



Explorer les corpus :
vue d'ensemble et navigation
dans de larges corpus documentaires

Jean-Daniel Fekete
AVIZ
INRIA Saclay – Île-de-France
www.aviz.fr



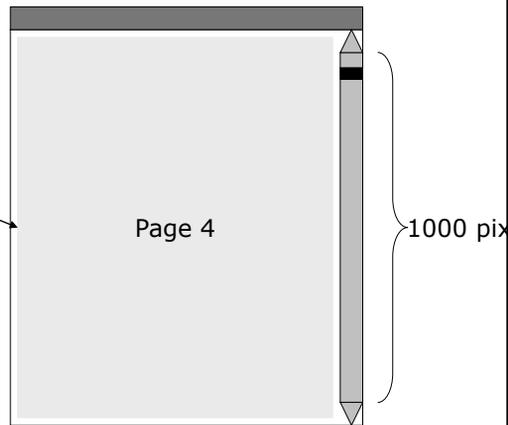
Grand corpus, vue d'ensemble et navigation

- Les documents s'agrègent en corpus
- Les corpus s'agrègent en collections de corpus
- Les centres de ressources numérique recensent des collections de corpus

- Comment peut-on avoir une vue d'ensemble ?

Idée 1 : étendre la barre de défilement

- Les 37 pièces de Shakespeare mises bout-à-bout représentent 150 000 lignes de texte
- Si on affiche 40 lignes par page (écran), il faut 3750 pages
 - Bouger la barre d'1 pixel fait sauter le texte de 3.75 pages



OrthoZoom Scroller Navigation Multi-échelle en 1D

Caroline Appert
Jean-Daniel Fekete
INRIA / LRI
Caroline.Appert@lri.fr
Jean-Daniel.Fekete@inria.fr

Question scientifique

- Quelle est la limite de la technique ?
- Quelles sont les prérequis ?
- Problème de la sélection en 1D
 - Sélectionner 1 parmi N
 - Théorie de l'information, l'indice de difficulté (ID) correspond à la quantité d'information à fournir au système pour lui préciser le 1 parmi N
 - $ID = \log_2(150\,000/40) = 12$ bits
 - Pour sélectionner une base dans un brin d'ADN humain, $ID=33$
 - Comment approcher pratiquement cette limite théorique ?

- L'informatique n'est pas plus la science des ordinateurs que l'astronomie n'est la science des télescopes.

[Hal Abelson ou Edsger W. Dijkstra ou Michael R. Fellows]

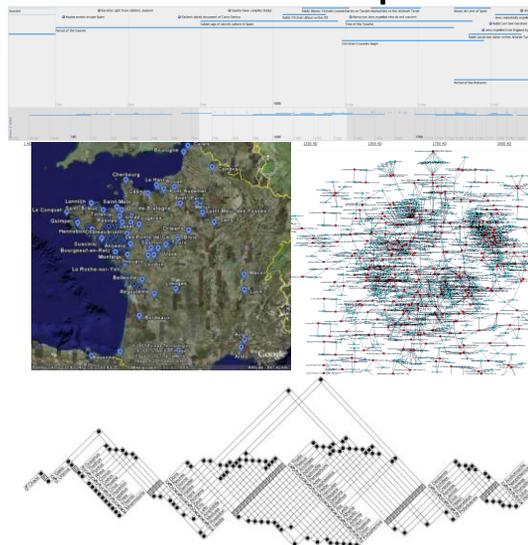
Les facettes des corpus

- Un corpus a plusieurs facettes
 - Table des matières, index des noms, des lieux
- En général, les sites ne donnent pas ces informations au premier niveau
 - Je dois savoir quoi chercher
- Ça ne marche plus lorsqu'on agrège les sites



