Outil de Veille Dynamique sur

l'Attente Sociétale en matière de Recherche et d'Innovation

Vademecum

Service de la Stratégie de la Recherche et de l'Innovation

Secteur Sciences et Société

David Chavalarias — Jean-Philippe Cointet — Camille Roth

1 À propos de cette plateforme

Cette plateforme présente une reconstruction des discussions sur le thème “Santé et Environnement” à partir d’un échantillon représentatif de contenus dynamiques du web français sur la période du 03 juillet 2010 au 16 janvier 2011.

Par reconstruction, il faut entendre l’extraction, à partir d’une analyse des productions de la blogosphère, de motifs temporels qui mettent en saillance certains sujets, en explicitent les modalités de production et permettent de suivre leur évolution.

Cet échantillon comporte 110 188 documents (billets de blogs, articles de journaux électroniques, flux RSS d’institutions) provenant de 2 304 sources sélectionnées pour leur coloration thématique. La liste de ces sources est accessible à partir du lien “SOURCES” qui figure dans le menu de navigation de la plateforme. Plusieurs pages du site proposent d’étendre la recherche de contenu à l’ensemble du web via l’icône “Google”.

Le périmètre thématique associé à “Santé et Environnement” a été défini en utilisant des outils de fouille de texte à partir d’une analyse des documents. Il a été élargi au fur et à mesure de la veille et est actuellement constitué de 2 476 termes. Cette liste est mise à jour régulièrement pour suivre l’actualité. La liste de ces termes est accessible à partir du lien “TERMES” qui figure dans le menu de navigation.

2 Ontologie de l’étude

Cette section décrit les concepts fondamentaux sur lesquels l’étude s’appuie.

2.1 Le Web en tant que réseau tripartite

Nous appréhendons spécifiquement le *Web* comme un réseau dynamique d’acteurs *et* de thématiques ; les documents textuels que l’on y trouve présentent dans leur grande majorité les descripteurs suivants (cf. Fig. 1) :

- Une ***source*** de contenu, représentant un acteur de notre domaine d’étude et pouvant représenter un individu ou un collectif (par exemple “`pedagotice.blogspot.com`”),
- Un ***document*** (billets de blogs, articles de journaux) écrit par une source et caractérisé par les termes qu’il contient ;
- Des ***termes***, qui désignent ici des groupes nominaux constitués de un ou plusieurs mots ; par exemple, “*IRD*”, “*industrie nucléaire*” ou bien “*gaz à effet de serre*”. Les termes que nous avons retenus sont à la fois parmi les plus signifiants et/ou les plus typiques du corpus étudié ;
- Des ***liens*** vers d’autres acteurs, vers d’autres sources ;
- Une ***date*** de publication.

Ces données et cette ontologie définissent un multi-réseau dynamique sur l’ensemble des documents, multi-réseau tripartite car constitué des trois sous-structures suivantes : (i) le ***réseau sémantique***, dans lequel des phrases-clés ou termes sont liés en fonction de leur co-apparition dans le document (données de type “co-occurrence”), (ii) le ***réseau socio-sémantique*** des relations entre les acteurs et les termes qu’ils manipulent, et (iii)

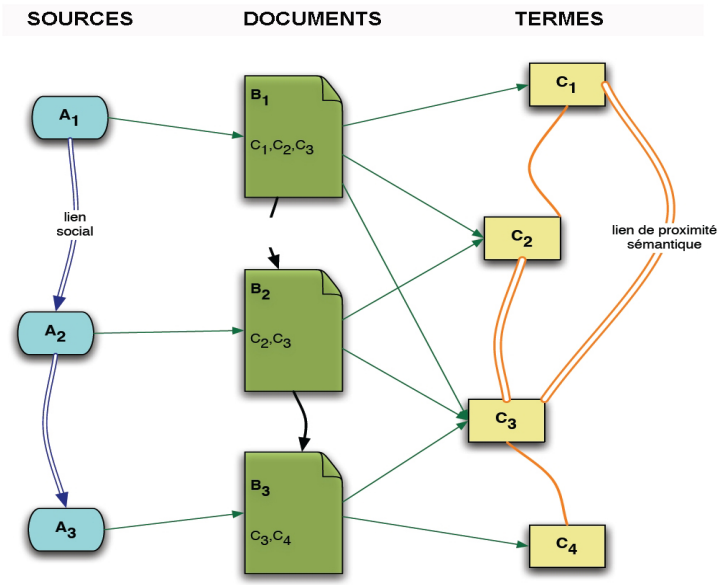

FIGURE 1 – *Description des données de base. L'unité textuelle de base : le billet : B, met en relation au sein d'un réseau socio-sémantique des acteurs (A) et des concepts (C), qui se retrouvent eux-mêmes liés au sein d'un réseau social et sémantique.*

le ***réseau social*** des acteurs, obtenu en s'appuyant sur les hyperliens que ces acteurs établissent vers d'autres acteurs, par le biais des documents et des contenus qu'ils citent.

