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

Open data in general (governments)

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

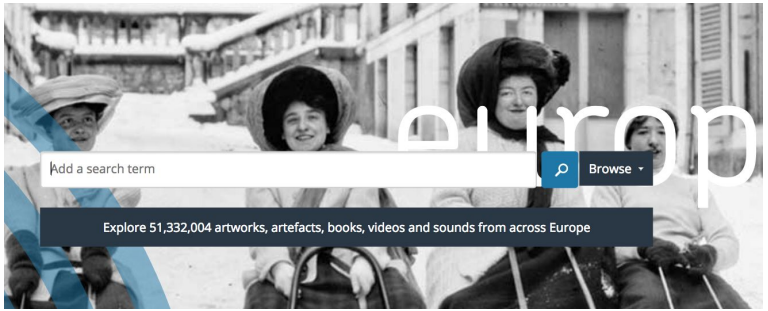
Open Data and the Knowledge Society (2017) - OKN's definition

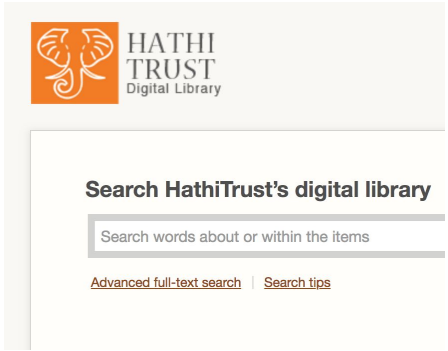
- **Availability and access:** data available as a whole and in a convenient and modifiable form.
- **Reuse and redistribution:** data must be provided under terms that permit reuse and redistribution including the intermixing with other datasets. The data must be machine-readable.
- **Universal participation:** 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.

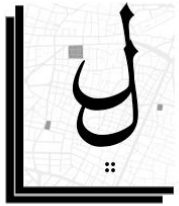
Pause

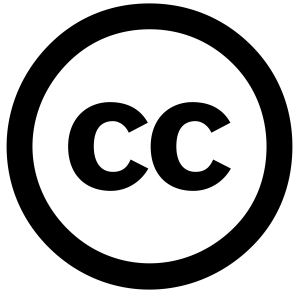
How might these ideas of Open Data and Open Knowledge be related to our research, our intellectual work?

Examples of Open Data in the humanities









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;

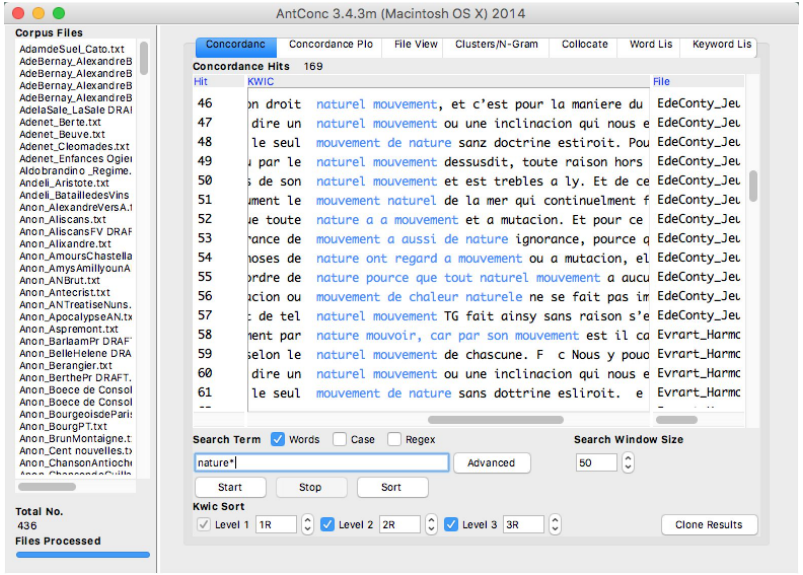
4. it depends on both quantitative and qualitative analytical techniques.

