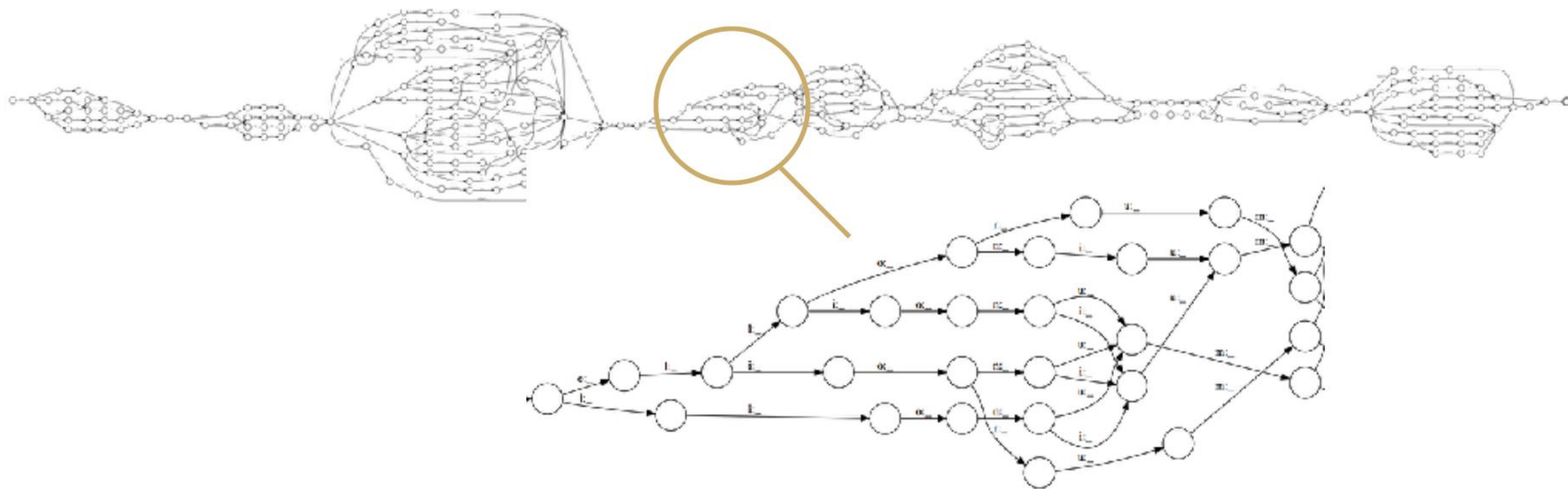


Fig. 1. Neural network architecture.

Indexation profonde (Deep Indexing)



© C. Kermorvant

You are here: [home](#) » [chancery](#) » JJ026 » page 174

Sorry, the query "noel" does not have results in this level...

[← Previous](#) | [Next →](#)

Provide the word ✕

Word:

Options:

- cancellarii
- cancellaria
- cancellaria**
- cancellania
- cancellario
- cancellari
- tancellarii
- cancellan
- cancellata
- tancellara
- gancellarii
- tancellaria
- concclarii
- gancellara
- tancellania
- concclara
- tancellario
- ancclarii
- gancellaria
- concclaria

us. filij Radulphi de Gen
ad ut perpetue stabi
presentem paginam
et Regij nominis ka
ato fecimus comunni.
anno incarnationis dnice
no sexto. mense octobri.
incefimo. Asparibus in
supposita sunt et signa.
rephani buticularij.
Johannis Camerarij. Constabulario nullo.