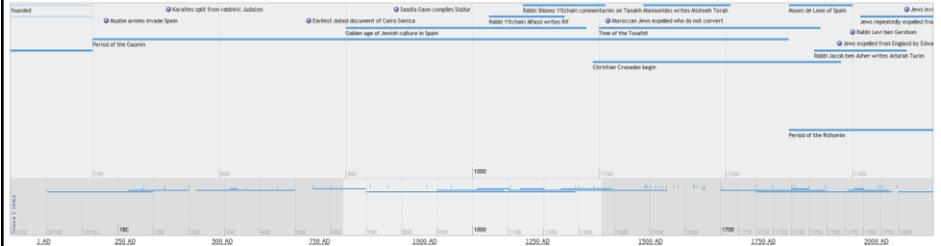
Visualiser les facettes d'un corpus

- Table des matières
- Index
- Ligne de temps / Chronologies
- Cartes
- Réseaux sociaux
 - Généalogies
- Évolution des textes



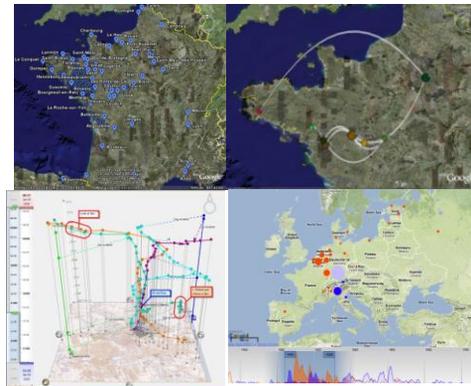
Lignes de temps / Chronologies

- Quelles époques sont couvertes/référencées ?
- Avec quelles densités ?
- Deux niveaux de détails
 - vue d'ensemble+zoom local
- Voir <http://www.simile-widgets.org/> ou d3.js ou protovis.org



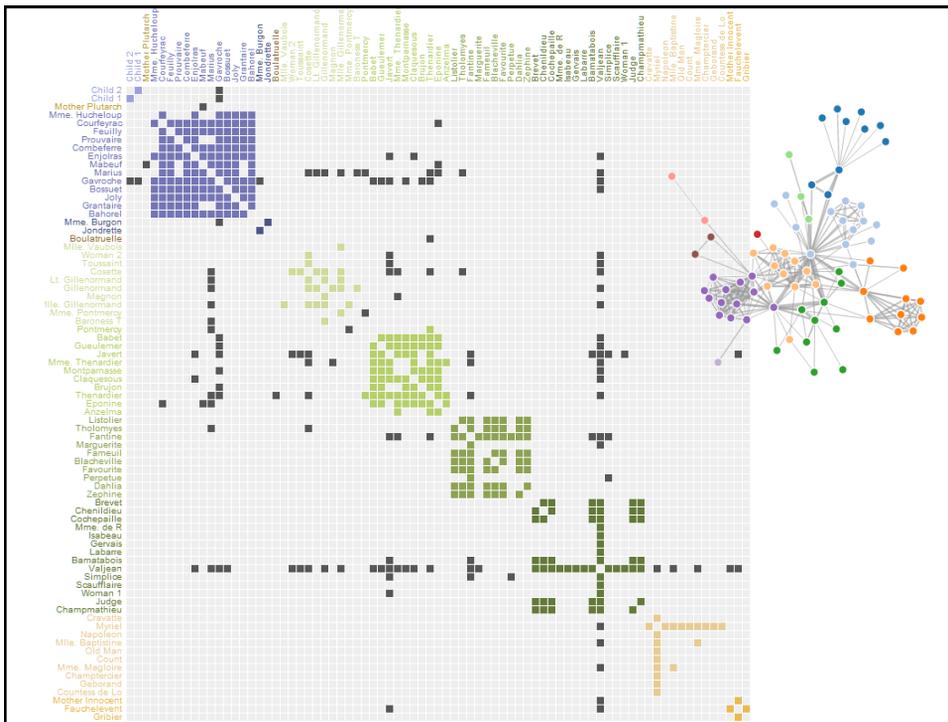
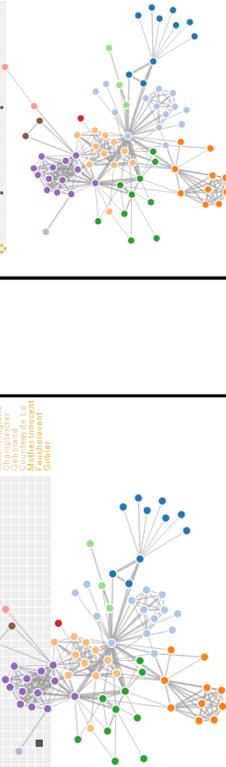
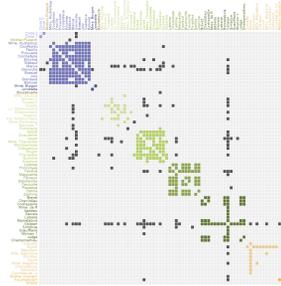
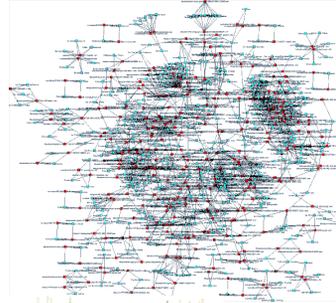
Cartes