(*italics mine*)

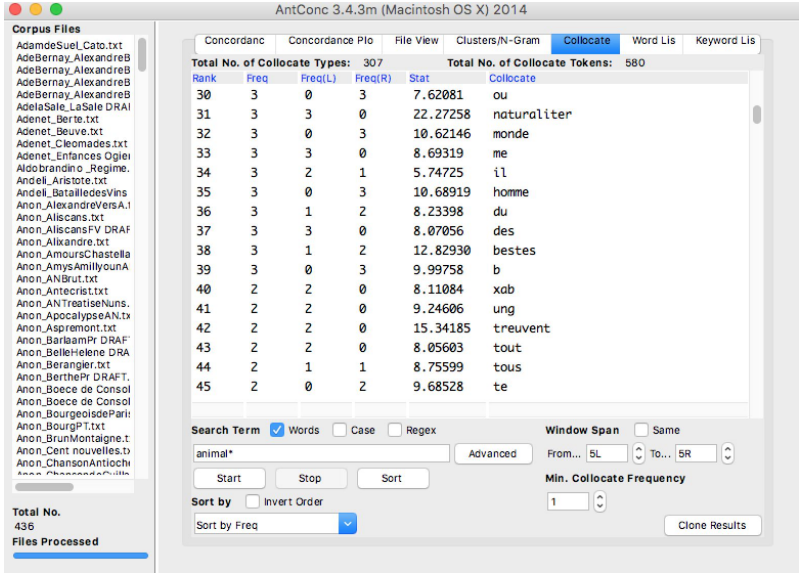
NB: A corpus is not a canon

- 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



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



