K-WORDS LAB : un outil d'analyse des mots clés permettant d'explorer les dynamiques d'un domaine scientifique.

Audrey Baneyx*, Philippe Breucker**

*Université Paris-Est, IFRIS, LATTS UMR CNRS 8134, France
baneyx@ifris.org
**INRA Sens, Université Paris-Est, IFRIS, France
breucker@ifris.org

1 Introduction

L'utilisation par les chercheurs des Sciences Humaines et Sociales (SHS) d'outils de fouille de données, d'extraction et d'analyse des connaissances est aujourd'hui une nécessité pour comprendre et étudier les dynamiques de certains domaines scientifiques (Turenne et Barbier, 2004). C'est dans ce contexte que s'inscrit le développement de l'outil K-WORDS LAB au sein de la plateforme CorTexT[1] de l'Institut francilien « Recherche, Innovation et Société ». K-WORDS LAB est un outil construit pour analyser les dynamiques qui témoignent de l'émergence d'un nouveau domaine scientifique. Nous pensons qu'il est possible de suivre les dynamiques qui se forment au sein d'un domaine scientifique en constitution en analysant sa production (projets, articles scientifiques, brevets, ...) (Baneyx et al., 2010).

2 Matériel et méthodologie

K-Words Lab est un outil basé sur la technologie php5/MySql5. L'interface web est écrite en php/HTML/CSS et certaines parties en flash (technologie Flex) et JavaScript. La méthodologie mise en œuvre dans K-WORDS LAB mobilise les apports du traitement automatique de la langue, de l'intelligence artificielle, des statistiques et de la sociologie des sciences. L'outil propose une chaîne de traitement automatique qui fait appel à plusieurs modules et va de l'analyse du corpus à la représentation dynamique et interactive des clusters de concepts qui caractérisent le domaine étudié, ainsi que son évolution.
(1) **Le module d'importation** analyse les différents types de corpus (notices bibliographiques, notices de dépôt de brevets, ...), les stocke et les indexe dans la base de données.
(2) **Le module linguistique** a pour objectif de réduire le nombre d'occurrences de mots clés[2] aux mots clés distincts compris dans le corpus.
(3) **Le module de clustering** implémente un clustering linguistique et un clustering contextuel. Le clustering linguistique regroupe les mots clés ayant des formes orthographiques proches les

[1] http ://www.cortext.fr
[2] Nous avons pour l'instant testé jusqu'à 10 000 000 occurrences de mots clés.