- Index de toponymes sur google map, liés à des documents
 - Possibilité de montrer les flux
- Cartes dynamiques
 - Europeana4D
- Combiner l'espace et le temps
- Info sur le contenu du corpus
- Delimite une recherche dans le corpus



Réseaux sociaux

- Relations entre personnes / organisations / documents etc.
 - Facilement illisible
- La représentation matricielle facilite la lecture des réseaux denses



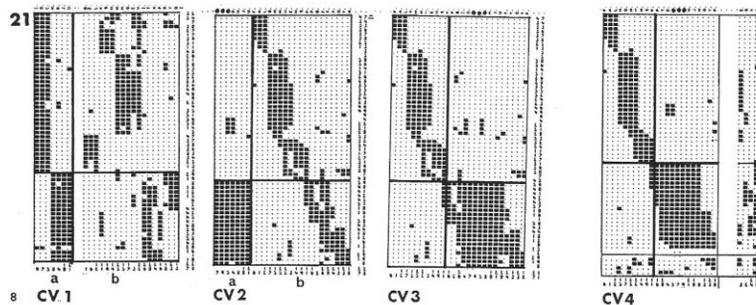
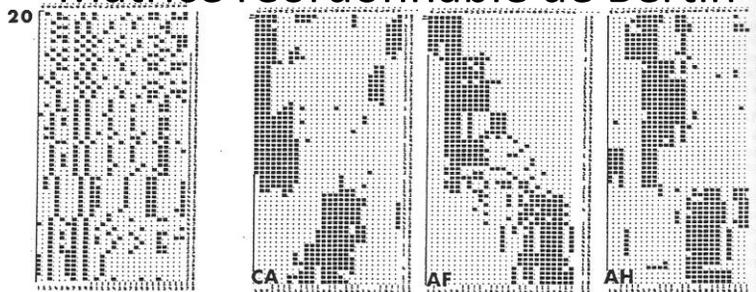
En mémoire de Jacques Bertin (27 juillet 1918 -- 3 mai 2010)



