Développement communautaire de logiciels pour les SHS : l'exemple du logiciel open-source TXM

Serge Heiden

UMR ICAR – ENS de Lyon / CNRS

Colloque n&m's
IFÉ Lyon, 17-18 juin 2013

Plan de l'exposé

- 1. Sciences humaines et sociales
  - Disciplines et démarche expérimentale
  - Observables : représentations textuelles
- 2. Outils : exemple de la plateforme TXM
  - Formats de représentation textuelle
  - Modèle de corpus progressif de TXM
  - Architecture de TXM
- 3. Développement open-source
  - Modèle et stratégie
  - Infrastructure de développement et de diffusion

1. Travaux en Sciences humaines et sociales

Disciplines participant aux ateliers TXM

Linguistique	38	allemande 1, anglaise 1, TAL 1
Géographie	17	
Sciences de l'éducation	11	SVT 1, linguistique 1
Documentation	11	publication 2, édition 2, archives 1
Langues	11	allemand 3, italien 3, russe 2, anglais 1, arabe 1, espagnol 1
Littérature	10	française 6, nord-américaine 1, moderne 1
Histoire	9	contemporaine 1, de l'art 1
Informatique	6	
Sciences de l'information et de la communication	5	

Ateliers TXM sept 2012- juin 2013

Disciplines participant aux ateliers TXM

Psychologie	4	sociale 1, sciences de l'éducation 1
Sociologie	3	des sciences et des techniques 1
Statistiques	3	
Sciences politiques	3	
Philosophie	2	Renaissance et âge classique 1
Sciences économiques	2	
Urbanisme	2	aménagement 1
Sciences de gestion	1	
Sémiotique	1	
Transport et environnement	1	
Études transculturelles	1	

Ateliers TXM sept 2012- juin 2013

Types de corpus (sources) : représentations textuelles

- Textes :
  - Œuvres, articles scientifiques, témoignages
  - Documentation technique, manuels
  - Articles de presse : Factiva, Europresse, archives dépt...
- Autres types de textes :
  - Discours politique : transcription, site web, tracts, débats parlementaires...
  - Forums, pages web, tweet, sms
  - Chansons
- Transcriptions :
  - Entretiens d'enquête
  - Vidéos de cours
- Autres logiciels : Alceste, Hyperbase, Cordial...

2. Exemple d'outil : la plateforme TXM aujourd'hui

Types de corpus de TXM (formats)

- Texte
  1) TXT – Unicode : texte brut tout venant
  2) XML : texte structuré et encodé → plans textuels (hors texte, locuteurs, notes...)
  3) XML-TEI : texte struct. et enc. standard
  4) XML-TEI-TXM : texte enrichi linguistiquement
- Transcription
  - Transcriber : texte synchronisé, locuteurs, évt.
  - *Word : locuteurs, sections*
- Corpus multilingues alignés – relation de traduction (FR // EN, etc)

Modèle de corpus de TXM

- **Unités textuelles** (roman, article…)
  - **Métadonnées** (auteur, date, domaine, genre…)
  - **Structure** interne (phrases, paragraphes, sections…)
  - Plans textuels
    - Notes, discours rapporté, réplique/tour de parole
    - Langue principale (français…), langue (latin…)
    - Hors-texte (commentaires, apparat critique…)
- **Unités lexicales** (formes graphiques, lemme, description morpho-syntaxique…)
- Outils de TAL impliqués (lemmatiseurs…)
- **Édition**
  - Mise en page (sauts de page, disposition)
  - Rendu (styles)
- Références bibliographiques
- Points de synchronisation (transcriptions)
- Alignement (corpus multilingues alignés)

