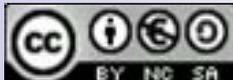




AGENCE NATIONALE DE LA RECHERCHE
ANR

La Textométrie devient Open Source

Serge Heiden, ENS Lyon
slh@ens-lyon.fr



Séminaire CÉDITEC, Créteil, 19 Mars 2010



Projet Textométrie 2007 - 2010



<http://textometry.ens-lsh.fr>

- Partenaires contractants initiaux :
 - Université de Nice Sophia-Antipolis - BCL
 - Université de Besançon - LASELDI
 - Université de Paris 3 - ILPGA
 - ENS-LSH Lyon – ICAR : hébergeant U. of Oxford – OUCS and U. de Montréal - ATO
- Participants au titre de :
 - Concepteurs de logiciels de Lexicométrie / Textométrie
 - Utilisateurs experts en Textométrie (politologie, linguistique, littérature, histoire...)



Concepteurs de logiciels

- **DTM** : **Ludovic Lebart** (développements logiciels innovants dans la suite de composants réalisés pour SPAD et SPAD-T, important logiciel de statistique et d'analyse des données diffusé par la société SPADsoft, Paris : <http://www.spadsoft.com>)
- **HYPERBASE** : **Etienne Brunet** (diffusé en CDROM par l'U. de Nice : <http://ancilla.unice.fr/~brunet/pub/hyperbase.html>)
- **LEXICO** : **André Salem** (diffusion par le web, en shareware : <http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW>)
- **SATO** : **François Daoust** (diffusé par l'UQAM, en CDROM et par le web : <http://www.ling.uqam.ca/sato/outils/sato.htm>)
- **WEBLEX** : **Serge Heiden** (diffusé par l'ENS-LSH : application web à usage académique par projets de recherche : <http://weblex.ens-lsh.fr/doc/weblex.pdf>)
- **XAIRA** : **Lou Burnard** (diffusé par l'OUCS : par le web, en open source : <http://www.xaira.org>)
- **ASTARTEX** : **Jean-Marie Viprey** (outil académique, support d'expérimentation et d'illustration de fonctionnalités textométriques innovantes : http://laseldi.univ-fcomte.fr/document/viprey/page_JMV.htm)
- **TRAMEUR** : **Serge Fleury** : <http://tal.univ-paris3.fr/trameur>

TXM - Objectifs



- Construire une nouvelle plateforme logicielle commune ouverte :
 - une boîte à outils pour les services fondamentaux
 - des applications prototypes :
 - Locale : Windows, Linux, *Mac OS X*
 - Distante : Navigateur Web
- Déploiement aisé pour l'utilisateur final
- Efficace

TXM – Effort principal



- Construire un **modèle conceptuel commun**
 - Unités de base
 - Unités de structures
 - Focus & Sélection
 - Regroupement raisonné de fonctions
- Stabiliser une **terminologie** commune
- S'accorder sur un **niveau de qualité des données d'entrée**

TXM – Modèle courant



- Hiérarchie d'unités ayant d'importe quelle propriétés
- Mécanisme général de sélection d'unités
 - sélection de corpus
 - sélection de focus
- Composant d'entrée scriptable pour la construction aisée de plugins d'importation de tout format : CVS-TXT-XML-XML/TEI...