Semiologie Graphique
1^{re} édition: 1967
Dernière édition anglaise: 2010

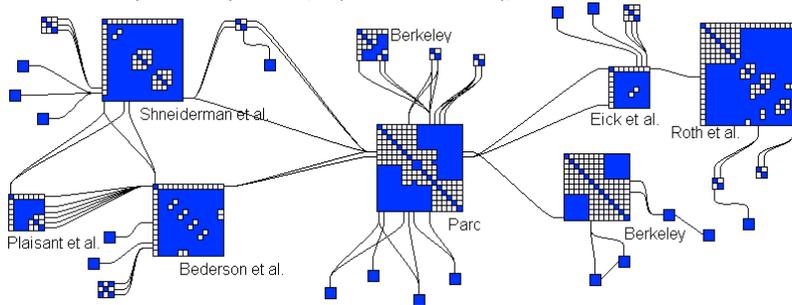


Matrice réordonnable de Bertin



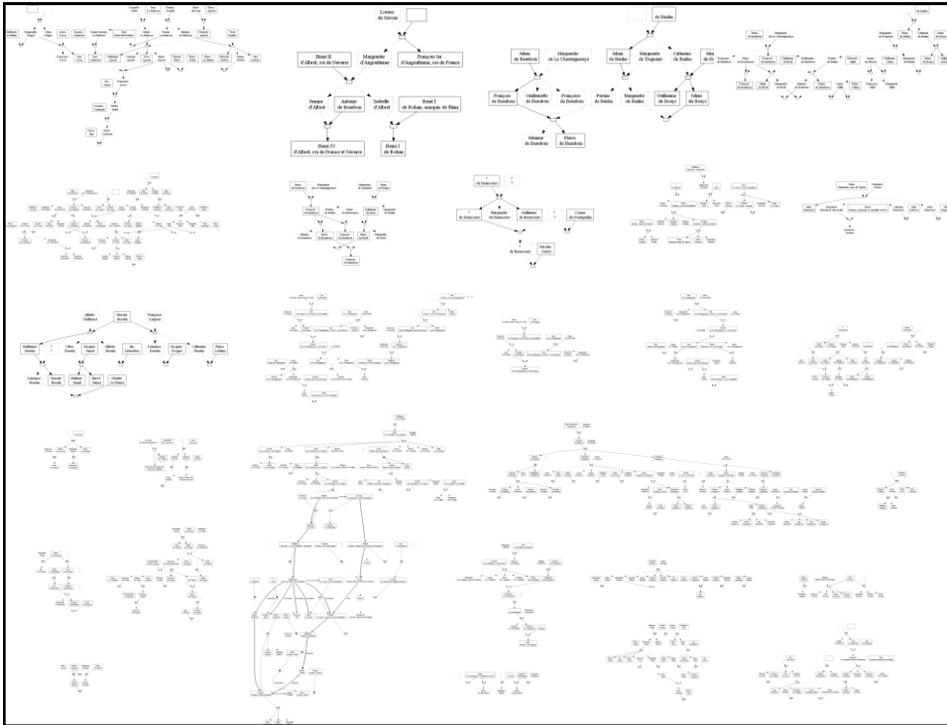
Réseaux sociaux

- Recherches récentes qui améliorent encore la visualisation de réseaux sociaux
 - Nathalie Henry, Jean-Daniel Fekete, and Michael J. McGuffin. 2007. NodeTrix: a Hybrid Visualization of Social Networks. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (November 2007), 1302-1309.



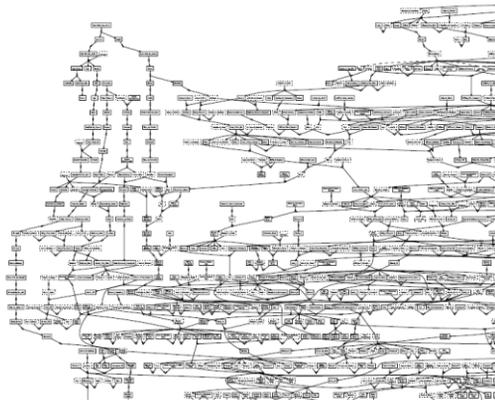
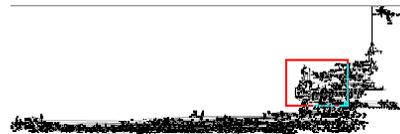
Généalogies





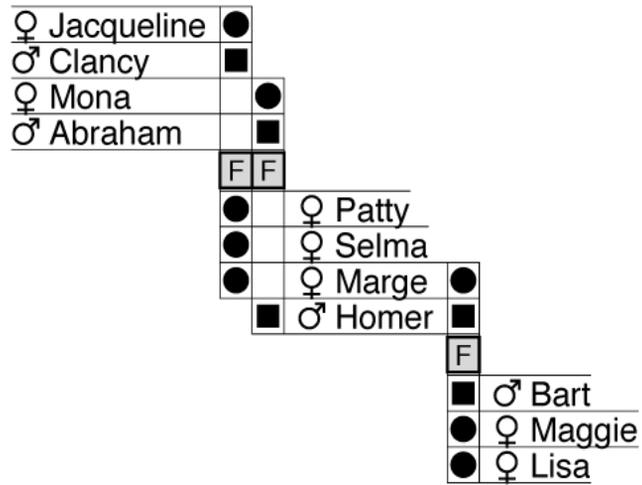
Généalogies

- Généalogie des familles royales européennes
 - 3100 personnes
 - 1422 familles
- Illisible en représentation traditionnelle
- Mais on peut faire mieux