Cahier des charges d'importation

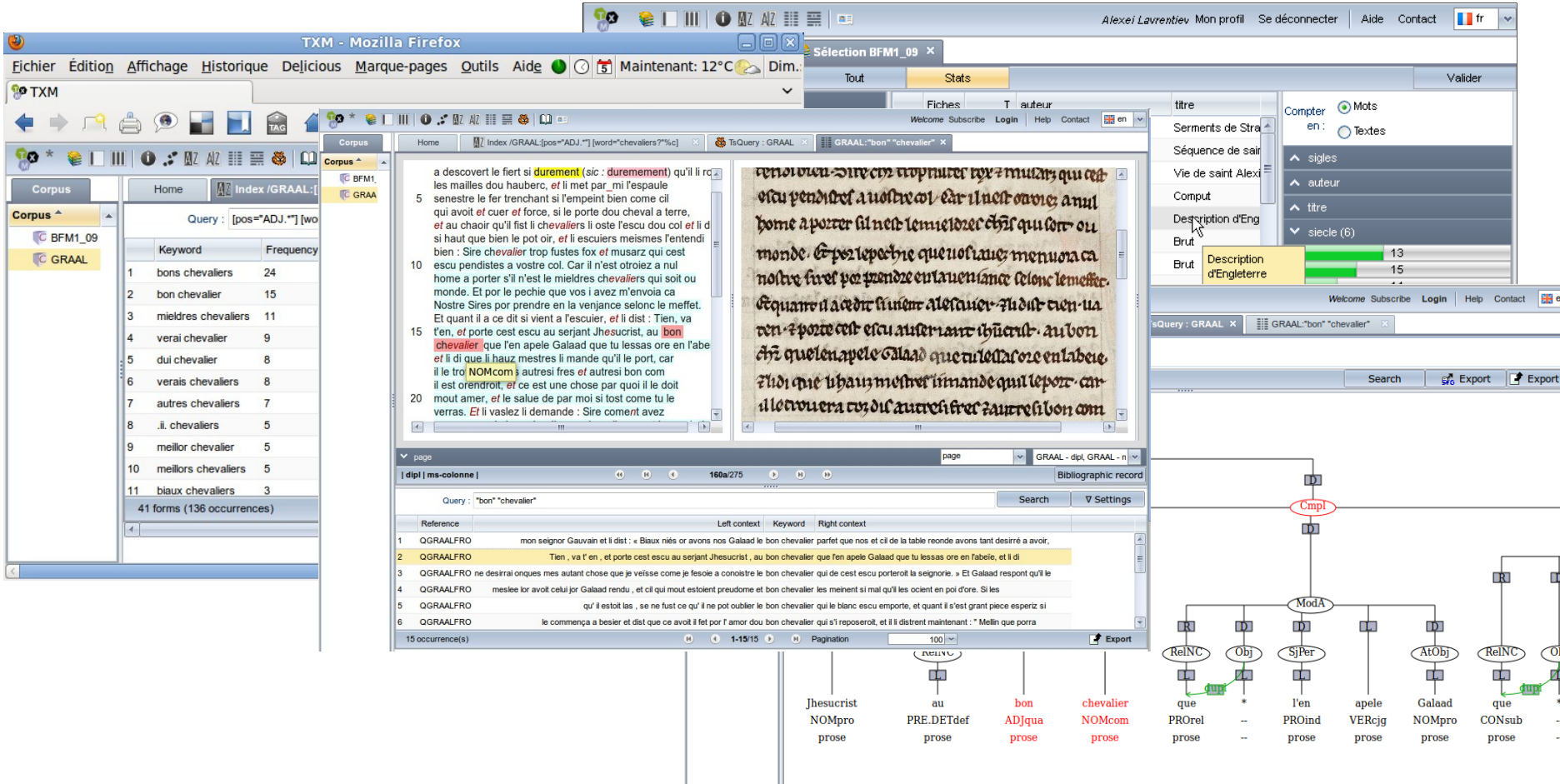
Niveau philologique progressif : niveaux de représentation / modèle de données

	TXT	**XML/w**	**XML-TEI**
Unités textuelles	fichiers	fichiers	fichiers
Métadonnées	Tableau CSV	Tableau CSV	Entête XML teiHeader
Mots + prop.	brut	<w>?	<w>?
Structures	-	Toutes balises	Balises spécifiques
Plans	-	transformations XSL	Balises spécifiques

Architecture de TXM pour gérer le modèle de données progressif

- A. Modules d'import multiples → format TEI TXM pivot
  - TXT+CSV (textes)
  - XML/w+CSV
  - XML-TEI BFM, BVH, Frantext, Perseus, PUC-Revues.org...
  - Transcriber+CSV (transcriptions)
  - TMX (corpus alignés)
- B. Boite à Outils
  - B.1 Moteur de recherche :
    - CQP - langage CQL : unités lexicales + unités de structures
      - création de focus (Tgen) : motifs de séquences
      - gestion de configurations au mot près : sous-corpus, partitions
    - *TigerSearch (syntaxe)*
  - B.2 Moteur statistique R
- D. Interfaces utilisateurs :
  - version bureau TXM
  - version portail TXM en ligne