Sancti vacan

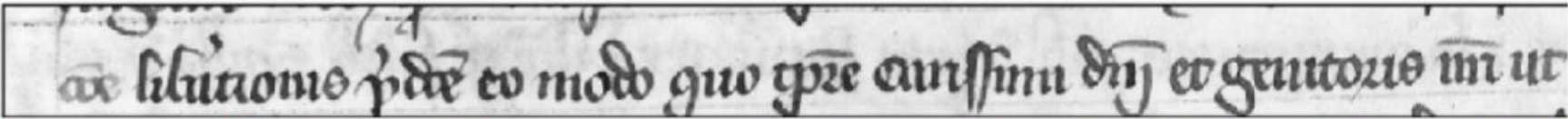
Cancellaria.

Indexation profonde (Deep Indexing)

- **Décoder et optimiser l'indexation**
 - Modèles linguistiques
 - N-grams (5-grams)
 - Dictionnaires
 - Lemmatisation

Modèle optique + modèle de langue :

GT : casione liberationis predicte, eo modo quo tempore carissimi Domini et genitoris nostri, ut



RNN: colme liblrationis predicte| eo modo quo temple carissim| Domini et genitoris nostri, ut
+LM: comme liberacionis predicte| eo modo quo tempore carissimi Domini et genitoris nostri, ut

Évaluer

	Transcription available (ground truth)		Transcription not available
	Training	Evaluation	
Time Span	1302-1361		1302-1483
# Acts	341	95	68 000
# Lines	6 061	1 733	2 692 339*
# Words	117 709	33 097	39 623 581*

* estimated

Character Error Rate: 18%

Évaluer

- Évaluation quantitative

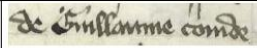
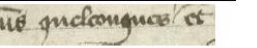
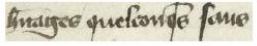
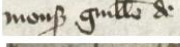
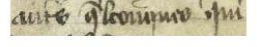
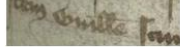
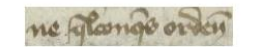
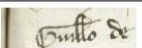
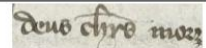
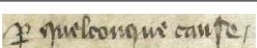
Keyword	Guillaume	chevalier	livres	quelconques
AP	0.79	0.89	0.79	0.91
$t: d, h, m, f$	0.45: 25,22,6,3	0.45: 39,34,3,5	0.20: 77,60,14,17	0.35: 11,10,1,1
$\pi/\rho @ t$	0.88 / 0.69	0.87 / 0.92	0.78 / 0.81	0.91 / 0.91
Full form				
Abbreviated				
				
				
False Positives				

Fig. 6. Examples of modernized (expanded) keyword queries and corresponding spotting results. For each query: AP is the Average Precision, t is the confidence threshold used for the example images below, d is the number of detected spots, h is the number of correct hits, m is the number of missing spots, and f is the number of false positives, all at the threshold t . The precision and recall obtained from these values are π and ρ , respectively. Selected examples of correct spotted images, both in full form and abbreviated, and examples of false positives, are shown for each query.

- Évaluation qualitative
 - Études historiques antérieures
 - Mots et thématiques de test
 - Évaluation par les utilisateurs

Perspectives

Recherche et nouveaux projets
Environnements de recherche
Métiers



Des possibilités infinies

← → ↺ ↻ <https://himanis.hypotheses.org/254> 80 %

Himanis ABOUT JPI CULTURAL HERITAGE

VISIT OUR NEW HIMANIS WEBSITES

HIMANIS website

THE HIMANIS PROJECT

NEW NEW NEW : Search the plain text of the Chancery registers and give your feedback on the PRHLT HIMANIS search interface

Search ...

The blog HIMANIS is devoted to image analysis and Hand-written text recognition for medievalists. Basing on a re-research program conducted on the registers of the French Royal Chancery in the 14th and 15th centuries, it covers several fields from the Humanities (paleography, diplomatics, history of the governments) and Computer Science. The re-research is funded by the European Commission through "Heritage Plus - JPI on Cultural Heritage and global change".

