

Techniques for the creation, enrichment and analysis of digital texts

David Joseph Wrisley, Assoc Prof of Digital Humanities, NYU AD,
@DJWrisley, NYU Buenos Aires, 22 March 2018

جامعة نيويورك أبوظبي



NYU | ABU DHABI

Useful links for my talk

<https://tinyurl.com/tallerhdNYUBA>

Outline

- PART ONE
 - What are open data in general (for humanities/medieval studies in particular)?
 - What is a corpus? What is an open corpus?
 - What forms of corpus-based research have emerged in modern literary studies? What forms are emerging in medieval studies?
- PART TWO
 - Historical (& current) visions of a “body” of medieval French
 - The vision of OpenMedFr & open knowledge theory
 - Present workflow
 - Participating in the project
 - A short demo video (8 mins)

Open data

Definición de Conocimiento Abierto

Versión: 2.1

La Definición de Conocimiento Abierto aporta precisión al significado del término «abierto» (open) cuando se aplica al conocimiento, promueve un procomún robusto en el que cualquiera puede participar y maximiza su interoperabilidad.

Resumen: *El conocimiento es abierto si cualquiera es libre de acceder a él, usarlo, modificarlo y compartirlo, estando sujeto a lo sumo a medidas que preserven su autoría y su apertura.*

Este significado esencial se corresponde con el de «abierto» (open) cuando se refiere al software, como en la [Open Source Definition](#), y es un sinónimo de «libre» (free) tal y como se usa en la [Definición de Software Libre](#) y en la [Definición de Obras Culturales libres](#).

El término **obra** se utilizará para denominar el ítem o porción de conocimiento a transferir.

El término **licencia** se refiere a las condiciones legales bajo las cuales está disponible la obra.

La expresión **dominio público** denota la ausencia de derechos de autor (copyright) y de restricciones similares, ya sea por defecto o por la renuncia a dichas condiciones.

<http://opendefinition.org/od/2.1/es/>

¿Qué son los datos abiertos?

Los **datos (oficiales) abiertos** —también denominados «información del sector público» y «Open Data»— son aquellos recopilados, generados o financiados por organismos públicos y que pueden reutilizarse para cualquier fin sin coste alguno. La licencia establece las condiciones de uso. Los principios por los que se rigen los datos abiertos se explican detalladamente en la página [Open Definition](#) .

Se entiende por **información del sector público** aquella que pertenece a las administraciones públicas. La Directiva relativa a la reutilización de la información del sector público ofrece un marco legal unificado para el establecimiento del mercado europeo de datos en poder de las instituciones públicas. Este instrumento se asienta sobre los pilares básicos del mercado interior: la libre circulación de datos, la transparencia y la competencia leal. Conviene puntualizar que no todos los contenidos del sector público son datos abiertos.

<https://www.europeandataportal.eu/es/what-we-do/our-activities>

@DJWrisley NYU Buenos Aires 22-23 March, 4

Open data in general (governments)

- Agriculture
- Transport
- Environment
- Education
- Health
- Geography
- Cities
- Culture
- ...



Ministerio de Modernización
Presidencia de la Nación

Datasets Organizaciones APIs Acerca -

Datos Argentina

624
CONJUNTOS DE DATOS

022
ORGANIZACIONES
CON DATOS

¿Qué datos buscás?

¡Ponemos a tu alcance datos públicos en formatos abiertos para que puedas usarlos, modificarlos y compartirlos. **Estos datos son tuyos.** Podés crear visualizaciones, aplicaciones y grandes herramientas con ellos.

Agroganadería, pesca y forestación | Asuntos internacionales | Ciencia y tecnología | Economía y finanzas | Educación, cultura y deportes | Energía | Gobierno y sector público | Justicia, seguridad y legales | Medio ambiente | Población y sociedad | Regiones y ciudades | Salud



Newsletter | FAQ | Search | Contact | Cookies | Legal notice | Login | English (en)

EUROPEAN DATA PORTAL

Search site content...

European Data Portal

What we do - Data - Providing Data - Using Data - Resources -

Search Datasets

Enter keywords... Search Q

SPARQL Search

Open Data and the Knowledge Society (2017) - OKN's definition

- **Availability and access:** data available as a whole and in a convenient and modifiable form.
- **Reuse and redistribution:** data must be provided under terms that permit reuse and redistribution including the intermixing with other datasets. The data must be machine-readable.
- **Universal participation:** everyone must be able to use, reuse and redistribute — there should be no discrimination against fields of endeavour or against persons or groups. For example, ‘non-commercial’ restrictions that would prevent ‘commercial’ use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

Pause

How might these ideas of Open Data and Open Knowledge be related to our research, our intellectual work?

Examples of Open Data in the humanities



D P L A DIGITAL PUBLIC LIBRARY OF AMERICA

Discover 20,961,350 images, texts, videos, and sounds from across the United States

Search the collection Search



Add a search term Browse

Explore 51,332,004 artworks, artefacts, books, videos and sounds from across Europe

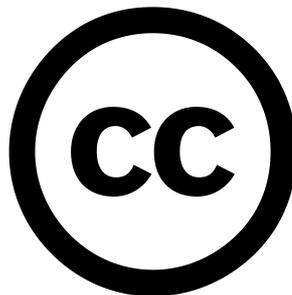


HATHI TRUST Digital Library

Search HathiTrust's digital library

Search words about or within the items

[Advanced full-text search](#) | [Search tips](#)



Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more.



Selected examples in open medieval studies

Digitized manuscripts (Parker on the Web (manuscripts))

Harvard Digital Atlas

Pleiades Gazetteer

OPenn (high-res archival images)

Open edition of *Speculum*

Open textual data

- Hispanic Seminary Digital Library (Old Iberian transcriptions)

- Princeton Charrette project (multi witness transcriptions)

- Exploring Place in the French of Italy (geodata)

- Archivio della latinità italiana del medioevo (Latin plain text)

- Open Medieval French



Outline

- PART ONE
 - What are open data (for humanities, and medieval studies in particular)?
 - **What is a corpus? What is an open corpus?**
 - What forms of corpus-based research have emerged in modern literary studies? What forms are emerging in medieval studies?
- PART TWO
 - Historical (& current) visions of a “body” of medieval French
 - The vision of OpenMedFr & open knowledge theory
 - Present workflow
 - Participating in the project
 - A short demo video (8 mins)

Aspects of corpus analysis (Biber, Conrad, & Reppen 1998)

1. it is empirical, analyzing the actual patterns of language use in natural texts (*--> it is a generally more of a surface than a symptomatic reading*);
2. it utilizes a *large and principled collection* of natural texts, known as a “corpus,” as the basis for analysis;
3. it makes extensive use of computers for analysis, using both automatic and interactive techniques;
4. it depends on both quantitative and qualitative analytical techniques.