Interface Graphique Utilisateur : http://portal.textometrie.org/demo



Composants open-source et langages de programmation utilisés dans TXM

- A. Modules d'import XML, scripts et macros : *Groovy* (projet open-source http://groovy.codehaus.org), XSLT (W3C)
- B. Boite à outils et applications : Java (consortium JCP http://jcp.org/en/home/index), IDE Eclipse
  - B1. Moteur de recherche CQP : C CWB (projet open-source http://cwb.sourceforge.net)
  - B2. Moteur statistique : C et R (projet open-source http://www.r-project.org)
    - Algorithme Reinert Iramuteq R
- D. Interfaces utilisateur
  - D.1 version bureau : Framework Eclipse RCP (projet open-source http://wiki.eclipse.org/index.php/Rich_Client_Platform)
  - D.2 version portail : Framework GWT (projet open-source http://code.google.com/intl/fr/webtoolkit)
    - Interface des commandes statistiques (UMR GREYC - PUC)

3. Développement open-source

Modèle de développement logiciel open-source

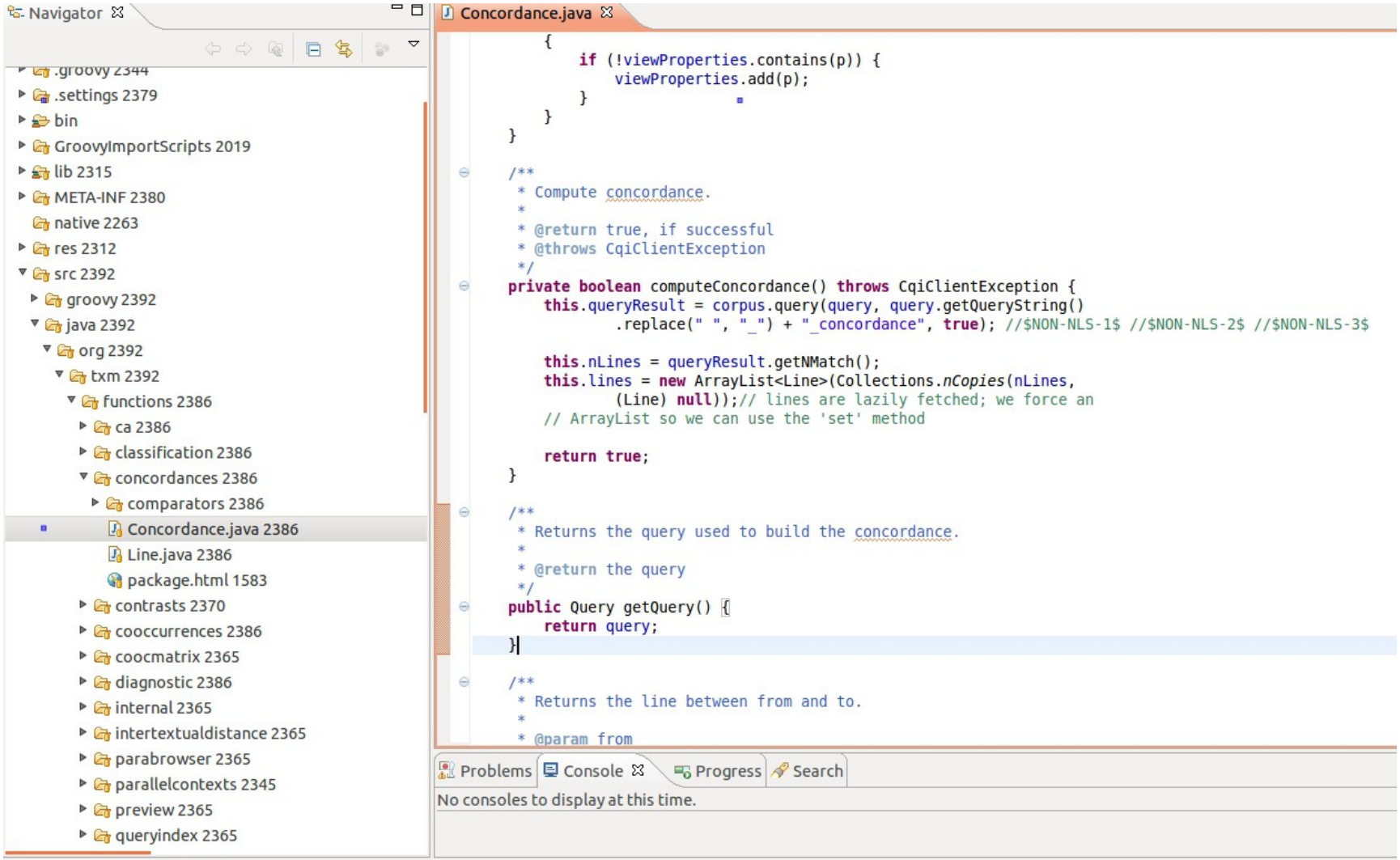
- Co-développement d'outils et de composants, intégration / recyclage de composants, décomposition de l'évolution / maintenance
Exemple Iramuteq / TXM : algorithme Reinert – R / formats
- Capitalisation, pérennisation communautaire (activités par programmes : projets ANR...)
Exemple : projet ANR-DFG PRESTO → lemmatisation
- Accès aux procédés de calcul (sources de logiciels : standard Java)