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

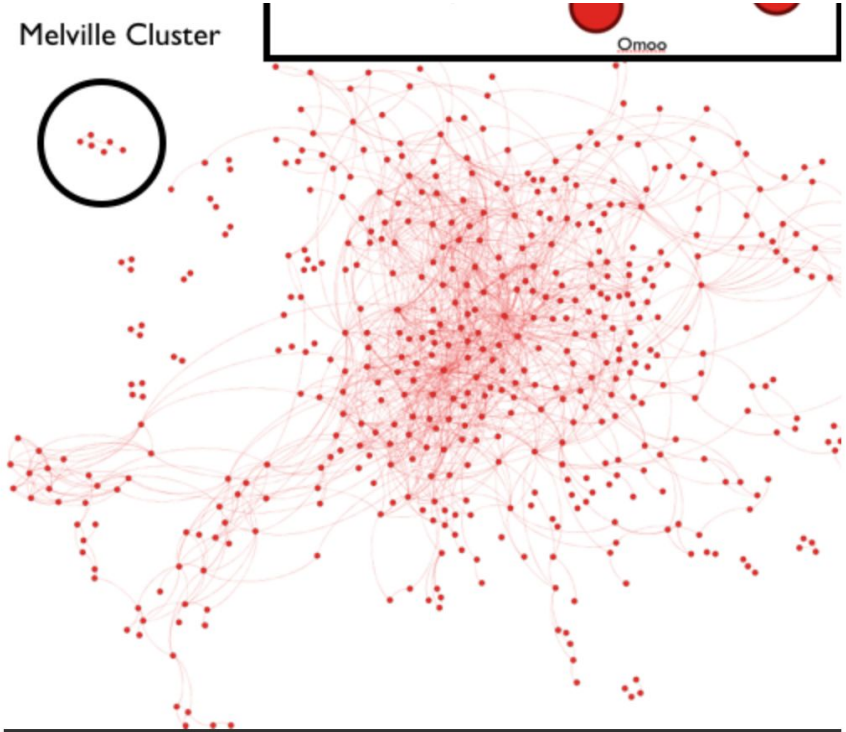
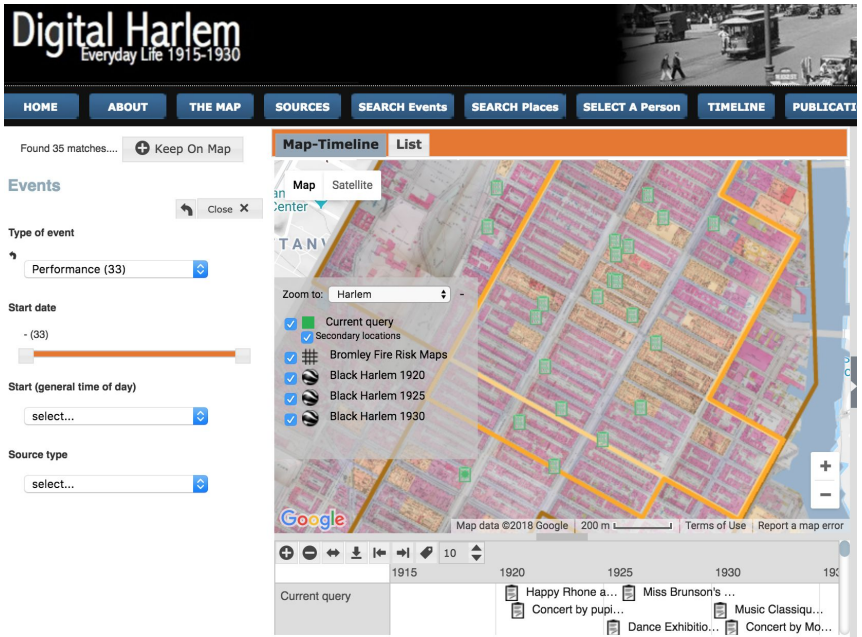
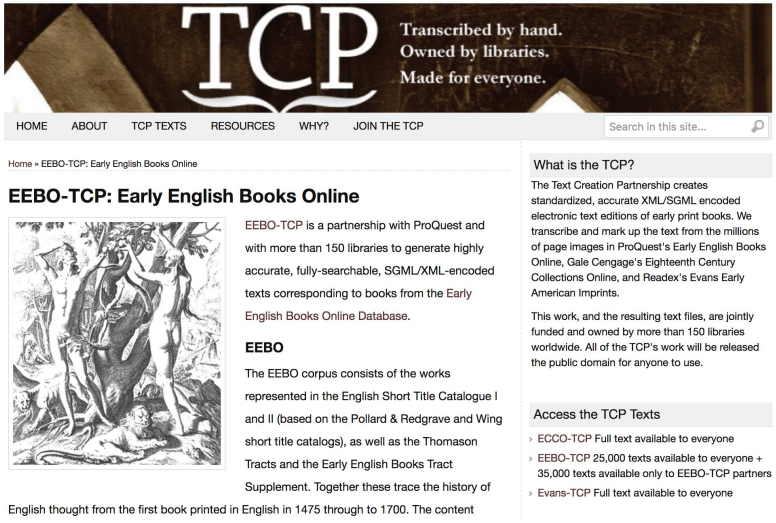