(italics mine)

NB: A corpus is not a canon

- A canon:
 - works of a particular author or artist recognized as genuine
 - A list of works considered to be of the highest quality of all texts
- A corpus:
 - A corpus is a sample, *designed* to represent a textual domain in a language.
 - A digital corpus is a collection of “machine-readable” texts that enables different forms of computer-based manipulation and analysis.
 - A corpus will contain some canonical works, but it is not designed as a digital container for a canon.
 - A corpus will be sustained as long as a community works to grow it. It is not a fixed body.

A corpus search for “natur*” in medieval French

AntConc 3.4.3m (Macintosh OS X) 2014

Corpus Files

Concordanc | Concordance Plo | File View | Clusters/N-Gram | Collocate | Word Lis | Keyword Lis

Concordance Hits 169

Hit	Kwic	File
46	on droit naturel mouvement, et c'est pour la maniere du	EdeConty_Jeu
47	dire un naturel mouvement ou une inclination qui nous e	EdeConty_Jeu
48	le seul mouvement de nature sanz doctrine estroit. Pou	EdeConty_Jeu
49	par le naturel mouvement dessusdit, toute raison hors	EdeConty_Jeu
50	de son naturel mouvement et est trebles a ly. Et de ce	EdeConty_Jeu
51	ment le mouvement naturel de la mer qui continuelment f	EdeConty_Jeu
52	ve toute nature a a mouvement et a mutacion. Et pour ce	EdeConty_Jeu
53	rance de mouvement a aussi de nature ignorance, pource q	EdeConty_Jeu
54	oses de nature ont regard a mouvement ou a mutacion, el	EdeConty_Jeu
55	ordre de nature pource que tout naturel mouvement a aucu	EdeConty_Jeu
56	acion ou mouvement de chaleur naturel ne se fait pas ir	EdeConty_Jeu
57	de tel naturel mouvement TG fait ainsy sans raison s'e	EdeConty_Jeu
58	ment par nature mouvoir, car par son mouvement est il ca	Evrart_Harmc
59	selon le naturel mouvement de chascune. F c nous y pouc	Evrart_Harmc
60	dire un naturel mouvement ou une inclination qui nous e	Evrart_Harmc
61	le seul mouvement de nature sans doctrine eslroit. e	Evrart_Harmc

Search Term Words Case Regex

Search Window Size 50

Kwic Sort Level 1 1R Level 2 2R Level 3 3R

Total No. 436
Files Processed

Collocations in context of a private
Corpus of 14/15th court writers

AntConc 3.4.3m (Macintosh OS X) 2014

Corpus Files

Concordanc | Concordance Plo | File View | Clusters/N-Gram | Collocate | Word Lis | Keyword Lis

Total No. of Collocate Types: 307
Total No. of Collocate Tokens: 580

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
30	3	0	3	7.62081	ou
31	3	3	0	22.27258	naturaliter
32	3	0	3	10.62146	monde
33	3	3	0	8.69319	me
34	3	2	1	5.74725	il
35	3	0	3	10.68919	homme
36	3	1	2	8.23398	du
37	3	3	0	8.07056	des
38	3	1	2	12.82930	bestes
39	3	0	3	9.99758	b
40	2	2	0	8.11084	xab
41	2	2	0	9.24606	ung
42	2	2	0	15.34185	treuvent
43	2	2	0	8.05603	tout
44	2	1	1	8.75599	tous
45	2	0	2	9.68528	te

Search Term Words Case Regex

Window Span Same

From... 5L To... 5R

Min. Collocate Frequency 1

Sort by Invert Order

Total No. 436
Files Processed

Collocating words as a list

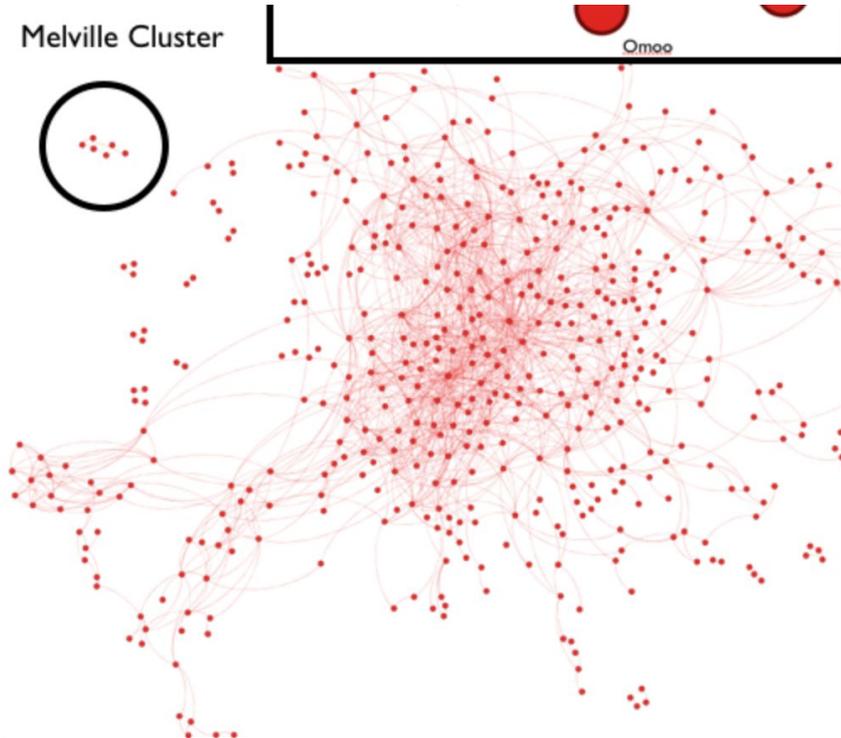
Widner, “Towards Text-Mining the Middle Ages” (2018)

Distant reading and *macroanalysis* are terms for modes of analysis that provide “the possibilities to answer long-standing, seemingly irresolvable questions, to generate previously unthought questions, and to explore the field of literature in as yet unforeseen ways ... Given the lure of text-mining, *where are the research projects using these techniques for new insights in medieval literature?* Unlike scholars of the eighteenth and nineteenth centuries, which have proven particularly fecund periods for distant reading, there are very few medievalists doing such research.” (p 132).