Le carnet de recherche HIMANIS est consacré à l'analyse d'image et la reconnaissance d'écriture manuscrite pour les

NEWS

MORE MEDIEVAL MUSIC (AN JJ 87 – JJ 99) / ENCORE PLUS DE MUSIQUE MÉDIÉVALE : LES MÉNESTRELS DES REGISTRES DE LA CHANCELLERIE AN JJ 87 – JJ 99

04/01/2018 DOMINIQUE STUTZMANN LEAVE A COMMENT



Dijon BM ms 493, fol. 15v (bvmn.irht.cnrs.fr/consult/consult.php?VUE_ID=1252786)

Voici la suite de nos ménestrels, en passant à des registres non dépouillés (retrouver [l'épisode 1 ici](#)). Au passage, rappelons que [Menestrel](#) est un site de ressource capital pour les médiévistes.

- Musique et musiciens
- Bâtardise
- Sorcellerie
- Marine et amirauté
- Vocabulaire diplomatique
- ...

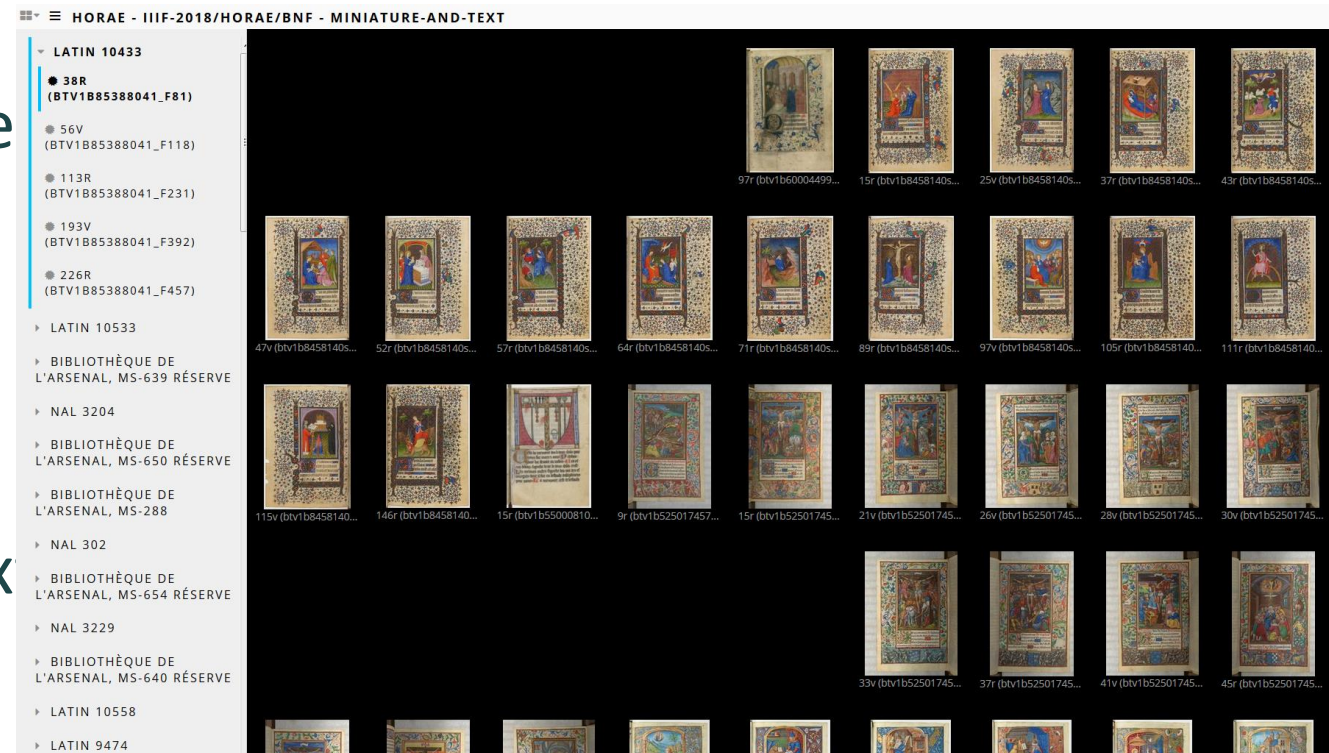
Futurs projets à l'IRHT

- **HOME**

- Projet européen (2019-2021)
- 2400 cartulaires
- Environnement de recherche

- **HORAE**

- Projet ANR (2018-2021)
- 1000 livres d'heures
- Identifier et comparer les tex



IIIF et (méta)données ouvertes

The screenshot displays the Paris Archives nationales IIIF viewer interface. The top navigation bar includes a hamburger menu, the title "PARIS, Archives nationales, JJ 040 (Reproduction intégrale)", and links for "Changer la disposition" and "Plein écran". Below the navigation bar, there are tabs for "Index", "Anno.", and "Rechercher", along with a "Calques" button. The left sidebar shows a tree view of the manuscript's