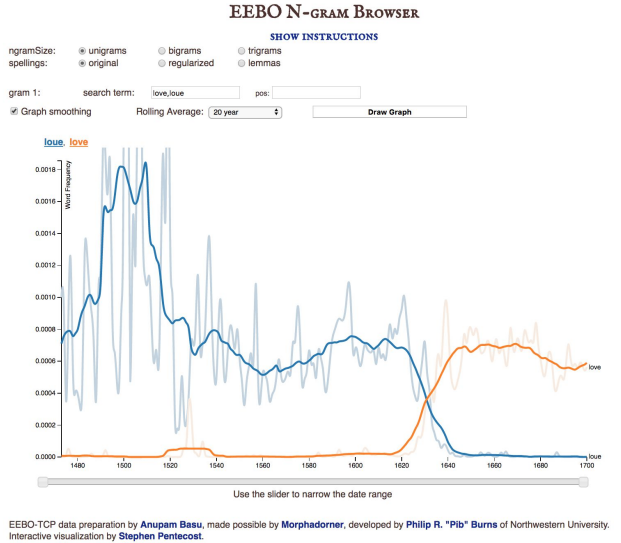
Photo: Matthew Jockers / University of Nebraska-Lincoln



A successful large corpus EEBO (1475-1700)

http





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/

Using a medieval corpus beyond linguistic corpus analysis…

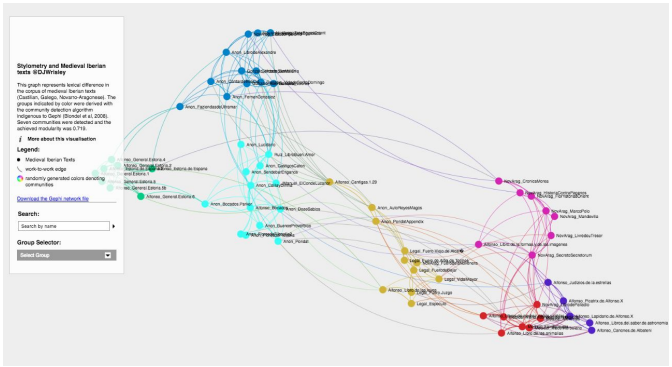
1. kafli

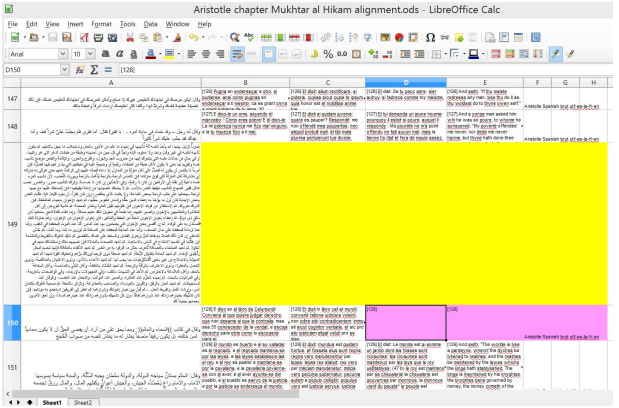
Það var á c
svarta, Guðr
matarilla, Ey
maður kom
neðan Fljóts
Hann var þá

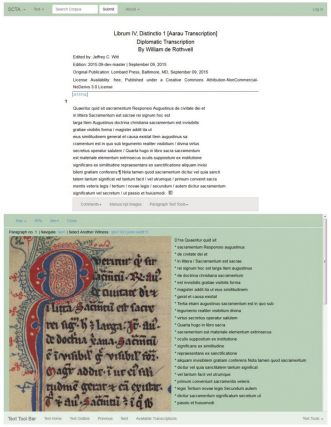
Hallfreður
Arnþrúður h
færði Hallfr
Geitdal. Og e