Outline

- PART ONE
 - What are open data (for humanities, and medieval studies in particular)?
 - What is a corpus? What is an open corpus?
 - What forms of corpus-based research have emerged in modern literary studies? What forms are emerging in medieval studies?
- PART TWO
 - Historical (& current) visions of a “body” of medieval French
 - The vision of OpenMedFr & open knowledge theory
 - Present workflow
 - A short demo video (8 mins)

Modern examples



Digital Harlem
Everyday Life 1915-1930

HOME ABOUT THE MAP SOURCES SEARCH Events SEARCH Places SELECT A PERSON TIMELINE PUBLICATI

Found 35 matches... Keep On Map

Events

Type of event
Performance (33)

Start date
- (33)

Start (general time of day)
select...

Source type
select...

Map-Timeline List

Map Satellite

Zoom to: Harlem

- Current query
- Secondary locations
- Bromley Fire Risk Maps
- Black Harlem 1920
- Black Harlem 1925
- Black Harlem 1930

Map data ©2018 Google | 200 m | Terms of Use | Report a map error

1915 1920 1925 1930 1935

Current query
Happy Rhone a... Miss Brunson's ...
Concert by pupi... Music Classiqu...
Dance Exhibiti... Concert by Mo...

The screenshot shows the Digital Harlem website interface. It features a navigation bar with links like HOME, ABOUT, THE MAP, SOURCES, SEARCH Events, SEARCH Places, SELECT A PERSON, TIMELINE, and PUBLICATI. The main content area displays search results for "Performance (33)" with filters for start date and source type. A map-timeline interface is visible, showing a map of Harlem with a timeline from 1915 to 1935. The map shows various locations marked with green icons, and the timeline shows a list of events such as "Happy Rhone a...", "Miss Brunson's ...", "Concert by pupi...", "Music Classiqu...", "Dance Exhibiti...", and "Concert by Mo...".

Photo: Matthew Jockers / University of Nebraska-Lincoln

A successful large corpus EEBO (1475-1700)

http:



HOME ABOUT TCP TEXTS RESOURCES WHY? JOIN THE TCP Search in this site...

Home » EEBO-TCP: Early English Books Online

EEBO-TCP: Early English Books Online



EEBO-TCP is a partnership with ProQuest and with more than 150 libraries to generate highly accurate, fully-searchable, SGML/XML-encoded texts corresponding to books from the Early English Books Online Database.

EEBO

The EEBO corpus consists of the works represented in the English Short Title Catalogue I and II (based on the Pollard & Redgrave and Wing short title catalogs), as well as the Thomason Tracts and the Early English Books Tract Supplement. Together these trace the history of

English thought from the first book printed in English in 1475 through to 1700. The content

What is the TCP?

The Text Creation Partnership creates standardized, accurate XML/SGML encoded electronic text editions of early print books. We transcribe and mark up the text from the millions of page images in ProQuest's Early English Books Online, Gale Cengage's Eighteenth Century Collections Online, and Readex's Evans Early American Imprints.

This work, and the resulting text files, are jointly funded and owned by more than 150 libraries worldwide. All of the TCP's work will be released the public domain for anyone to use.

Access the TCP Texts

- » ECCO-TCP Full text available to everyone
- » EEBO-TCP 25,000 texts available to everyone + 35,000 texts available only to EEBO-TCP partners
- » Evans-TCP Full text available to everyone

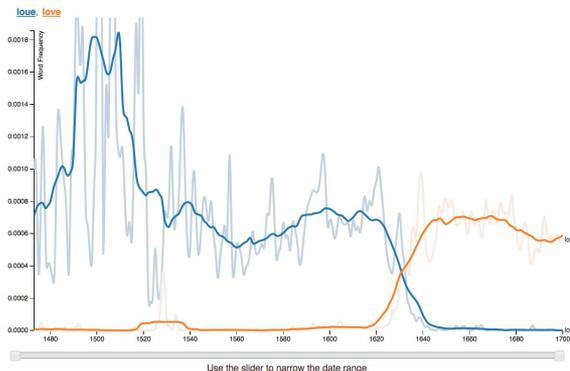
EEBO N-GRAM BROWSER

SHOW INSTRUCTIONS

ngramSize: unigrams bigrams trigrams
spellings: original regularized lemmas

gram 1: search term: love,love pos: _____

Graph smoothing Rolling Average:

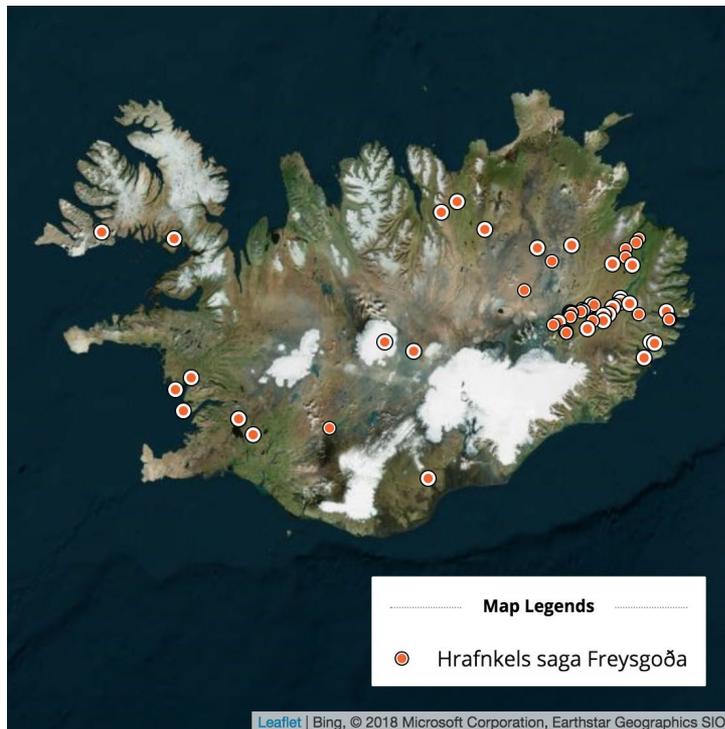


EEBO-TCP data preparation by Anupam Basu, made possible by Morphadorner, developed by Philip R. "Pib" Burns of Northwestern University. Interactive visualization by Stephen Pentecost.