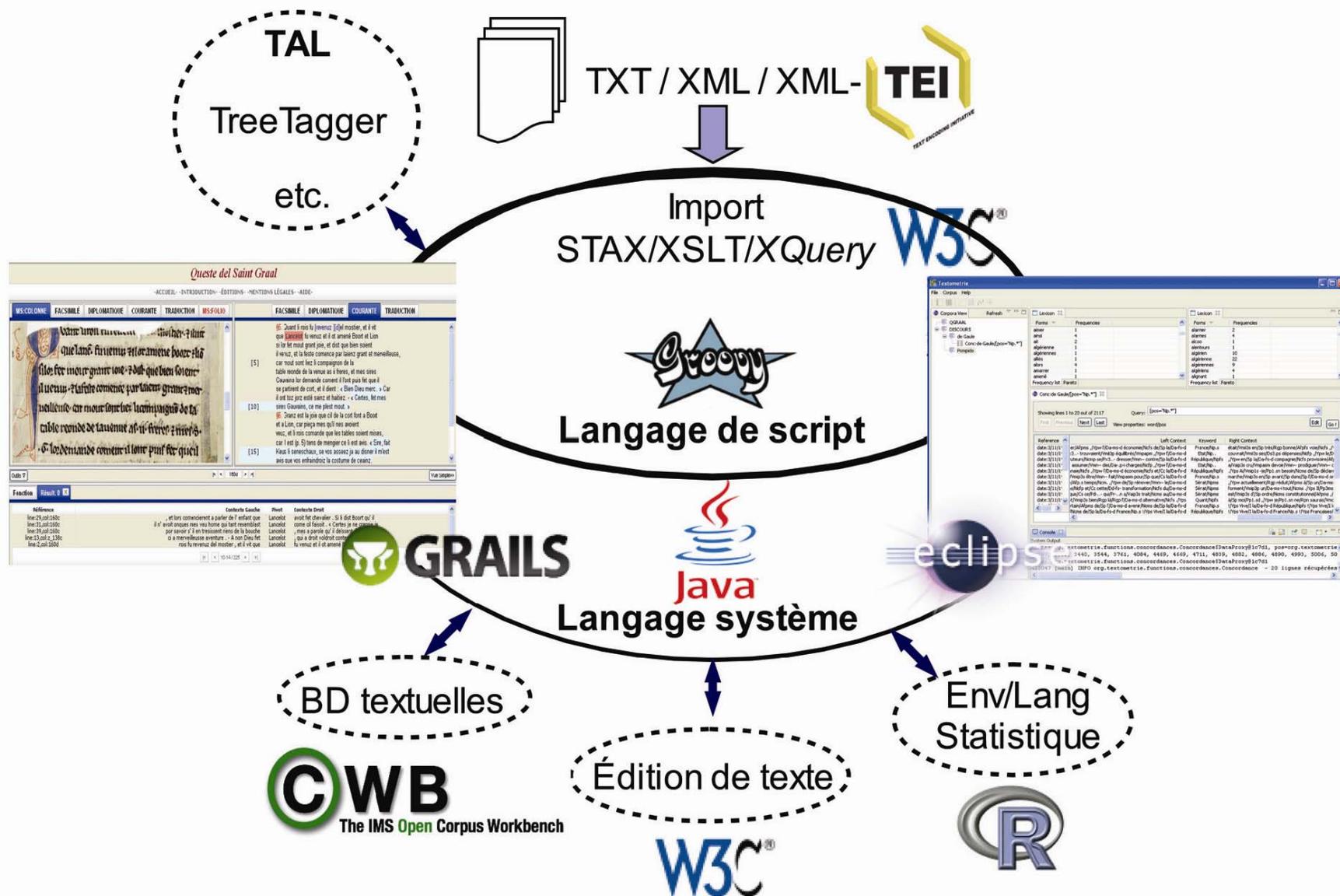
TXM - Choix



- Plateforme basée sur des standards ouverts : **Logiciel Open Source**
 - Interopérabilité
 - Extensibilité
 - Capitalisation
 - Utiliser des composants Open Source
 - IMS CWB
 - Environnement statistique R
 - TAL ouvert...