2.2 Notions fondamentales

Niveau *micro*, *sources* et *termes* À partir de ces données, nous distinguons ensuite trois niveaux de description de la structure dynamique des conversations du web. Le niveau le plus bas est le niveau *micro* des termes et des sources. Les termes et les sources que nous avons retenus sont à la fois parmi les plus fréquents, les plus signifiants et/ou les plus typiques du corpus étudié. Les relations sémantiques entre termes se structurent selon une "topologie de la connaissance" que les étapes ultérieures visent à révéler et représenter. Les relations de citation entre sources se structurent selon une "topologie sociale" que les étapes ultérieures visent à révéler et représenter.

Comme sur la plupart des plateformes de ce type, il est possible d'obtenir le profil et les productions d'une source donnée (par exemple `pedagotice.blogspot.com`) ; ainsi que le profil d'un terme (evolution, billets associés, etc.), comme par exemple "rwe".

Niveau *meso* - *les champs thématiques* L'étude des relations de proximité dans le corpus de termes nous permet ensuite de définir un niveau à la fois intermédiaire et fondamental, celui du champ thématique. Nous supposons en effet que l'unité signifiante fondamentale n'est pas le terme, mais un ensemble de termes qui se contextualisent les uns les autres. Le champ thématique est ainsi constitué d'un ensemble de termes apparaissant de manière conjointe et récurrente dans le contenu des billets étudiés. Par exemple, *{changement climatique, Copenhague, crise climatique, crise environnementale, enjeux de*

Copenhague, fiscalité écologique} et *{agriculture, azote, bassins versants, nitrates, pollution des eaux }* représentent tous deux des champs thématiques du fait d'une forte proximité de ces termes les uns vis-à-vis des autres, proximité qui est directement liée à la co-apparition fréquente des termes en question au sein de notre corpus de billets.

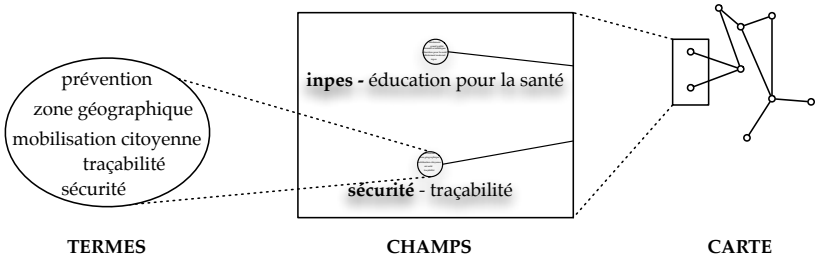

FIGURE 2 – *Ontologie des différents niveaux d'observation : les **cartes générales** regroupent des **champs** constitués de **termes**.*

Les champs peuvent être caractérisés par un nombre très variable de termes qui en donnent la coloration thématique. Nous choisissons alors d'étiqueter chaque champ à l'aide de leurs deux termes les plus représentatifs. Pour des raisons de lisibilité, ceci nous permet en effet de désigner les champs thématiques par une ***double étiquette*** constituée d'une paire de termes plutôt que l'ensemble de ses termes.

Les champs thématiques sont liés à un ensemble de contenus provenant des sources. Cette recherche de contenu peut-être étendue à tous le web depuis la page de profil d'un champ thématique via l'icône Google . La liste des champs thématiques est accessible via le lien "CHAMPS THÉMATIQUES" dans le menu de navigation.

Niveau *macro* - *les cartes* Les cartes (niveau 'macro') montrent l'articulation des champs thématiques à une période donnée, la proximité des problématiques abordées. Pour chaque période, une carte représente les champs thématiques sous forme de réseau dont chaque nœud est un champ repéré par sa double étiquette. Cette carte est spatialisée de manière à faire apparaître les champs proches sémantiquement à proximité "géographique" les uns des autres. Les cartes sont accessibles via le lien CARTES dans le menu de navigation.

Dans l'exemple précédent, le premier champ est ainsi désigné par "***sécurité*** – *traçabilité*", le second par "***inpes*** – *éducation pour la santé*", où la double étiquette correspond au couple de termes "**le plus générique** – le plus spécifique".

Niveau *macro-dynamique* - *les fils thématiques* Le thème couvert par un ***champ thématique*** donné à une période donnée évolue généralement au cours du temps. Par exemple, le scandale du Mediator a d'abord été un scandale sanitaire pour évoluer vers un scandale financier (médicament remboursé par la sécurité sociale) et politique (les décideurs étaient au courant). Ceci est représenté par le fait que des champs thématiques de coloration légèrement différente apparaissant à des périodes distinctes pourront être liés par

un *fil thématique* permettant d'explorer la mutation des débats au cours du temps (cf. fig. 3).