structure, including "PARIS, Archives nationales, JJ 040 (Reproduction intégrale)" and a list of folios from "plat supérieur" to "003r". The main area displays a high-resolution image of a medieval manuscript page in Gothic script. The text is overlaid with a grid of blue rectangular frames, representing IIIF frames. A white tooltip box with the text "TENE BUNTUR" is visible over one of the frames. The bottom right corner of the image shows some UI controls for zooming and panning.

Interfaces flexibles

← → ↺ 🏠 ⓘ 🔒 https://arkindex.teklia.com/pages?q=gratia&corpus=&score=0.7 90 % ... ⚙️ ☆ 🔍 Rechercher 📖 📄 📅 📌 📎 📏 📐 📑 📔 📕 📖 📗 📙 📚 📛 📜 📝 📞 📟 📠 📡 📢 📣 📤 📥 📦 📧 📨 📩 📪 📫 📬 📭 📮 📯 📰 📱 📲 📳 📴 📵 📶 📷 📸 📹 📺 📻 📼 📽 📾 📿 📰 📱 📲 📳 📴 📵 📶 📷 📸 📹 📺 📻 📼 📽 📾 📿

ArkIndex Browse Search ▾

Search in pages

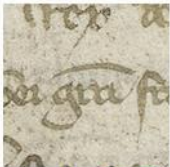
All corpuses ▾ gratia 🔍 SEARCH

Confidence level 70 %

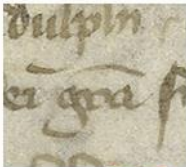
Page 1 : Showing items 1 to 20 out of 76 results

PREVIOUS NEXT

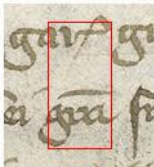
Démo Publique / JJ040 / JJ040 / Page no. 78, recto • 4 transcriptions • [View page](#)



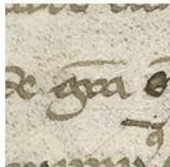
GRATIA 98%



GRATIA 96%



GRATIA 91%



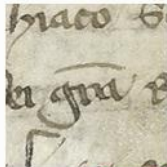
GRATIA 85%

2d4222b7-e063-432f-b7d2-9e3631634024

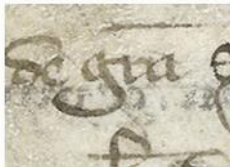
Démo Publique / JJ040 / JJ040 / Page no. 62, recto • 4 transcriptions • [View page](#)