/ /

Accès aux choix d'encodage (sources de textes : standard XML-TEI)

Scientificité

Exemple de code source Java



Exemple d'encodage XML-TEI(-TXM)

```
207  <lb n="52"/>sa hanste brise par asteles.
208  <lb n="53"/>E Gorm<supplied rend="crochets">un</supplied>d ad l'espee traite,
209  <lb n="54"/>si l'ad feru sur le heaume :
210  <lb n="55"/>la teste en fist voler a destre,<pb n="6"/>
```

```
2103                <lb n="53"/>
2104                <s n="17" id="s_17">
2105                    <w id="w_360">
2106                        <txm:form>E</txm:form>
2107                        <interp resp="#txm" type="#ttpos">CONcoo</interp>
2108                        <interp resp="#txm" type="#ttlemma">-</interp>
2109                    </w>
2110                    <w id="w_361">
2111                        <txm:form>Gorm<supplied rend="crochets">un</supplied>d</txm:form>
2112                        <interp resp="#txm" type="#ttpos">NOMpro</interp>
2113                        <interp resp="#txm" type="#ttlemma">Gormund</interp>
2114                    </w>
2115                    <w id="w_362">
2116                        <txm:form>ad</txm:form>
2117                        <interp resp="#txm" type="#ttpos">VERcjg</interp>
2118                        <interp resp="#txm" type="#ttlemma">-</interp>
2119                    </w>
2120                    <w id="w_363">
2121                        <txm:form>l&amp;apos;</txm:form>
2122                        <interp resp="#txm" type="#ttpos">DETdef</interp>
2123                        <interp resp="#txm" type="#ttlemma">-</interp>
2124                    </w>
```

Modèle de développement logiciel open-source

- Co-développement d'outils et de composants, intégration / recyclage de composants, décomposition de l'évolution / maintenance
Exemple Iramuteq / TXM : algorithme Reinert – R / formats
- Capitalisation, pérennisation communautaire (activités par programmes : projets ANR...)
Exemple : projet ANR-DFG PRESTO → lemmatisation
- Accès aux procédés de calcul (sources de logiciels : standard Java)
/ /
Accès aux choix d'encodage (sources de textes : standard XML-TEI)

Scientificité

Modèle de valorisation du logiciel open-source

- Co-développement, capitalisation communautaire
- Implication des utilisateurs
  - Publications de travaux réalisés avec l'aide de TXM
  - Retours de bugs : logiciel, documentation
  - Propositions d'améliorations : logiciel, documentation
  - Communication sur l'outil, entraide entre utilisateurs
- *Intégration des logiciels, lexiques et corpus dans la production scientifique (groupe corpus-écrits)*