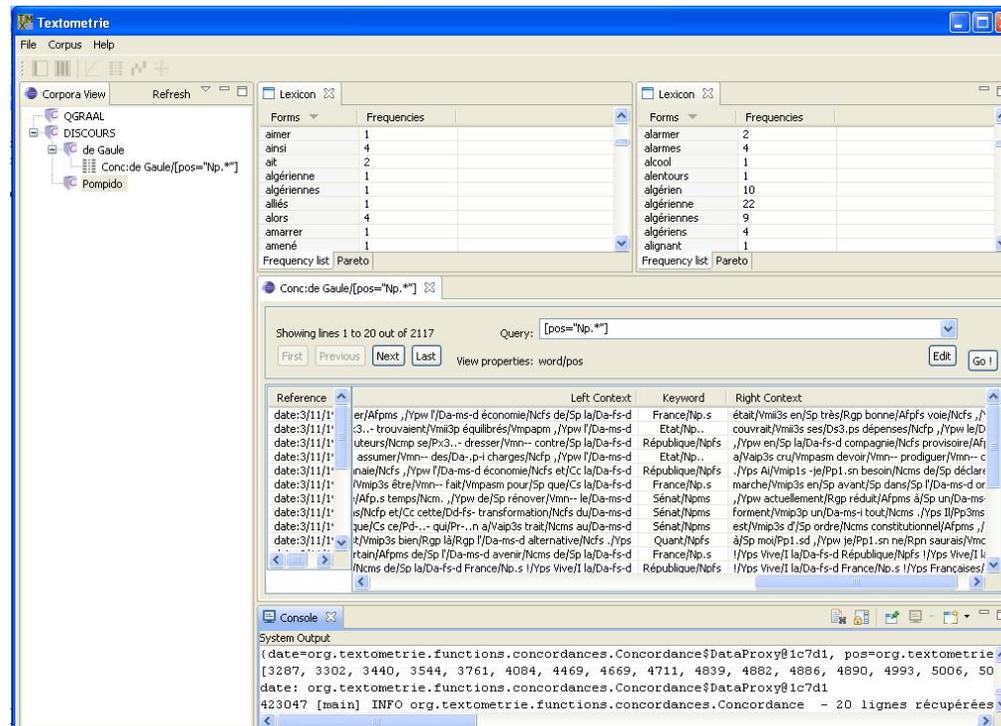
- Compatible XML-TEI : **Corpus Open Source**
 - Interopérabilité
 - Extensibilité
 - Capitalisation
 - Valider sur des corpus existants

TXM - architecture



TXM - application Eclipse RCP

- Windows XP : Setup_0.4.0.exe
- Linux Ubuntu : TXM-0.4.0.tgz



TXM Eclipse RCP application

Textometrie

File Corpus Help

Corpora View Refresh

- QGRAAL
- DISCOURS
 - de Gaule
 - Conc:de Gaule,[pos="Np.*"]
 - Pompidu

Lexicon

Forms	Frequencies
aimer	1
ainsi	4
ait	2
algérienne	1
algériennes	1
alliés	1
alors	4
amarrer	1
amené	1

Frequency list Pareto

Lexicon

Forms	Frequencies
alarmer	2
alarmes	4
alcool	1
alentours	1
algérien	10
algérienne	22
algériennes	9
algériens	4
alignant	1

Frequency list Pareto

Conc:de Gaule,[pos="Np.*"]

Showing lines 1 to 20 out of 2117 Query: [pos="Np.*"]

First Previous Next Last View properties: word/pos Edit Go!