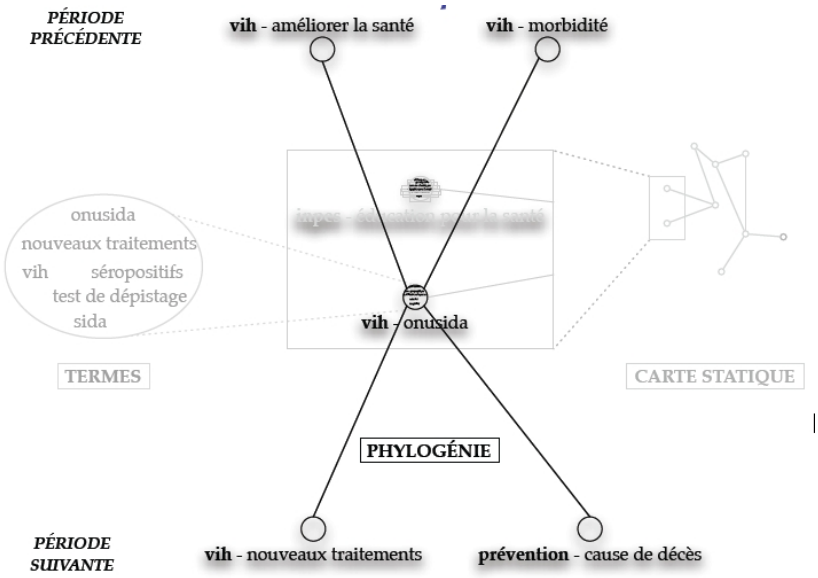

FIGURE 3 – *Les champs thématiques sont intégrés au sein de fils thématiques qui permettent de suivre leurs evolutions et leurs mutations. Au sein d'un fil thématique, un champ thématique peut provenir d'une évolution linéaire ou être le fruit de la rencontre de plusieurs thématiques antérieures. Il peut également donner naissance à une ou plusieurs champs thématiques sur les périodes ultérieures.*

Le champ thématique constitue ainsi le niveau de référence statique — en tant qu'agrégat de termes et unité de base des cartes — et dynamique — en tant que forme cristallisée d'une controverse à un moment donné.

On peut alors comparer les fils thématiques à des colliers dont les perles seraient les champs thématiques, chaque perle étant caractérisée par sa couleur thématique (les termes qui constituent les champs thématiques).

À cet égard, la liste des fils thématiques est accessible via le lien "FILS THÉMATIQUES dans le menu de navigation.

Les fils thématiques attirent l'attention de la blogosphère de manière variable au cours du temps. Leur popularité évolue. Il est ainsi possible de représenter l'évolution de la popularité de l'ensemble des fils thématiques par un graphique "alluvial" (cf. fig. 9). Le scandale du Mediator, ainsi, a connu plusieurs rebondissements depuis son amorce (début septembre sur la fenêtre d'observation, cf. fig. 5).

3 Exploration à l'aide des fils thématiques

Contrairement aux explorations s'appuyant sur des termes choisis a priori ayant pour entrée le niveau 'micro', les entrées par les niveaux 'meso' et 'macro' permettent de sélectionner des sujets à partir de leur profil d'évolution et donc de suggérer des parcours et

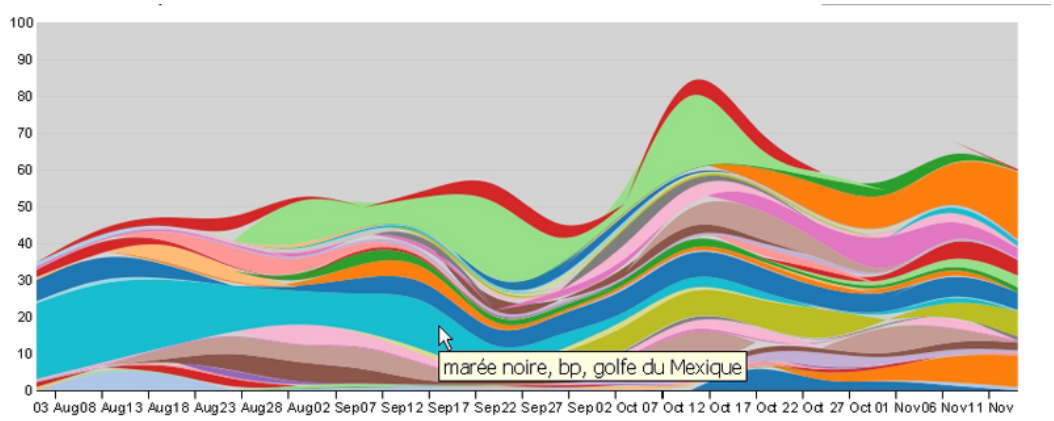

FIGURE 4 – *Exemple de visualisation d'un ensemble de fils thématiques. Ceux-ci intègrent les champs thématiques dans une structure temporelle qui permet de suivre l'évolution de leur popularité. Chaque couleur représente un fil thématique, l'épaisseur des tubes étant proportionnelle à la popularité de la thématique concernée sur la période correspondante.*

