# Expérimentation de text-mining (TDM) au sein du LaDéHiS-CRH (CNRS UMR 8558/ EHESS)

Francine Filoche, Laura Pages

▶ **To cite this version:**

Francine Filoche, Laura Pages. Expérimentation de text-mining (TDM) au sein du LaDéHiS-CRH (CNRS UMR 8558/ EHESS) . Journée d'étude InSHS - mercredi 7 décembre 2016 – au siège du CNRS à Paris Thème : Rôle des professionnels de l'IST pour offrir une meilleure visibilité aux travaux des acteurs de la recherche en SHS., Journée d'étude InSHS - CNRS à Paris, Dec 2016, Paris, France. halshs-01780688

**HAL Id: halshs-01780688**
**https://shs.hal.science/halshs-01780688v1**

Submitted on 25 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

**17 octobre 2016**

**Journée d'étude InSHS - mercredi 7 décembre 2016 – au siège du CNRS à Paris**
**Thème : Rôle des professionnels de l'IST pour offrir une meilleure visibilité aux travaux des acteurs de la recherche en SHS.**

**Titre :** Expérimentation de text-mining (TDM) au sein du LaDéHiS-CRH (CNRS/EHESS) : faciliter la recherche d'information

**Intervenants :**

Francine Filoche a construit son parcours professionnel dans plusieurs domaines de la BAP F, au sein de l'École des Hautes Études en Sciences Sociales (EHESS) : l'édition en SHS, les collections patrimoniales au sein de deux bibliothèques de recherche et d'une UMR en histoire. Depuis quelques années s'est ajouté le domaine de l'IST à l'échelle d'un laboratoire (LaDéHiS-CRH). Elle est depuis 2014, chargée de mission services au public, de la co-construction des futurs services, de la bibliothèque du Campus Condorcet.

Laura Pagès a suivi des études d'économie à Sciences-Po Lyon et à l'Université Paris-Dauphine avant de devenir conservateur d'État des bibliothèques. Très attachée au développement du numérique en bibliothèque, elle a consacré son mémoire Enssib à l'étude des bibliothèques sans livres imprimés. Elle est depuis juillet 2016 chargée de mission ressources et innovation numériques au Campus Condorcet.

**Descriptif :**

L'expérimentation de fouille de textes qui est menée depuis janvier 2016 par le campus Condorcet est née de la rencontre entre deux besoins. Le LaDéHiS cherchait à améliorer et à systématiser les processus d'extraction et de traitement de l'information dans le cadre de ses travaux de recherche. Le Grand équipement documentaire souhaitait quant à lui expérimenter sur le terrain un service de préfiguration innovant, fondé sur la manipulation de documents numériques. Il s'agissait pour les deux partenaires, chercheurs et bibliothécaires, de se laisser la possibilité d'expérimenter une technologie numérique innovante appliquant à des corpus de texte numérisés des techniques analytiques puissantes permettant de dégager de nouvelles connaissances.

Une phase préliminaire a été consacrée à la préparation du corpus-source proposé par le chercheur. Chaque unité documentaire a ainsi été nettoyée, formatée et correctement nommée pour pouvoir tester les outils de fouille de texte selon une suite logique et progressive. L'expérimentation s'est ensuite déroulée en trois grandes étapes. La première étape dite de classement a permis de hiérarchiser et de classer automatiquement les documents composant le corpus en utilisant le logiciel OntoGen[1]. Ensuite, l'étape « sémantique » a permis d'élaborer plusieurs scénarios de concepts pouvant intéresser le chercheur en exploitant les potentialités du logiciel Tropes[2]. Enfin, grâce au logiciel Calliope[3], l'étape d' « indexation automatisée » a permis de constituer un lexique de référence respectant les orientations scientifiques du chercheur.

À chaque étape, les interactions avec le chercheur ont joué un rôle clé dans le bon déroulé de cette expérimentation et ont permis d'obtenir, à partir du corpus-source confié, de nouveaux matériaux de recherche qu'il n'aurait pas été possible d'obtenir en explorant les données textuelles à la main.

[1] http://ontogen.ijs.si/
[2] http://tropes.fr/
[3] https://www.calliope-textmining.com/

**Mots clés :**

Données de la recherche, Droit d'auteur, Éditeur, Exploration de données, Information scientifique et technique, Licence, Open data, Open Science, Propriété intellectuelle, Recherche, TDM (Text and Data Mining), Loi Numérique

**Pour aller plus loin:**

- Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique.
- *Liberating Data: How libraries and librarians can help researchers with TDM*, London School of Economics Library, 2016.
- Atelier du LIBER et de la British Library : «...The right to read is the right to mine »
- 2014 (Nov). Actualisation du rapport de Kenneth Crews sur les exceptions au droit d'auteur en faveur des bibliothèques réalisé en 2008
- *TDM Report from the Expert Group*. Commission européenne (DG Research and Innovation), 2014
- *Study on the legal framework of text and data mining (TDM)*, Jean-Paul Triaille, Jérôme de Meeüs d'Argenteuil et Amélie de Francquen, Commission européenne (DG Internal Market and services, 2014.