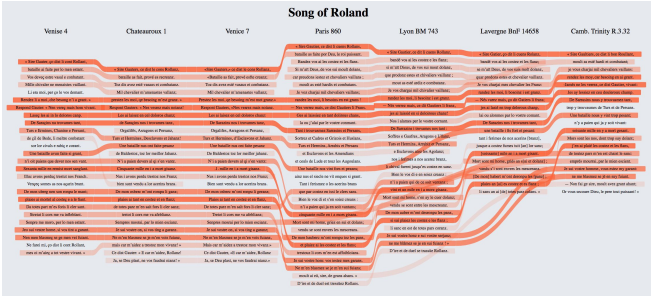
"Þar liggu
vestur yfir La

Eftir það
síðan heitir
göltur og ha
húsin, og týr



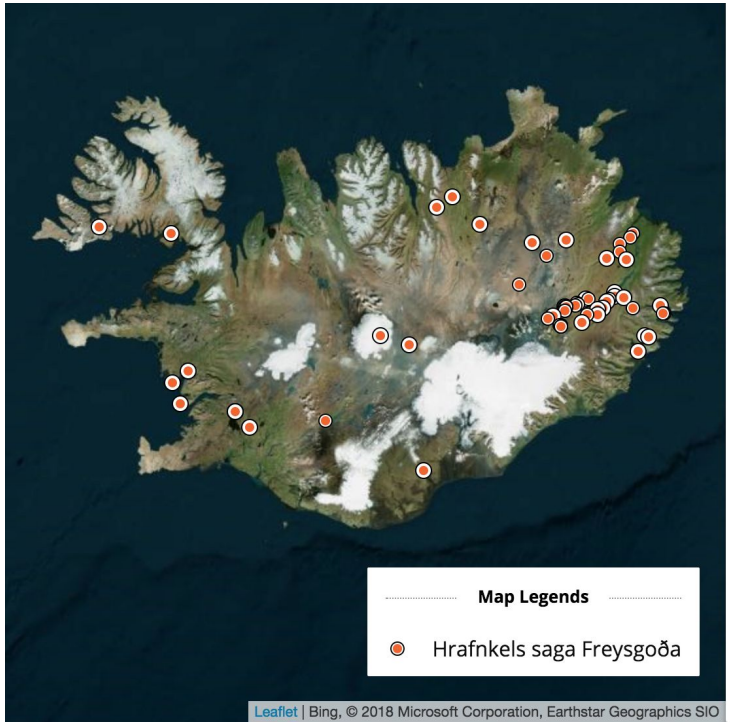






@DJWrisley NYU Buenos Aires 22-23 March, 18

Icelandic saga map



1. kafli

Það var á d
svarta, Guðr
matarilla, Ey
maður kom
neðan Fljóts
Hann var þá

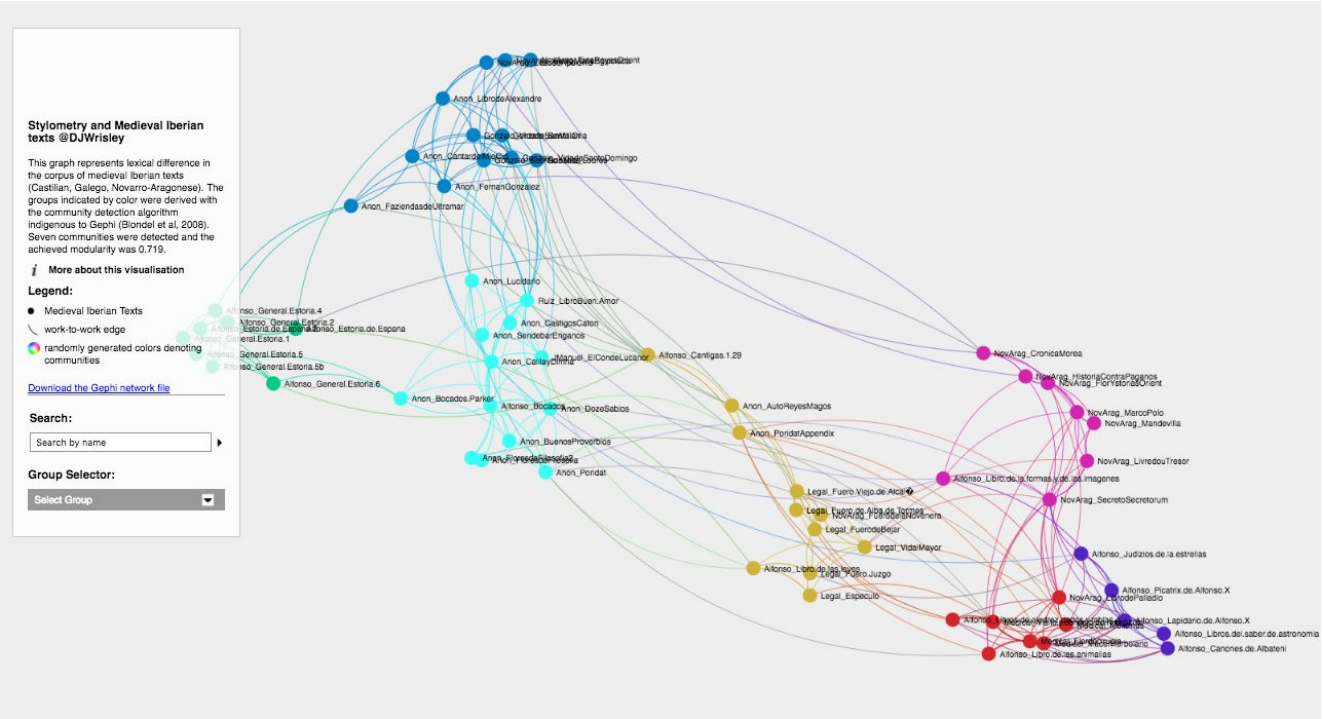
Hallfreðu
Arnþrúður h
færði Hallfr
Geitdal. Og e