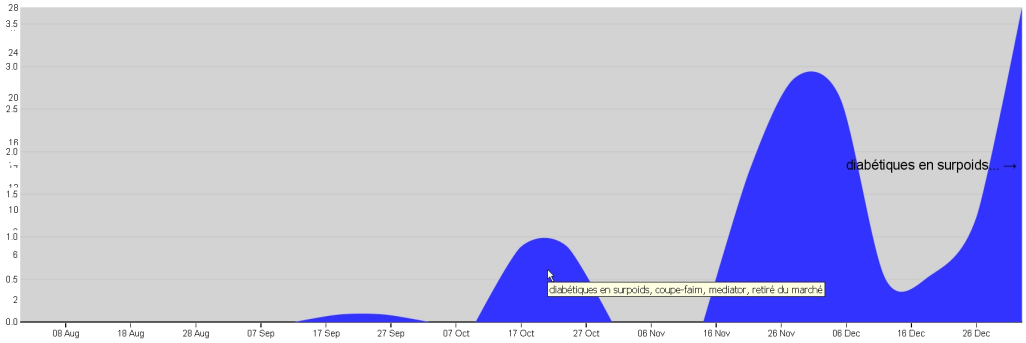

FIGURE 5 – *Exemple d'évolution du fil thématique autour du scandale du médiator. On voit nettement une amorce en septembre.*

des sujets à étudier. Les fils thématiques ont des scores, représentés par des étoiles qui indiquent le degré d'attention maximal que la blogosphère leur a accordé.

La liste des fils thématiques permet d'accéder aux champs thématiques (cf. fig. 6).

FIGURE 6 – *Extrait de la liste des champs thématiques.*

- Un clic sur le score d'un fil thématique permet de se rendre directement à la période où celui-ci a suscité le plus d'activité (cf. fig. 7),
- Un clic sur le nom d'un fil thématique donne des informations sur le fil thématique

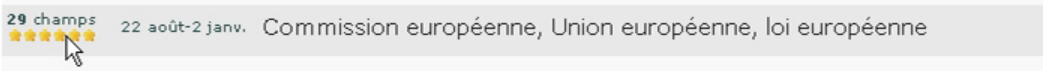

FIGURE 7

(nombre de champs, périodes couvertes, nombre de termes utilisés) et fournit des liens vers des périodes remarquables du fil thématique, sa période de popularité maximale ainsi que sa période la plus récente (cf. fig. 8).

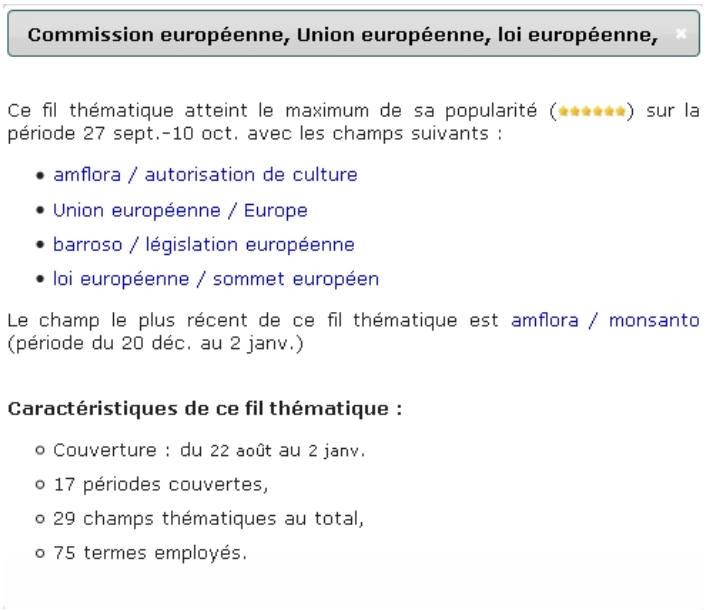

FIGURE 8

Lors du parcours des champs thématiques, le nom du fil thématique auquel celui-ci est rattaché est rappelé en dessous du nom de celui-ci et est également cliquable afin d'obtenir des informations (cf. fig. 9).

FIGURE 9