Reference	Left Context	Keyword	Right Context
date:3/11/1'	er/Afpms ,/Ypw l'/Da-ms-d économie/Ncfs de/Sp la/Da-fs-d	France/Np.s	était/Vmii3s en/Sp très/Rgp bonne/Afpfs voie/Ncfs ,/'
date:3/11/1'	:3.- trouvaient/Vmii3p équilibrés/Vmpapm ,/Ypw l'/Da-ms-d	Etat/Np..	couvrait/Vmii3s ses/Ds3.ps dépenses/Ncfp ,/Ypw le/D
date:3/11/1'	uteurs/Ncmp se/Px3.- dresser/Vmn-- contre/Sp la/Da-fs-d	République/Npfs	,/Ypw en/Sp la/Da-fs-d compagnie/Ncfs provisoire/Afj
date:3/11/1'	assumer/Vmn-- des/Da-p-i charges/Ncfp ,/Ypw l'/Da-ms-d	Etat/Np..	a/Vaip3s cru/Vmpasm devoir/Vmn-- prodiguer/Vmn-- c
date:3/11/1'	inaie/Ncfs ,/Ypw l'/Da-ms-d économie/Ncfs et/Cc la/Da-fs-d	République/Npfs	,/Yps Ai/Vmip1s -je/Pp1.sn besoin/Ncms de/Sp déclar
date:3/11/1'	/Vmip3s être/Vmn-- fait/Vmpasm pour/Sp que/Cs la/Da-fs-d	France/Np.s	marche/Vmip3s en/Sp avant/Sp dans/Sp l'/Da-ms-d or
date:3/11/1'	/Afp.s temps/Ncm. ,/Ypw de/Sp rénover/Vmn-- le/Da-ms-d	Sénat/Npms	,/Ypw actuellement/Rgp réduit/Afpms à/Sp un/Da-ms
date:3/11/1'	s/Ncfp et/Cc cette/Dd-fs- transformation/Ncfs du/Da-ms-d	Sénat/Npms	forment/Vmip3p un/Da-ms-i tout/Ncms ./Yps Il/Pp3ms
date:3/11/1'	que/Cs ce/Pd-.- qui/Pr-..n a/Vaip3s trait/Ncms au/Da-ms-d	Sénat/Npms	est/Vmip3s d'/Sp ordre/Ncms constitutionnel/Afpms ,/
date:3/11/1'	t/Vmip3s bien/Rgp là/Rgp l'/Da-ms-d alternative/Ncfs ./Yps	Quant/Npfs	à/Sp moi/Pp1.sd ,/Ypw je/Pp1.sn ne/Rpn saurais/Vmc
date:3/11/1'	rtain/Afpms de/Sp l'/Da-ms-d avenir/Ncms de/Sp la/Da-fs-d	France/Np.s	!/Yps Vive/I la/Da-fs-d République/Npfs !/Yps Vive/I l
date:3/11/1'	/Ncms de/Sp la/Da-fs-d France/Np.s !/Yps Vive/I la/Da-fs-d	République/Npfs	!/Yps Vive/I la/Da-fs-d France/Np.s !/Yps Françaises/

Console

```
System Output
{date=org.textometrie.functions.concordances.Concordance$DataProxy@1c7d1, pos=org.textometrie
[3287, 3302, 3440, 3544, 3761, 4084, 4469, 4669, 4711, 4839, 4882, 4886, 4890, 4993, 5006, 50
date: org.textometrie.functions.concordances.Concordance$DataProxy@1c7d1
423047 [main] INFO org.textometrie.functions.concordances.Concordance - 20 lignes récupérées
```



TXM Eclipse RCP application



algérienne 1
 algériennes 1
 alliés 1
 alors 4
 amarrer 1
 amené 1

aléntours 1
 algérien 10
 algérienne 22
 algériennes 9
 algériens 4
 alignant 1

Conc:de Gaule/[pos="Np.*"]

Showing lines 1 to 20 out of 2117 Query: [pos="Np.*"]