Modèle de diffusion open-source

A) Objets
  (1) Logiciels, applications
  (2) Bibliothèques, scripts, macros
  (3) Documentation
  (4) Corpus, lexiques
B) Déclaration de propriété intellectuelle
  (1) Patrimoniale : institutions (universités, écoles), laboratoires, personnes
  (2) Moral : personnes
C) Déclaration de licence de diffusion
D) Plateformes de diffusion en ligne

Liste des contrats ayant développé la plateforme TXM

1) jan 2007 – déc 2010 : projet ANR Textométrie #ANR-06-CORP-029 - conception et développements initiaux, partenaires impliqués dans TXM : UMR ICAR - **Lyon**, UMR BCL - **Nice**, EA LASELDI - **Besançon**, EA SYLED – **Paris 3**
2) juin – aout 2009 : contrat région Rhône-Alpes Cluster 13 / ENS de Lyon - prototype TXM/Grails pour la mise en ligne de la Queste del Saint Graal
3) sept 2009 – déc 2011 : contrat ANR CORPTEF / ENS de Lyon - développement de TXM en version portail GWTmport XML-Transcriber, GUI pour l'environnement R, partenaires impliqués dans TXM : UMR ICAR - Lyon, UMR **EVS** – Lyon
4) avril 2011 : contrat CNRS (DGLFLF) / UMR ICAR - import et traitements pour le corpus GGHF (Grande Grammaire Historique du Français)
5) juin – juil 2012 : contrat ANR-DFG SRCMF / ENS de Lyon - conception du module Tiger Search, import et concordances syntaxiques
6) juin – juil 2012 : contrat Equipex Matrice / Univ. Paris 1 - développements de la plateforme pour l'infrastructure nationale pour les historiens, partenaires **Paris 1**...
7) juin 2013 – déc 2015 : contrat ANR-DFG PRESTO - développement de la préparation de corpus (TAL), partenaires impliqués Univ. **Cologne**
8) Sept 2013 – août 2015 : contrat Labex ASLAN, Lyon – développement portail, RCP

Détermination de la Propriété Intellectuelle

- Brevets ?
- Accords de consortium (ANR ?)

- Services juridiques d'établissements ?
- Services de valorisation : PRES de Lyon
- Incubateurs d'entreprises, INPI ?

Licences de diffusion open-source

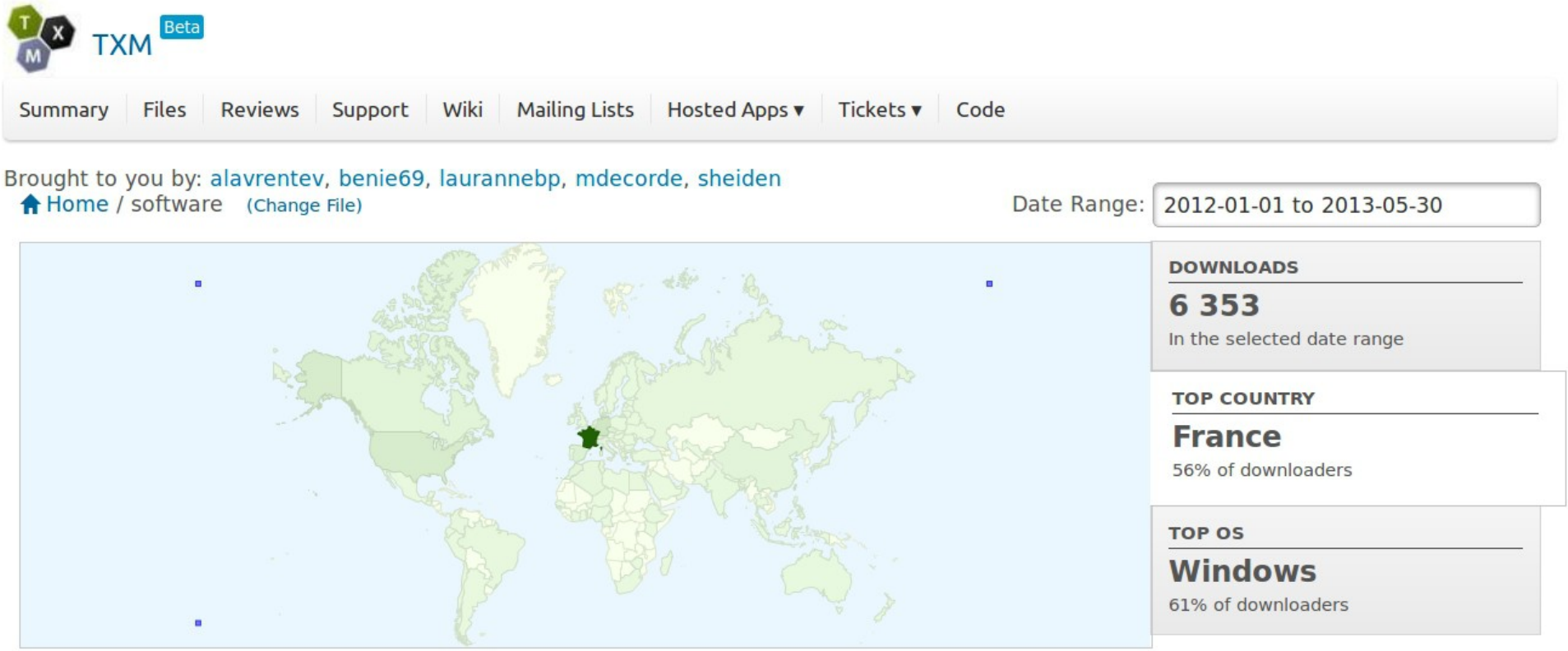
- Logiciel
  - Type GNU Public Licence (GPL v 3.0)
    - Citation
    - Rétrocession de sources améliorées
    - Licence de diffusion identique
  - Type BSD
    - Citation
- Corpus, Lexique & Documentation
  - Type Creative Commons : BY-NC-SA
  - Type LGPPL (Unitex)