How was EEBO created? *In partnership with ProQuest & 150 libraries (2000-present), slow collective transcription and progressive release into the public domain*

<https://earlyprint.wustl.edu/>

Icelandic saga map



1. kafli

Það var á d
svarta, Guðr
matarilla, Ey
maður kom
neðan [Fljóts](#)
Hann var þá

Hallfreður
Arnþrúður þ
færði Hallfr
[Geitdal](#). Og €

"þar liggju
vestur yfir [Lá](#)

Eftir það
síðan heitir .
göltur og ha
húsin, og týr

<http://sagamap.hi.is/is/>

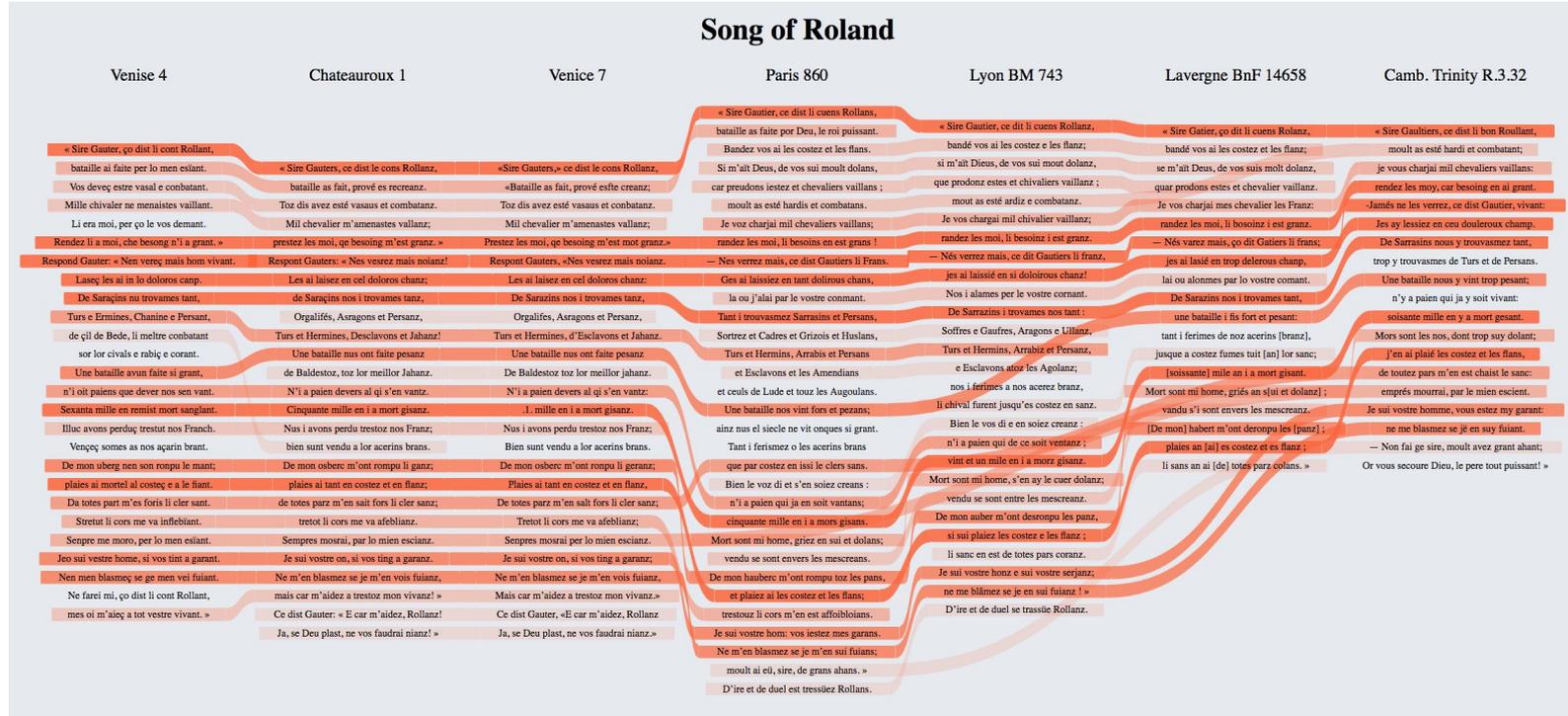
Mapping “rewriting”

-comp

d prose)



Visualizing *mouvance*



Digital Editions

-Scholastic Commentaries Text Archive (SCTA)

SCTA - Text - Search Corpus - Submit - About - Log in

Librum IV, Distinctio 1 [Aarau Transcription]
Diplomatic Transcription
By William de Rothwell

Edited by Jeffrey C. Witt
Edition: 2015.09-dev-master | September 09, 2015
Original Publication: Lombard Press, Baltimore, MD, September 09, 2015
License Availability: free, Published under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License
[A157ra]