GeneaQuilts

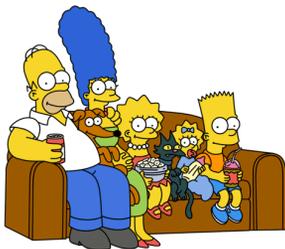
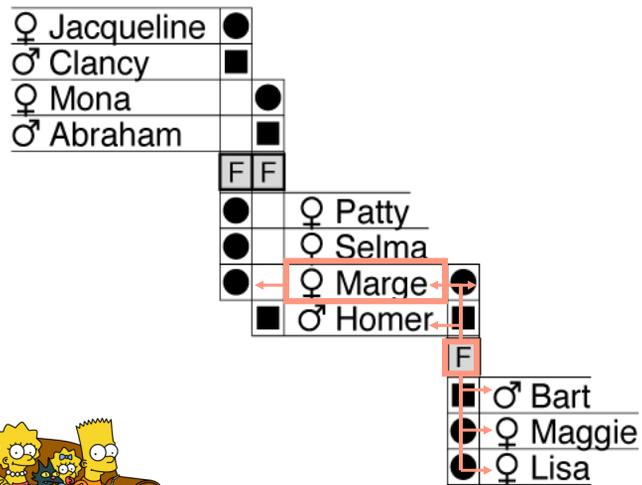
GeneaQuilts



19

GeneaQuilts

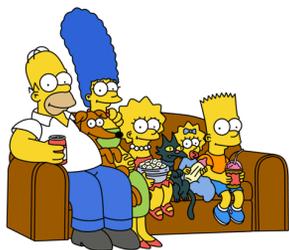
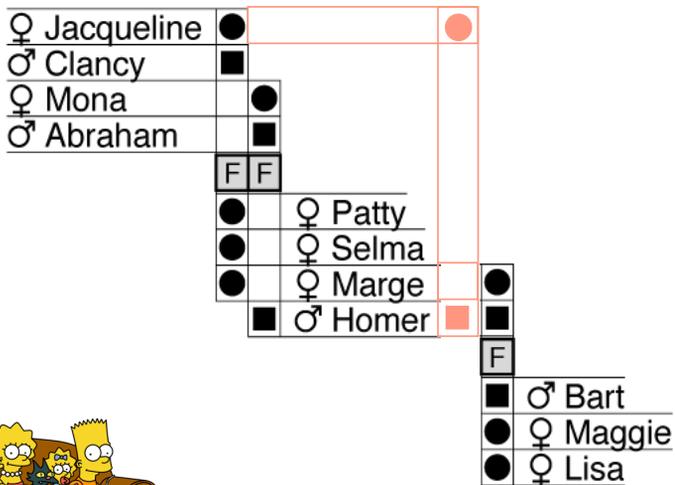
GeneaQuilts