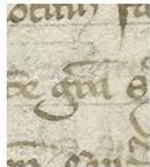
GRATIA 99%



GRATIA 91%



GRATIA 87%



GRATIA 75%

 Teklia

PREVIOUS

NEXT

Act 1

Act 2

Act 3

Act 4

Act 5

Act 6

Act 7

Act 8

Act 9

Act 10

Act 11

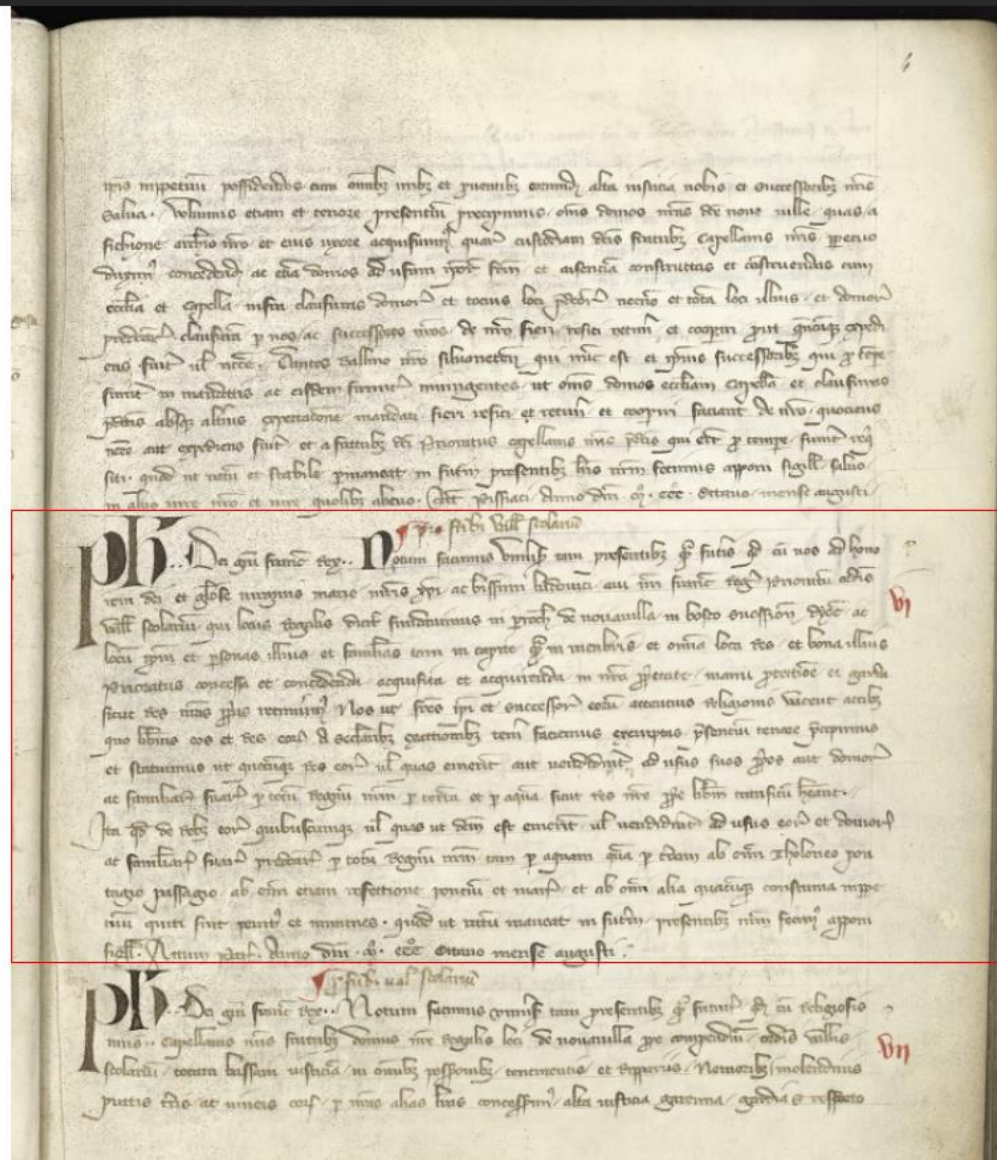
Act 12

Act 13

Act 14

Act 15

Act 16



Metadata

date 1308-08

folio 4

id JJ22840

language lat.