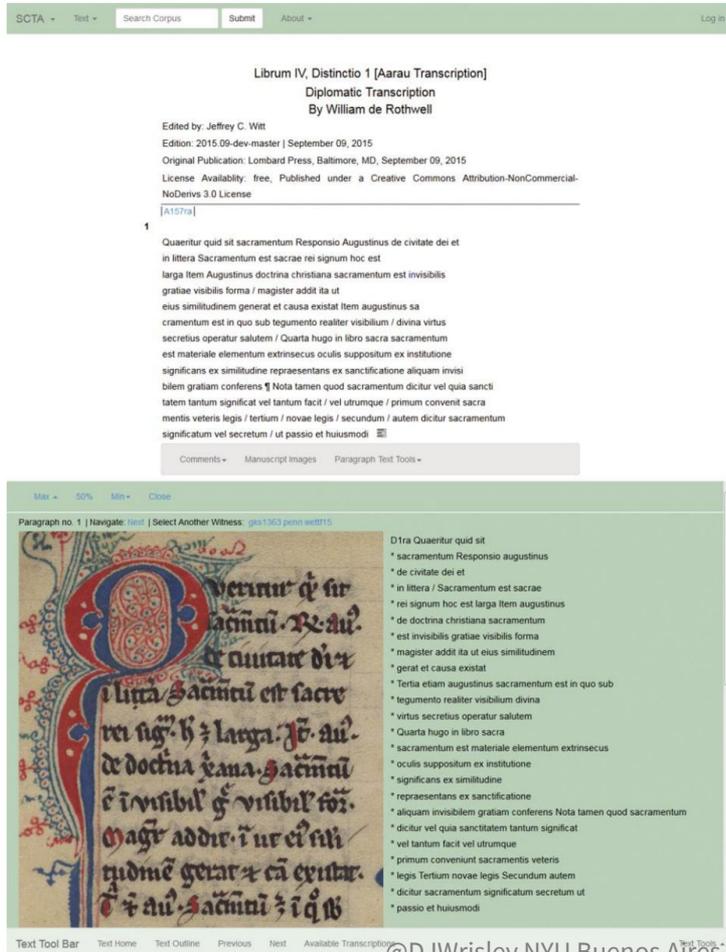
1

Quaeritur quid sit sacramentum Responso Augustinus de civitate dei et
in littera Sacramentum est sacrae rei signum hoc est
larga Item Augustinus doctrina christiana sacramentum est invisibilis
gratiae visibilis forma / magister addit ita ut
eius similitudinem generat et causa existat Item augustinus sa
cramentum est in quo sub tegumento realiter visibilium / divina virtus
secretus operatur salutem / Quarta hugo in libro sacra sacramentum
est materiale elementum extrinsecus oculis suppositum ex institutione
significans ex similitudine representans ex sanctificatione aliquam invis
ibilem gratiam conferens ¶ Nota tamen quod sacramentum dicitur vel quia sancti
tatem tantum significat vel tantum facit / vel utrumque / primum convenit sacra
mentis veteris legis / tertium / novae legis / secundum / autem dicitur sacramentum
significatum vel secretum / ut passio et huiusmodi

Comments - Manuscript Images - Paragraph Text Tools -

Para - 50% - Min - Close

Paragraph no. 1 | Navigate: text | Select Another Witness: gms1363 penn wett15



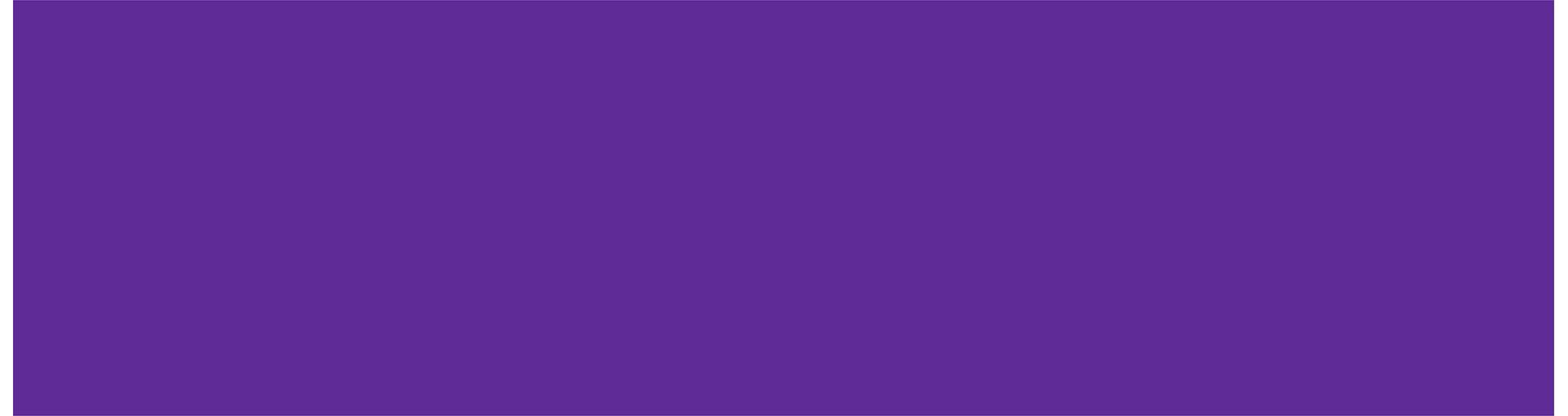
Dra Quaeritur quid sit
* sacramentum Responso augustinus
* de civitate dei et
* in littera / Sacramentum est sacrae
* rei signum hoc est larga Item augustinus
* de doctrina christiana sacramentum
* est invisibilis gratiae visibilis forma
* magister addit ita ut eius similitudinem
* gerat et causa existat
* Tertia etiam augustinus sacramentum est in quo sub
* tegumento realiter visibilium divina
* virtus secretus operatur salutem
* Quarta hugo in libro sacra
* sacramentum est materiale elementum extrinsecus
* oculis suppositum ex institutione
* significans ex similitudine
* representans ex sanctificatione
* aliquam invisibilem gratiam conferens Nota tamen quod sacramentum
* dicitur vel quia sanctitatem tantum significat
* vel tantum facit vel utrumque
* primum convenit sacramentis veteris
* legis Tertium novae legis Secundum autem
* dicitur sacramentum significatum secretum ut
* passio et huiusmodi

Text Tool Bar - Text Home - Text Outline - Previous - Next - Available Transcriptions - Text Tools -

Pause

¿Preguntas o comentarios?

Towards an open corpus of medieval French



Outline

- PART ONE
 - What are open data (for humanities, and medieval studies in particular)?
 - What is a corpus? What is an open corpus?
 - What forms of corpus-based research have emerged in modern literary studies? What forms are emerging in medieval studies?
- PART TWO
 - **Historical (& current) visions of a “body” of medieval French**
 - The vision of OpenMedFr & open knowledge theory
 - Present workflow
 - Participating in the project
 - A short demo video (8 mins)

ARLIMA

2700+ texts

in medieval French

(11th-15th c)

In 2014



DJ Wrisley

@DJWrisley



[@arlima_le_site](#) En 2007 ARLIMA comprenait 1700 notices, mais combien de textes en français sont répertoriés dans ARLIMA à présent?

Translate from French

6:17 AM - 18 Jul 2014 from [Lebanon](#)



1



Add another Tweet



Arlima [@arlima_le_site](#) · 18 Jul 2014



Replying to [@DJWrisley](#)

[@DJWrisley](#) En français seulement, il y a aujourd'hui 2702 textes répertoriés. Une 20aine de plus sont bilingues (français et surtout latin)

Translate from French



1



@DJWrisley NYU Buenos Aires 22-23 March, 27

***Looking back to the past of medieval
French***

Visions of a medieval French corpus (1)

- **Academy Publications**
- **Third Republic Learned Society:**
Société des anciens textes français (SATF):
“qui a pour « *but de publier des documents de toute nature rédigés au Moyen Âge en langue d'oïl ou en langue d'oc*” (1875)
- **Publication series:** Editions Honoré Champion (1875, bought by Slatkine, 1973) -> **Classiques français du moyen âge (CFMA)** (1910), Germany (**Altfranzösische Bibliothek**) and others
- **Theses and inaugural dissertations**
- **Journals:** *Romania*; *Romanische Forschungen*



In these cases, the corpus is not electronic, it is spread across different publication series.