"Þar liggu
vestur yfir La

Eftir það
síðan heitir
göltur og ha
húsin, og týr

http://sagamap.hi.is/is/

Stylometry in old Spanish



http://djwrisley.com/networks/HDH2015/castellano/

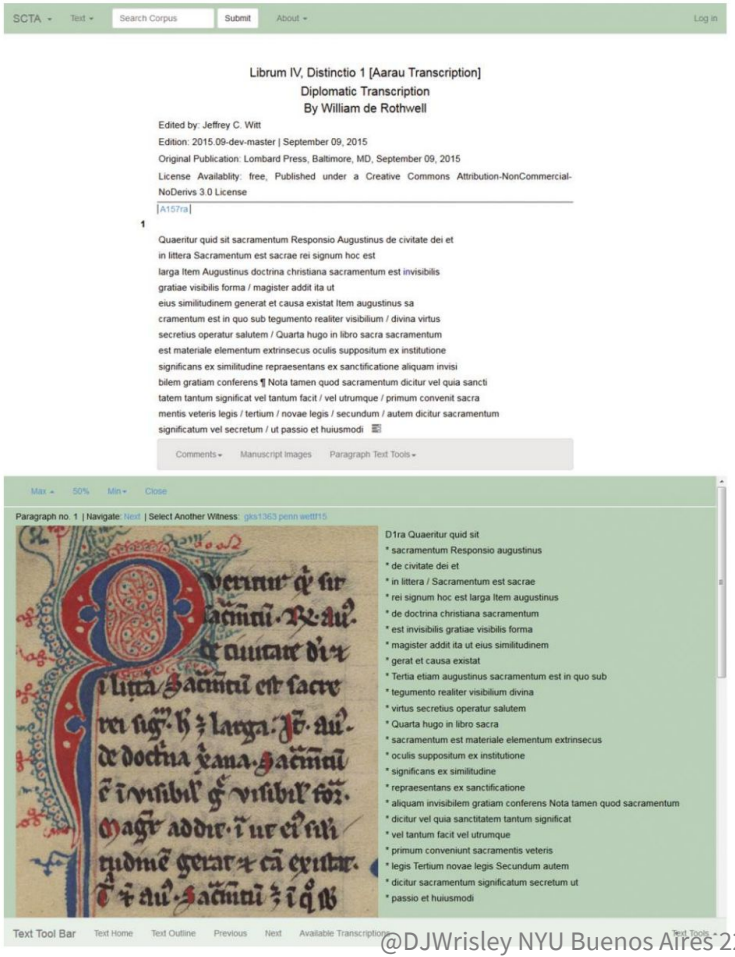
Mapping “rewriting”

-comp d prose)

Visualizing *mouvance*

Digital Editions

-Scholastic Commentaries Text Archive (SCTA)



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

Looking back to the past of medieval French

Visions of a medieval French corpus (1)

- **Academy Publications**
- **Third Republic Learned Society**:
  Sociéte 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.

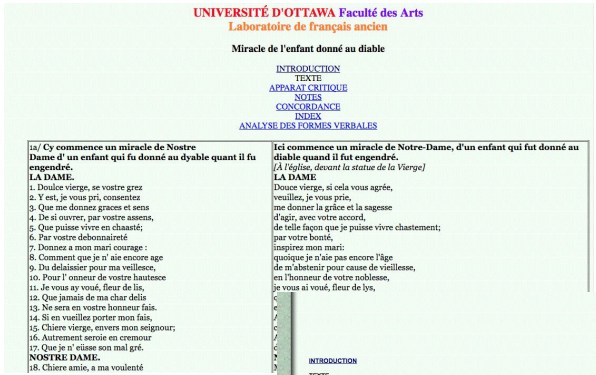
Visions of a medieval French corpus (2)

“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 marial 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* (Aurélie 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 Nostre 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 Raguidel* (May Plouzeau)

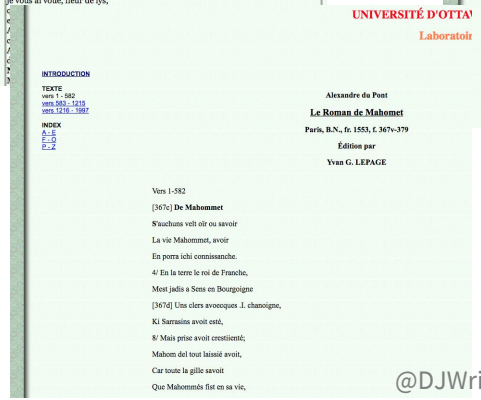


UNIVERSITÉ D'OTTAWA Faculté des Arts
Laboratoire de français ancien

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.**	**Ici commence un miracle de Notre-Dame, d'un enfant qui fut donné au diable quand il fut engendré.**
LA DAME.	*[À l'église, devant la statue de la Vierge]*
1. Doulce vierge, se vostre grez	**LA DAME**
2. Y est, je vous pri, consentez	Douce vierge, si cela vous agrée,
3. Que me donnez graces et sens	veuillez, je vous prie,
4. De si ouvrer, par vostre assens,	me donner la grâce et la sagesse
5. Que puisse vivre en chasté;	d'agir, avec votre accord,
6. Par vostre debonnaireté	de telle façon que je puisse vivre chastement;
7. Donnez a mon mari courage :	par votre bonté,
8. Comment que je n' aie encore age	inspirez mon mari:
9. Du delaissier pour ma veillesce,	quoique je n'aie pas encore l'âge
10. Pour l' onneur de vostre hautesce	de m'abstenir pour cause de vieillesse,
11. Je vous ay voué, fleur de lys,	en l'honneur de votre noblesse,
12. Que jamais de ma char delis	je vous ai voué, fleur de lys,
13. Ne sera en vostre honneur fais.	
14. Si en vueilles porter mon fais,	
15. Chiere vierge, envers mon seignour;	
16. Autrement seroie en cremour	
17. Que je n' eüsse son mal gré.	
NOSTRE DAME.	
18. Chiere amie, a ma voulenté	