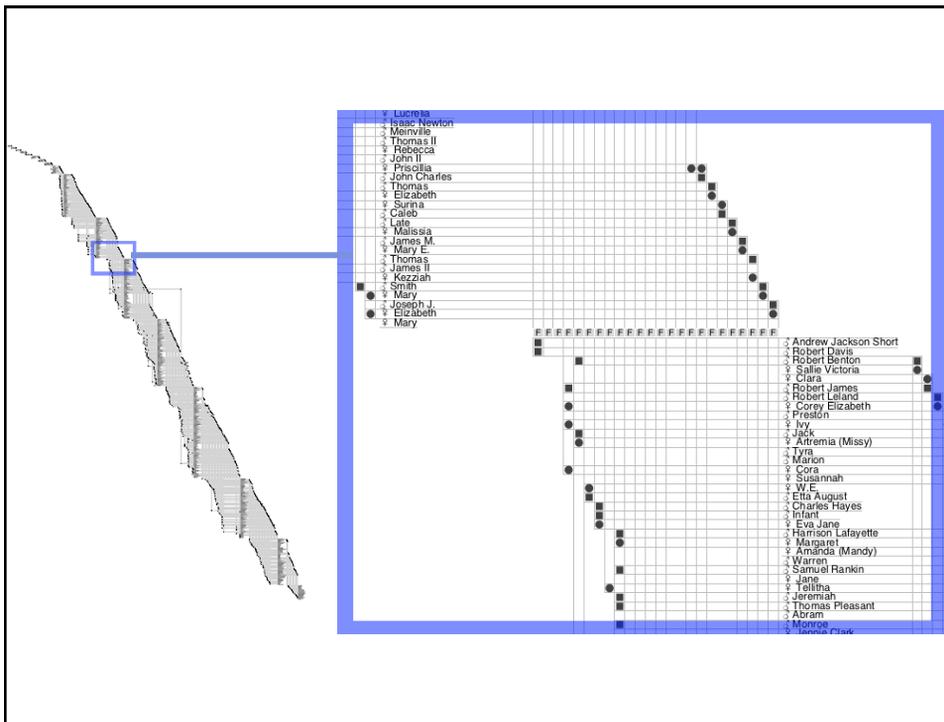
20

GeneaQuilts

GeneaQuilts



21



GeneaQuilts

GeneaQuilts system

Genealogy Quilt: /Users/anastasiab/eclipse-workspace/geneaquilt/data/royal92.ged

File View Edit

534 1000 1500 1992

♂ Ingeborg //
 ♂ Philip II Augustus //
 ♀ Isabella of Hainault //
 ♂ Alfonso IX //
 ♂ Sancho VI //
 ♀ Eleanor of Aquitaine //
 ♀ Constance of Toulouse //
 ♂ Baldwin //
 ♂ Eustace of Boulogne //
 ♂ Maria //
 ♀ Mary of Boulogne //
 ♂ William of Boulogne //
 ♂ Matthew of Alsace //
 ♂ Henry II Curmishie //
 ♂ Geoffrey VI of Anjou //
 ♂ William //
 ♀ Isabel de Warrenne //
 ♂ Aymer of Angouleme /Taillefer//
 ♀ Alice de Courentev //
 ♂ William //
 ♂ William of Gloucester //
 ♂ Henry of Huntingdon //
 ♀ Ada //
 ♂ Hugh /Kerstock//
 ♂ William De Braose//
 ♀ Bertha //
 ♂ Lywelyn Fawr the Great//
 ♀ Unknown //
 ♂ Richard (Strongbow) //
 ♂ Gruffydd //
 ♀ Aoife (Eva) //
 ♂ Urchard //
 ♂ Donnell More //
 ♂ Conan of Brittany //