Diffusion de la plateforme TXM : http://sourceforge.net/projects/txm

- 3 Applications de bureau (pour travail sur poste local) GPL v3 :
  - version Windows XP, Windows Vista et Windows 7 (en 32 et 64bits)
  - version Mac OS X (32+64 bits)
  - version Linux Ubuntu (en 32 et 64bits)
- 1 Application portail web pour serveur (pour l'accès aux corpus en ligne) GPL v3
- Corpus exemples CC-BY-NC
- Documentations CC-BY-NC

Connaître ses utilisateurs (1/3)
Téléchargements mensuels jan 2012 – mai 2013

Connaître ses utilisateurs (2/3)
Téléchargements mensuels jan 2012 – mai 2013



Connaître ses utilisateurs (3/3) tél. mens.

Country	Linux	Mac	Win	Total
France	17%	14%	42%	3,616
Germany	15%	3%	58%	423
United States	4%	20%	51%	319
Spain	3%	10%	20%	206
Belgium	8%	23%	50%	173
Czech Republic	2%	3%	65%	159
Canada	12%	24%	46%	116
China	14%	0%	55%	111
Italy	8%	22%	39%	106
Morocco	5%	2%	46%	100
Tunisia	0%	4%	73%	70
Switzerl	10%	46%	34%	63
United Kingdom	15%	13%	48%	46
India	0%	5%	51%	41
United Arab Emirates	0%	0%	100%	40
Egypt	3%	0%	94%	33
Netherlan	6%	3%	68%	31
Colombia	48%	6%	19%	31
Brazil	3%	13%	60%	30
Greece	0%	0%	72%	29
Finland	0%	5%	91%	22
Hungary	82%	0%	14%	22
Australia	15%	5%	50%	20
Turkey	16%	5%	37%	19

Infrastructures (1/2) : besoins

- Hébergement pour le développement : SVN, trackers, wiki, liste de diffusion
  - Projets de développement informatique
  - Projets de constitution de corpus
- Hébergement pour la diffusion des logiciels, lexiques et corpus
  - Assistance juridique pour la CPI
  - Assistance juridique pour les licences
  - Assistance pour la mise en ligne 24/7
  - Assistance pour la valorisation

Infrastructures (2/2) : aujourd'hui

- Nationales
  - Laboratoires : UMR, EA
  - Sites personnels
  - Centres de ressources
    - CNRTL (ATILF-BVH)
    - Telma (IRHT-ENC)
  - TGE Adonis + TGIR, Corpus = HumaNum ?
- Internationales
  - CLARIN, DARIAH, OTA
  - ELRA / ELDA ?

Merci !