UNIVERSITÉ D'OTTA
Laboratoi

INTRODUCTION

TEXTE
vers 1 - 582
vers 583 - 1215
vers 1216 - 1997

INDEX
A - E
F - O
P - Z

Alexandre du Pont

Le Roman de Mahomet

Paris, B.N., fr. 1553, f. 367v-379

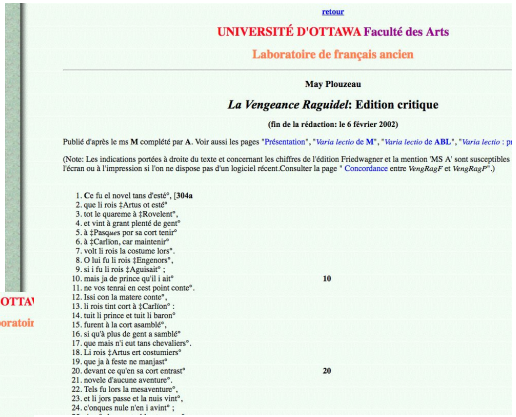
Édition par

Yvan G. LEPAGE

Vers 1-582

[367c] **De Mahommet**

S'auchuns velt oïr ou savoir
La vie Mahommet, avoir
En porra ichi connissanche.
4/ En la terre le roi de Franche,
Mest jadis a Sens en Bourgoigne
[367d] Uns clers avoecques .I. chanoigne,
Ki Sarrasins avoit esté,
8/ Mais prise avoit crestïienté;
Mahom del tout laissié avoit,
Car toute la gille savoit
Que Mahommès fist en sa vie,



retour

UNIVERSITÉ D'OTTAWA Faculté des Arts
Laboratoire de français ancien

May Plouzeau

***La Vengeance Raguidel*: Edition critique**

(fin de la rédaction: le 6 février 2002)

Publié d'après le ms **M** complété par **A**. Voir aussi les pages "Présentation", "*Varia lectio* de **M**", "*Varia lectio* de **ABL**", "*Varia lectio* : p

(Note: Les indications portées à droite du texte et concernant les chiffres de l'édition Friedwagner et la mention 'MS A' sont susceptibles l'écran ou à l'impression si l'on ne dispose pas d'un logiciel récent.Consulter la page " Concordance entre *VengRagF* et *VengRagF'*.)

1. Ce fu el novel tans d'esté, [**304a**
2. que li rois ‡Artus ot esté
3. tot le quareme à ‡Rovelent",
4. et vint à grant plenté de gent
5. .ì. à ‡Pasques por sa cort tenir
6. à ‡Carlion, car maintenir
7. volt li rois la costume lors".
8. O lui fu li rois ‡Engenors",
9. si i fu li rois ‡Aguisaïs" ;
10. mais ja de prince qu'il i ait" **10**
11. ne vos tenrai en cest point conte".
12. Issi con la matere conte",
13. li rois tint cort à ‡Carlion" ;
14. tuit li prince et tuit li baron"
15. furent à la cort asamblé",
16. si qu'à plus de gent a semblé"
17. que mais n'i eut tans chevaliers".
18. Li rois ‡Artus ert costumiers"
19. que ja à feste ne manjast"
20. devant ce qu'en sa cort entrast" **20**
21. novele d'aucune aventure".
22. Tels fu lors la mesaventure",
23. et li jors passe et la mais vint",
24. c'onques nule n'en j avint" ;