First Previous Next Last View properties: word/pos Edit

Reference	Left Context	Keyword	Right Context
date:3/11/1'	er/Afpms ,/Ypw l'/Da-ms-d économie/Ncfs de/Sp la/Da-fs-d	France/Np.s	était/Vmii3s en/Sp très/Rgp bonne/Afpfs voie
date:3/11/1'	3...- trouvaient/Vmii3p équilibrés/Vmpapm ,/Ypw l'/Da-ms-d	Etat/Np..	couvrait/Vmii3s ses/Ds3.ps dépenses/Ncfs ,/Y
date:3/11/1'	uteurs/Ncmp se/Px3...- dresser/Vmn-- contre/Sp la/Da-fs-d	République/Npfs	,/Ypw en/Sp la/Da-fs-d compagnie/Ncfs provis
date:3/11/1'	assumer/Vmn-- des/Da-.p-i charges/Ncfs ,/Ypw l'/Da-ms-d	Etat/Np..	a/Vaip3s cru/Vmpasm devoir/Vmn-- prodiguer/
date:3/11/1'	inaie/Ncfs ,/Ypw l'/Da-ms-d économie/Ncfs et/Cc la/Da-fs-d	République/Npfs	,/Yps Ai/Vmip1s -je/Pp1.sn besoin/Ncms de/Sp
date:3/11/1'	/Vmip3s être/Vmn-- fait/Vmpasm pour/Sp que/Cs la/Da-fs-d	France/Np.s	marche/Vmip3s en/Sp avant/Sp dans/Sp l'/Da-
date:3/11/1'	/Afp.s temps/Ncm. ,/Ypw de/Sp rénover/Vmn-- le/Da-ms-d	Sénat/Npms	,/Ypw actuellement/Rgp réduit/Afpms à/Sp un
date:3/11/1'	is/Ncfs et/Cc cette/Dd-fs- transformation/Ncfs du/Da-ms-d	Sénat/Npms	forment/Vmip3p un/Da-ms-i tout/Ncms ,/Yps Il
date:3/11/1'	que/Cs ce/Pd-...- qui/Pr-...n a/Vaip3s trait/Ncms au/Da-ms-d	Sénat/Npms	est/Vmip3s d'/Sp ordre/Ncms constitutionnel/A
date:3/11/1'	it/Vmip3s bien/Rgp là/Rgp l'/Da-ms-d alternative/Ncfs ./Yps	Quant/Npfs	à/Sp moi/Pp1.sd ,/Ypw je/Pp1.sn ne/Rpn saur
date:3/11/1'	rtain/Afpms de/Sp l'/Da-ms-d avenir/Ncms de/Sp la/Da-fs-d	France/Np.s	!/Yps Vive/I la/Da-fs-d République/Npfs !/Yps
date:3/11/1'	/Ncms de/Sp la/Da-fs-d France/Np.s !/Yps Vive/I la/Da-fs-d	République/Npfs	!/Yps Vive/I la/Da-fs-d France/Np.s !/Yps Frar

TXM Grails web application



- Groovy / J2EE based + Java TXM toolbox
- YUI & Prototype AJAX javascript toolkits

Queste del Saint Graal

-ACCUEIL- -INTRODUCTION- -ÉDITIONS- -MENTIONS LÉGALES- -AIDE-

MS:COLONNE FACSIMILÉ DIPLOMATIQUE COURANTE TRADUCTION MS:FOLIO FACSIMILÉ DIPLOMATIQUE COURANTE TRADUCTION

The screenshot displays the TXM Grails web application interface. It features a header with the title "Queste del Saint Graal" and navigation links. Below the header, there are tabs for "MS:COLONNE", "FACSIMILÉ", "DIPLOMATIQUE", "COURANTE", "TRADUCTION", and "MS:FOLIO". The main content area is split into two columns: the left column shows a manuscript page with a large blue initial 'B' and text in Old French, and the right column shows the corresponding translation in modern French. The translation includes line numbers [5], [10], and [15] in light blue boxes. At the bottom, there is a search bar with "Fonction" and "Résultat: 0" and a table with columns for "Référence", "Contexte Gauche", "Pivot", and "Contexte Droit".