location Paris

name 6

summary

html

Concession de libre passage et exemption de droits de transit dans tout le royaume accordées, pour leur usage, aux religieux du prieuré de Royallieu.

persons

text

- Paris, actes passés à

places

text

- Royallieu-lès-Compiègne, Regalis Locus (prieuré du Val-des-Écoliers de) [Oise, con et cne Compiègne], prieur et religieux, chapelains du roi

subjects

text

Évaluation par les utilisateurs

The screenshot displays the HIMANIS web application interface. At the top, there is a navigation bar with the HIMANIS logo and a 'BETA' badge. Below the navigation bar, a search bar contains the text 'pelerinage || pelerinages'. To the right of the search bar, there is a 'Search' button and a 'Confidence' slider. Below the search bar, a message states: 'You are here: [home](#) » [chancery](#) » [JJ098](#) » [page 440](#)'. Below this message, it says: '2 matches found for "pelerinage || pelerinages" with an average confidence of 39.3 !'. Below the message, there are links for 'Previous' and 'Next'. The main content area shows a snippet of a handwritten document. A yellow box highlights the word 'pelerinage' in the text. A confirmation dialog box is overlaid on the document, asking 'Was this word correctly spotted?' and 'Please confirm:'. The dialog box has three buttons: 'Correct', 'Incorrect', and 'Not sure'. At the bottom of the page, there is a footer with the text: '© 2017 HIMANIS (see also this and this) & PRHLT, UPV' and 'XHTML CSS AA Browse Happy'. The browser's taskbar is visible at the bottom, showing various application icons and the system clock.

Programa Libri e colo Mirador Vi Mirador Vi Tra ostentazio Università degi Università degi La Bulle du (12 218 no Himanis - C x Bitly | URL Himanis | Himanis - Char

prhlt-kws.prhlt.upv.es/himanis/index.php/ui/show/chancery/195/440?q=pelerinage+||+pelerinages&feedback=1

Rechercher

HIMANIS BETA

pelerinage || pelerinages

Need help?

You are here: [home](#) » [chancery](#) » [JJ098](#) » [page 440](#)

2 matches found for "pelerinage || pelerinages" with an average confidence of 39.3 !

← Previous | Next →

Was this word correctly spotted?

Please confirm:

Correct Incorrect Not sure

© 2017 HIMANIS (see also this and this) & PRHLT, UPV

XHTML CSS AA Browse Happy

FR 23:07 07/09/2017

Créer du sens

- **Créer du sens**
 - → Redéfinir le rôle de l'édition (critique) dans la recherche
 - → Des données et des standards ouverts
- **Nouveau cycle de métadonnées**
- **Nouveaux métiers**
 - La lecture (pas seulement distante)
 - Qui va exploiter cette information ?
 - Comment évaluer les discours tenus par les humains à partir de données de l'IA
 - Métiers de l'historien et des professionnels de l'information

Nouvelles médiations, nouveaux métiers

- **Connaître les corpus rendus disponibles par l'IA**
 - ... et ne pas oublier les autres
- **Un cercle vertueux, si...**
 - HTR > traitements linguistiques > analyse de données, traduction automatique
 - Mise en œuvre d'une stratégie de qualité (mesure, vérification, aide à la décision)

Nouvelles médiations, nouveaux métiers

- **Transformations (= crainte de la suppression de métiers)**
 - IA se nourrit de l'intelligence humaine
 - IA accélère la production de données / connaissances
 - Métiers de conception, médiation et réflexion
 - Transdisciplinaire – contractualisation public-privé
- **Source de financement / Responsabilités / Temporalités**
 - Standards (IIIF, TEI)
 - Infrastructure (Huma-Num, BVMM: pérennisation et permanence de l'accès)
 - Projet de recherche (Himanis)
 - « Services numériques innovants » (médiations)
- **Rôles et stratégies**
 - Consultance : technologies émergentes, identifier les actions pertinentes
 - Prestation de service / Recherche partagée