Visions of a medieval French corpus (2)

“Textes en liberté” (Ottawa) - web-based critical editions (late 90s, early 2000s)

Éditions de textes

Les textes publiés dans les TEL sont également interrogeables dans la base TFA.

- Bestiaire marital tiré du *Rosarius* (Angela Mattiacci)
- Arthur dans *Le roman de Brut* de Wace (Pierre Kunstmann)
- Les chansons attribuées au trouvère picard Raoul de Soissons : édition critique (Ineke Hardy)
- *Le chevalier au lion* de Chrétien de Troyes (LFA)
- *Le chevalier au lion* de Pierre Sala (Pierre Kunstmann)
- *Le conte du Graal* de Chrétien de Troyes (Pierre Kunstmann)
- *Les enfances Garin de Monglane* (Aurélié Kostka-Durand)
- *Erec et Enide* de Chrétien de Troyes: transcription du ms. C (BnF, fr. 794) (Carleton W. Carroll)
- *Le miracle de l'enfant donné au diable* (Pierre Kunstmann)
- *Les miracles de Notre Dame* (LFA)
- *Les miracles de Notre-Dame de Chartres* de Jean le Marchant (Pierre Kunstmann)
- *Les miracles de Notre-Dame de Jean le Conte* (Pierre Kunstmann et Yen Duong)
- Les miracles de Notre-Dame tirés du *Rosarius* (Pierre Kunstmann)
- *Le roman de Mahomet* d'Alexandre du Pont (Yvan G. Lepage)
- Section lyrique : les chansons de trouvères (Ineke Hardy)
- *La vengeance Raguide!* (May Plouzeau)

UNIVERSITÉ D'OTTAWA Faculté des Arts
Laboratoire de français ancien

Miracle de l'enfant donné au diable

INTRODUCTION
TEXTE
APPARAT CRITIQUE
NOTES
CONCORDANCE
INDEX
ANALYSE DES FORMES VERBALES

1a/ Cy commence un miracle de Nostre Dame d'un enfant qui fu donné au dyable quant il fu engendré.

LA DAME.
1. Douce vierge, se vostre grez
2. Y est, je vous pri, consentez
3. Que mie donnee graces et sens
4. Se si oyez, par vostre assens
5. Que puisse vivre en chaasté;
6. Par vostre debonnaité
7. Donnez a monz mari enoage:
8. Comment que si' aie encore age
9. De delaisier pour ma veillesse,
10. Pour l'onneur de vostre hautesce
11. Je vous sy voué, fleur de lis,
12. Que jamais de ma char delis
13. Ne sera en vostre honneur fais.
14. Si en vueillez porter mon fais,
15. Chiere vierge, envers mon seigneur;
16. Autrement seroit en crementour
17. Que je n'eusse souz mal gré.
NOSTRE DAME.
18. Chiere amie, a ma volenté

Ici commence un miracle de Notre-Dame, d'un enfant qui fut donné au diable quand il fut engendré.
(A l'église, devant la statue de la Vierge)
LA DAME
Douce vierge, si cela vous agré, veuille, je vous prie, me donner la grâce et la sagesse d'agir, avec votre accord, de telle façon que je puisse vivre chastement; par votre bonté, inspirez mon mari, quoique je n'aie pas encore l'âge de m'abstenir pour cause de vieillesse, en l'honneur de votre noblesse, je vous ai, voué, fleur de lys,

INTRODUCTION
TEXTE
vms 1- 582
vms 583- 1215
vms 1216- 1592

INDEX
A-E
F-Z

Vers 1-582
[367c] De Mahomet
S'achuns veit oïr ou savoir
La vie Mahommet, avoï
En porra ichi commensache.
4/ En la terre le roi de France,
Mez jadis a Sens en Bourgoigne
[367d] Uns clers avoques J. chanoigne,
Ki Sarraïns avoit eñt,
8/ Mais prise avoit crestitenté,
Mahom del tout laïsté avoit,
Car touz la gille savoit
Que Mahommet fist en sa vie,

UNIVERSITÉ D'OTTAWA Faculté des Arts
Laboratoire de français ancien

May Plouzeau
La Vengeance Raguide! Edition critique
(fin de la rédaction: le 6 février 2002)

Publié d'après le ms M complété par A. Voir aussi les pages "Présentation", "Miria lectio de MF", "Miria lectio de ABL", "Miria lectio: p1
(Note: Les indications portées à droite du texte et concernant les chiffres de l'édition Fiedrowager et la mention "MS A" sont susceptibles d'être ou à l'impression si l'on se dispose pas d'un logiciel récent Consulter la page "Concordance entre Woylaguif et Woylaguif".)

1. Ce fu el novel sans d'enté, [364a
2. que il rois d'Arms et eost"
3. Nō le quantes à l'Évoqueur",
4. et vint à grant plent de gent"
5. à l'Évoqueur por sa cort tenir"
6. à Carlon, car maltoizor"
7. vōit il rois la costume her"
8. O lui fu li rois l'Éngouzeur",
9. si fu li rois l'Éngouzeur"
10. mais je de prince epl'i i ar"
11. ne vos lezrai en cest poier come".
12. lui con la maere eostor",
13. li rois tint eort à Carlon";
14. lui li prinse et lui li baron"
15. furent à la cort asanté",
16. si qu'à plus de gent a samble"
17. que mais il eut tant chevalier",
18. Li rois d'Arms est costumier"
19. que à fente ne manje"
20. devant ce qu'en sa cort entrast"
21. avoie d'aucune aventure"
22. Tūc fu lors la creseventance",
23. et il jens passe et la mais vint",
24. Coupees male eort i avant".