Référence	Contexte Gauche	Pivot	Contexte Droit
line:29,col:160c line:31,col:160c line:39,col:160c line:13,col:z_138c line:2,col:160d	, et lors comencierent a parler de l' enfant que il n' avoit onques mes veu home qui tant ressemblast por savoir s' il en tresissent riens de la bouche ci a merveilleuse aventure . - A non Dieu fet rois fu revenuz del mostier , et il vit que	Lancelot Lancelot Lancelot Lancelot	avoit fet chevalier . Si li dist Boort qu' il come cil faisoit . « Certes je ne croeroie ja , mes a parole qu' il deissent de ceste chose , qui a droit voldroit conter le terme de cest fu venuz et il ot amené Boort et Lion si

Queste del Saint Graal web edition TXM Grails application



Queste del Saint Graal

-ACCUEIL- -INTRODUCTION- -ÉDITIONS- -MENTIONS LÉGALES- -AIDE-

MS:COLONNE FACSIMILÉ DIPLOMATIQUE COURANTE TRADUCTION MS:FOLIO

FACSIMILÉ DIPLOMATIQUE **COURANTE** TRADUCTION

Vant uiron tureu... mostier. 7 iluz
que lané. fu uenuz 7 il ot amene boort 7 ho
filoz fet mouit grant ioie. 7 dist que bien soient
il uenuz. 7 la feste comence par laienz grant 7 mer-
ueilleuse. car mouit sont liez li compaignoz de la
table reonde de la uenue as .ii. freres. 7 mes .s.
. 6. lor demande coment il l'ont puis fet que il

[5] §5. Quant li rois fu [revenu] [d]el mostier, et il vit
que Lancelot fu venuz et il ot amené Boort et Lion
si lor fet mouit grant joie, et dist que bien soient
il venuz, et la feste comence par laienz grant et merueilleuse,
car mouit sont liez li compaignon de la
table reonde de la uenue as ii freres, et mes sires
Gauvains lor demande coment il l'ont puis fet que il
se partirent de cort, et il dient : « Bien Dieu merci. » Car
il ont toz jorz esté sainz et haitiez. - « Certes, fet mes
sires Gauvains, ce me plest mouit. »

[10] §6. Granz est la joie que cil de la cort font a Boort
et a Lion, car pieça mes qu'il nes avoient
veuz, et li rois comande que les tables soient mises,
car il est (p. 5) tens de mengier ce li est avis. « Sire, fait
Keus li seneschaux, se vos asseez ja au disner il m'est
avis que vos enfreindroiz la costume de ceainz.

Outils ▾ < < 160d > > Vue Simple>>

Fonction Résult. 0 x

Référence	Contexte Gauche	Pivot	Contexte Droit
line:29,col:160c line:31,col:160c line:39,col:160c line:13,col:z_138c line:2,col:160d	, et lors comencierent a parler de l' enfant que il n' avoit onques mes veu home qui tant resemblast por savoir s' il en tresissent riens de la bouche ci a merueilleuse aventure . - A non Dieu fet rois fu revenu del mostier , et il vit que	Lancelot Lancelot Lancelot Lancelot Lancelot	avoit fet chevalier . Si li dist Boort qu' il come cil faisoit . « Certes je ne croie ja , mes a parole qu' il deissent de ceste chose , qui a droit voldroit conter le terme de cest fu venuz et il ot amené Boort et Lion si

< < 10-14 / 225 > >

TXM - documentation ouverte & sites web & wiki



- **Projet Scientifique :**
<http://textometry.ens-lsh.fr>
- **Développeurs informatique :**
http://sourceforge.net/apps/mediawiki/textometry/index.php?title=Main_Page
- **Site logiciel principal SourceForge :**
<http://sourceforge.net/projects/textometrie>



Thank you to partners and sponsors

■ Partners

- Contracting : Nice University, Besançon University, Paris 3 University, Oxford University (UK), Montréal University (CA)
- Chicago University (US) – Philologic/Philomine
- UPR IRHT, UMR CESR, UMS RISC (FR) - corpus
- Cultural Heritage Consortium - Izhevsk (RU) ? - corpus

■ Sponsors

- National French Research Agency (ANR)
- CNRS
- Rhône-Alps region Research Cluster 13 Agency