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

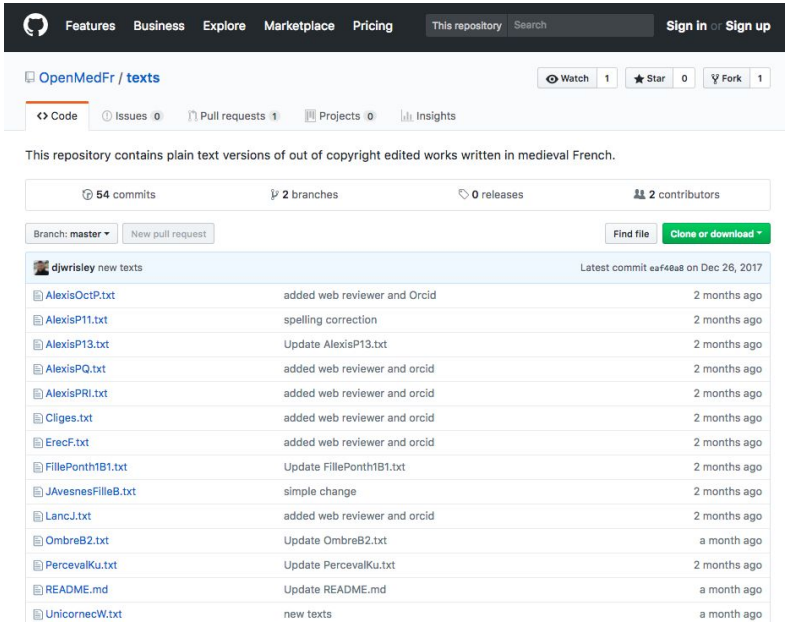
-open

-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

“schöpferische Zerstörung”

-creative destruction of the edition facilitating new forms of computer-assisted reading

-respects the edition “as is” - in normalized format

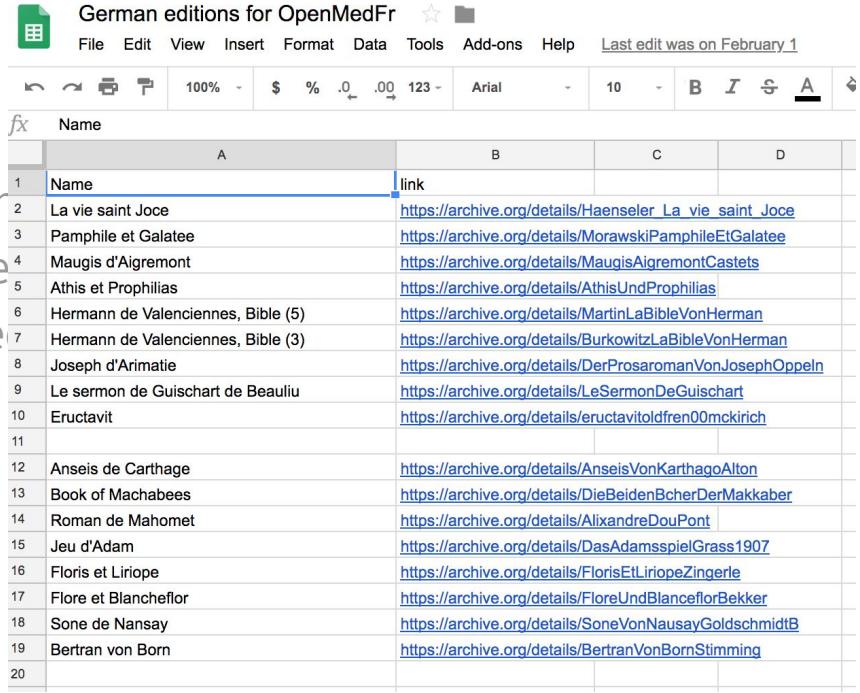
-does not eliminate the pdf of scans of archive.org, but it makes them actionable in digital environments.

-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

fx 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'Arimatie	https://archive.org/details/DerProsaromanVonJosephOppeln		
9	Le sermon de Guischart de Beauliu	https://archive.org/details/LeSermonDeGuischart		
10	Eructavit	https://archive.org/details/eructavitoldfren00mckirich		
11				
12	Anseis de Carthage	https://archive.org/details/AnseisVonKarthagoAlton		
13	Book of Machabees	https://archive.org/details/DieBeidenBcherDerMakkaber		
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/FlorisEtLiriopeZingerle		
17	Flore et Blancheflor	https://archive.org/details/FloreUndBlanceflorBekker		
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 or @DJWrisley / djwrisley.com

GitHub for project: github.com/OpenMedFr

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