Search

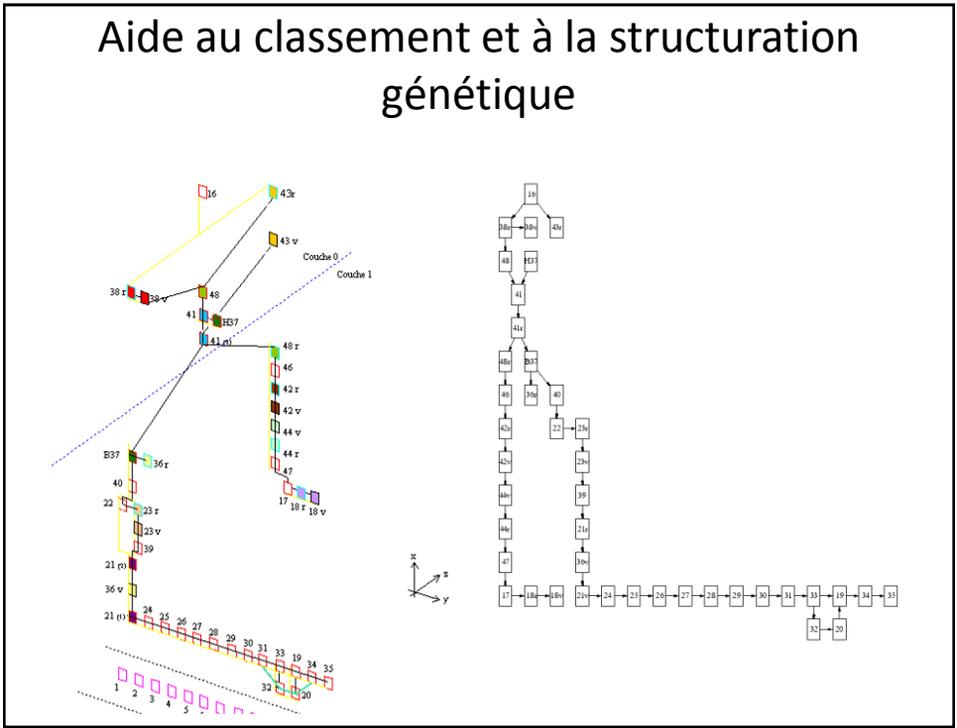
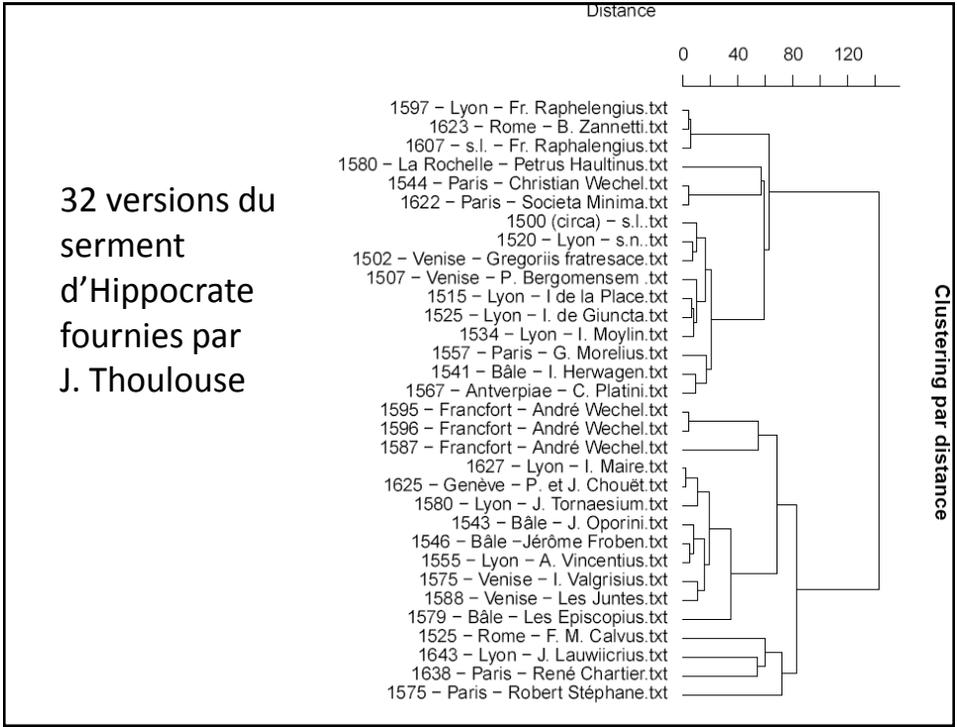
Attribute	Value
BIRT DATE	1130/1131
COMP	0
DEAT DATE	10 AUG 11
DEAT PLAC	Bury St Ed
ID	11400
LAYER	94
NAME	Eustace of
SEX	M
TITL	Count
COMP	0
ID	1532
LAYER	94
NAME	Alfonso IX

Filter

23

Évolution des textes

- Multiples générations de documents
- Génétique textuelle



Conclusion

- Ajoutez des vues d'ensemble à vos ressources numériques
 - Autorisez la sélection/filtrage à partir de ces vues
- Facilite la contextualisation des documents
- Facilite l'attractivité des sites

- Des solutions logicielles très riches existent déjà :
 - thejit.org, protovis.org, [mbostock.github.com/d3/](https://github.com/mbostock/d3),
Europeana4D