UNIVERSITÉ D'OTTAWA
Laboratoire de français ancien

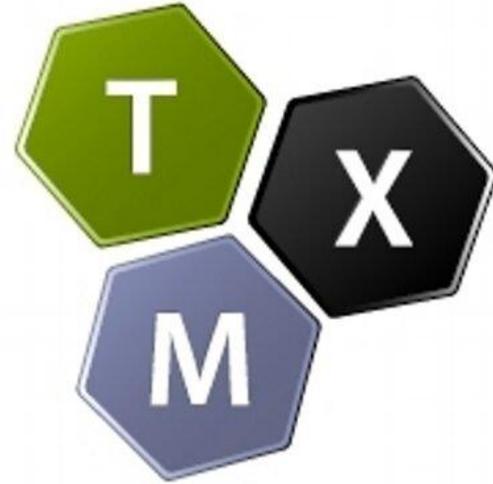
Alexandre du Pont
Le Roman de Mahomet
Paris, B.N., fr. 1553, f. 367v-379

Édition par
Yvan G. LEPAGE

Visions of a medieval French corpus (3)

Base de français médiéval (<https://cahier.hypotheses.org/bfm>)

La Base de français médiéval diffuse un corpus numérique de 153 textes français écrits entre le IXe et la fin du XV siècle (près de 4 100 000 occurrences-mots). Elle est accessible librement et gratuitement par le grand public comme par le monde académique (plus de 300 utilisateurs enregistrés). ***L'accès est médiatisé par la plateforme TXM***, qui offre un riche panel de fonctionnalités de recherche et d'exploitation des données numériques : création d'index, de concordances KWIC de mots ou de motifs textuels, vocabulaire de l'ensemble du corpus ou d'un texte particulier, navigation dans les pages d'édition, ***téléchargement de l'édition des textes libres de droits au format PDF***, calcul de mots spécifiques à un sous-corpus, calcul de collocations, etc.



Visions of a medieval French corpus (4)

Corpus de Littérature Médiévale des origines à la fin au 15e siècle, Éditions Champion Électronique / Classiques Garnier Numérique.

- Pay-walled, expensive subscription
- incorporates CFMA/Slatkine series
- 184 texts
- very limited interaction with texts

A next generation medieval French corpus

Vision of OpenMedFr <https://github.com/OpenMedFr>

- A corpus with textual material published both in and outside of France (Germany, England, Scandinavia, whatever *is in the public domain*) - “the data is the edition” (J. Flanders) -> **the open data is the public domain edition.**
- Strong definition of open (stronger than *textes en liberté*) -- explicit Creative Commons license, downloadable in their entirety in plain txt files, unique URLs. A **“Project Gutenberg for medieval French”**
- Data-first approach, **“minimizing cost and reducing complexity”** (DataFirst Manifesto), rely on past editorial work, rather mere a sprint to produce large corpora quickly (Niles/Poston).
- **Handling heterogeneous editorial principles** in the creation of a plain text archive, **expanded metadata.**
- **Community-sourced and curated in the open:** all acts of data creation (even student work) acknowledged.
- Expanded notion of metadata to include uncertainty of time, linked data and an **explicit attention to intellectual labor that is often hidden in humanities data** (ORCID, web publisher, web reviewer)

***Bringing the public domain past of
medieval French scholarship into the open***

Initial Workflow of OpenMedFr <https://github.com/OpenMedFr>

- Building a list of desiderata, as **complete** as possible from public domain
- Setting threshold of acceptable error
- Optical character recognition (OCR) of already scanned documents from openly available digital libraries (Gallica, archive.org)
- Pattern training in OCR to train the software with human input and situating the textual creation partnership within the community of practice (unlike EEBO that came into being, typed to “Asian countries.”)
- **Correct** against the original
- **Connect** appropriate metadata from recognized sources in the disciplines
- Publishing in GitHub with Creative Commons license
- Quality control via an editorial committee (to come)

Why GitHub?

-free platform for versioning of code/data

-open

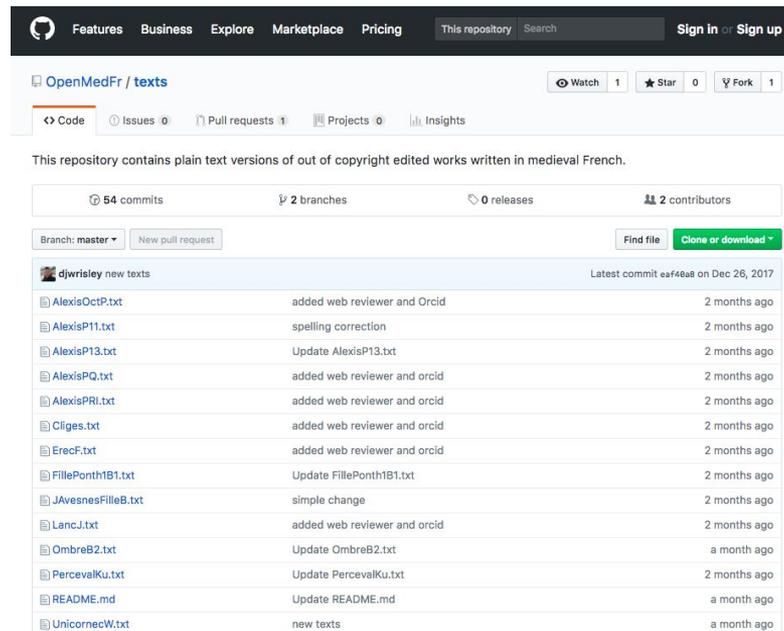
-versioning: continuous improvement of the corpus, publication of versions with DOI

-allows for geographically disparate teams

-static, sustainable website

-easily downloadable, unique URLs for texts

-used by other digital humanities projects: The Programming Historian, Open ITI, LLBeirut



The screenshot shows the GitHub interface for the repository 'OpenMedFr / texts'. The repository is described as containing 'plain text versions of out of copyright edited works written in medieval French'. It has 54 commits, 2 branches, 0 releases, and 2 contributors. The current branch is 'master'. A table of recent commits is displayed, showing the author 'djwrisley' and the commit message 'new texts' for the file 'UnicornecW.txt' committed a month ago. Other files listed include 'AlexisOctP.txt', 'AlexisP11.txt', 'AlexisP13.txt', 'AlexisPQ.txt', 'AlexisPRI.txt', 'Cliges.txt', 'ErecF.txt', 'FilePonth1B1.txt', 'JAvesnesFileB.txt', 'LancJ.txt', 'Ombreb2.txt', and 'PercevalKu.txt', all committed 2 months ago.

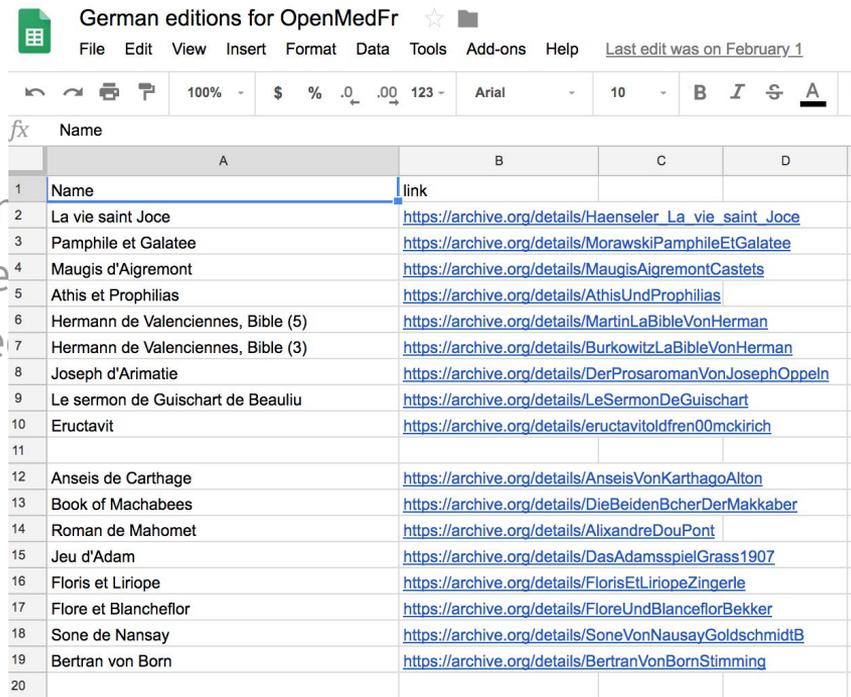
File	Commit Message	Time
UnicornecW.txt	new texts	a month ago
AlexisOctP.txt	added web reviewer and Orcid	2 months ago
AlexisP11.txt	spelling correction	2 months ago
AlexisP13.txt	Update AlexisP13.txt	2 months ago
AlexisPQ.txt	added web reviewer and orcid	2 months ago
AlexisPRI.txt	added web reviewer and orcid	2 months ago
Cliges.txt	added web reviewer and orcid	2 months ago
ErecF.txt	added web reviewer and orcid	2 months ago
FilePonth1B1.txt	Update FilePonth1B1.txt	2 months ago
JAvesnesFileB.txt	simple change	2 months ago
LancJ.txt	added web reviewer and orcid	2 months ago
Ombreb2.txt	Update Ombreb2.txt	a month ago
PercevalKu.txt	Update PercevalKu.txt	2 months ago
README.md	Update README.md	a month ago

“schöpferische Zerstörung”

- creative destruction of the edition facilitating new forms of computer-assisted reading
- respects the edition “as is” - in normalized format
- does not eliminate the pdf of scans of archive.org, but it makes them actionable in digital environments.
- resembles Project Gutenberg that provides plain text utf-8 editions that its founder “believed to be the format most likely to be readable in the extended future.”
- incorporates metadata/details of remediation in a “header” for automated work.

Proposed starting point

OpenMedFr: an expanded SATF + Altfranzösisch
inaugural dissertations and journal published e
Forschungen), targeting gaps in these combine



German editions for OpenMedFr

File Edit View Insert Format Data Tools Add-ons Help Last edit was on February 1

100% - \$ % .0 .00 123 - Arial - 10 - B I S A ↕

	Name			
	A	B	C	D
1	Name	link		
2	La vie saint Joce	https://archive.org/details/Haenseler_La_vie_saint_Joce		
3	Pamphile et Galatee	https://archive.org/details/MorawskiPamphileEtGalatee		
4	Maugis d'Aigremont	https://archive.org/details/MaugisAigremontCastets		
5	Athis et Prophilias	https://archive.org/details/AthisUndProphilias		
6	Hermann de Valenciennes, Bible (5)	https://archive.org/details/MartinLaBibleVonHerman		
7	Hermann de Valenciennes, Bible (3)	https://archive.org/details/BurkowitzLaBibleVonHerman		
8	Joseph d'Armatie	https://archive.org/details/DerProsaromanVonJosephOppeln		
9	Le sermon de Guischart de Beauliu	https://archive.org/details/LeSermonDeGuischart		
10	Eruclavit	https://archive.org/details/eructavitoldfren00mckrich		
11				
12	Anseis de Carthage	https://archive.org/details/AnseisVonKarthagoAlton		
13	Book of Machabees	https://archive.org/details/DieBeidenEcherDerMakkaber		
14	Roman de Mahomet	https://archive.org/details/AlixandreDouPont		
15	Jeu d'Adam	https://archive.org/details/DasAdamsspielGrass1907		
16	Floris et Liriope	https://archive.org/details/FlorisELiriopeZingerle		
17	Flöre et Blanche flor	https://archive.org/details/FlöreUndBlanceflorBekker		
18	Sone de Nansay	https://archive.org/details/SoneVonNausayGoldschmidtB		
19	Bertran von Born	https://archive.org/details/BertranVonBornStimming		
20				

Interested in participating in OpenMedFr?

- **Skills needed:** attention to detail, desire to contribute, some knowledge of French, no complex computer skills, ability to work on a screen and commit completing a text of 50-100 pages.
- **Tools needed:** some OCR software (we use Abbyy Finereader), two free downloads: text editor-Atom, GitHub desktop
- **Skills gained:** teamwork on a long-term, international digital humanities project, project management with GitHub, remediation of texts, introduction to using plain text for digital analysis
- **Attribution:** if you complete a text or designated portion of text, in the metadata you are listed as web author with your ORCID
- **Leadership:** In the near future we will have an editorial board made up of the most active members.
- ***Do you need to contribute to use the resource? NO***

RECAPITULATION

- PART ONE
 - What are open data in general (for humanities/medieval studies in particular)?
 - What is a corpus? What is an open corpus?
 - What forms of corpus-based research have emerged in modern literary studies? What forms are emerging in medieval studies?
- PART TWO
 - Historical (& current) visions of a “body” of medieval French
 - The vision of OpenMedFr & open knowledge theory
 - Present workflow
 - Participating in the project
 - A short demo video (8 mins)

Abbyy - free trial version

30 days

100 pages

I have prepared 3 page segments of the Roman de Troie to practice

Basic edition costs about \$100

End of Session 1

Contact: [djw12 \(at\) nyu \(dot\) edu](mailto:djw12@nyu.edu) or [@DJWrisley](https://twitter.com/DJWrisley) / djwrisley.com

GitHub for project: github.com/OpenMedFr

Check out the CFP for my proposed MAA panel 2019 ([Significatio sana in corpore sano](#))