Veille sociologique et flux d'informations numériques

Francis Chateauraynaud, Josquin Debaz

Veille sociologique et flux d'informations numériques

Une expérience de chronique automatique sur les thèmes sanitaires et environnementaux

Francis Chateauraynaud et Josquin Debaz
GSPR (EHESS)
chateau@msh-paris.fr debaz@ehess.fr

Texte pour l'atelier
« veille numérique, au carrefour des sciences »
EGC 2009

Décembre 2008

Depuis plus d'une dizaine d'années, les travaux de sociologie sur les alertes et les controverses publiques se sont accompagnés de développements informatiques spécifiques. Le principal objectif de ces travaux est d'outiller les enquêtes, en permettant de lier le suivi de grands dossiers, notamment des dossiers sanitaires, environnementaux ou technologiques (amiante, nucléaire, OGM, nanotechnologies, pesticides, champs électro-magnétiques), et la modélisation des formes de l'action et du jugement utilisées par de multiples protagonistes. Il ne s'agit pas de développer une sociologie des discours ou des textes mais bien de considérer les ensembles discursifs comme des lieux de cristallisation de processus sociaux complexes marqués par une profonde réflexivité des acteurs, puisque ceux qui s'imposent sont aussi ceux qui concentrent le plus d'expertise – ou de contre-expertise – sur les dossiers en question[1]. Dans le cadre de ces travaux, pour faire face à l'accumulation documentaire et aux changements de phases, comme lorsque se produit un emballement politique ou médiatique sur un sujet, il est décisif de disposer de protocoles de « veille ». La fonction de ces protocoles est d'assurer dans le même mouvement la mise à jour des bases documentaires et l'identification des changements de régimes argumentatifs, l'émergence de nouveaux acteurs ou d'événements et de prises de parole susceptibles de modifier les éléments centraux d'un dossier.

A travers la description d'un prototype de chroniqueur automatique, ce texte aborde différents problèmes posés par l'automatisation d'une veille sociologique sur des séries textuelles évolutives : quelles sources utiliser ? Comment éviter à la fois dispersion et redondance ? Quels outils statistiques et sémantiques mettre en oeuvre pour faire parler les séries collectées ? A quelles séries comparables, quels fonds documentaires ou quelles encyclopédies de connaissances doit-on rapporter les documents analysés pour caractériser leur contenu et surtout leur apport du point de vue pragmatique ? Quelles opérations évaluatives mettre en place et quelle place accorder à l'utilisateur, et donc à l'interprète dans la boucle ? Une veille non supervisée a-t-elle un sens ? Le produit d'un outil de veille est-il équivalent à un agrégateur sémantique ? Enfin, quel mode d'archivage et de réinterrogation doit-on se donner ? Vastes questions ! Dans cette courte contribution, nous

[1] F. Chateauraynaud, Prospéro : Une technologie littéraire pour les sciences humaines, Paris, CNRS, 2003 ; F. Chateauraynaud, « *Moteurs de (la) recherche et pragmatique de l'enquête. Les sciences sociales face au Web connexionniste* », in L'Historien face à l'ordre informatique, Matériaux pour l'histoire de notre temps, n°82, avril-juin 2006, (revue de la BDIC), p. 109-118.

proposons d'alimenter la discussion en examinant les pistes et les problèmes suscités par la réalisation d'un agent chroniqueur dédié au fil santé-environnement de nos travaux[2].

Suivre l'évolution de grands dossiers : entre acteurs hétérogènes et séries homogènes

Le programme socio-informatique développé depuis plusieurs années se place à la croisée de deux tendances lourdes : d'une part le déploiement de la « société de l'information » dont une des dimensions repose sur une conception coopérative et collective des artefacts cognitifs[3] ; de l'autre, les transformations massives de ce que l'on appelle la « société du risque »[4]. En moins de dix ans, la problématique des risques a changé complètement de nature et de forme. La période des crises, amorcée en France avec l'affaire du sang contaminé qui se déploie véritablement dans l'espace public en 1991, atteint une forme de point d'acmé avec l'enchaînement consécutif de quatre grands dossiers : l'amiante qui est de retour après quinze ans de « silence », le dossier nucléaire qui défraye la chronique avec des alertes à La Hague puis les dix ans de Tchernobyl (1996), la crise de la vache folle puis l'arrivée des premiers OGM en France – arrivée dénoncée par Greenpeace qui intercepte une cargaison en provenance des États-Unis à l'automne 1996. Tous ces dossiers ont un point commun : les sciences et les techniques y jouent un rôle décisif, et se pose de manière cruciale la question de la validité des expertises. C'est aussi au cours de l'année 1996 que se généralisent les références au « principe de précaution » déjà issu d'une longue série de travaux et de discussions, et en particulier le Sommet de Rio de 1992. Au tournant du siècle, les sources d'alerte et les événements dramatiques saturent les arènes politico-médiatiques, avec en outre une extension de la question des risques au terrorisme sous toutes ses formes. On enregistre, dans les interventions publiques, la présence de plus en plus forte des opérations de mise en série, pointant vers l'idée que l'on est entré dans une ère d'instabilité et d'incertitude accrues. Selon les contextes, se trouvent associés, selon des figures variables, des événements comme l'incendie du tunnel du Mont-Blanc (1999), l'explosion de l'usine d'AZF (septembre 2001), le crash du Concorde (juillet 2000), la crise de la vache folle (1996-2000), celle du SRAS (2003) puis de la grippe aviaire (2005-2007), la montée de l'alerte globale sur le climat et l'occurrence du « big one » avec le tsunami de décembre 2004, ou dans un autre registre les attentats du 11 septembre 2001 puis ceux de Madrid (mars 2004) ou de Londres (juillet 2005). Ici ou là, des ponts s'effondrent, des avions s'écrasent, des déchets toxiques se répandent, des ferries se renversent, des incendies font rage, des inondations dévastent des régions entières. Même ce qui est destiné à guérir, à traiter ou à remplacer (médicaments, produits de substitution, innovations technologiques, dont les fameux nanomatériaux) engendre méfiance et inquiétude, alerte et polémique. On note ainsi de nombreuses affaires de « retrait de produit », qu'il s'agisse de cocktails thérapeutiques ou de jouets pour enfants. Aucun milieu de vie et domaine d'activité ne semble épargné, de sorte que tout signe précurseur qui parvient à un degré suffisant de visibilité publique produit une vive agitation qui engendre à peu près toujours la même configuration : des acteurs annoncent l'imminence de la catastrophe, les journalistes convoquent des experts qui, généralement, ne sont pas d'accord, les pouvoirs publics et les industriels se déclarent vigilants et mettent en place des comités ou des commissions permettant de juguler le danger. C'est dans l'organisation collective de ces dispositifs que la problématique de la « veille » et des « signaux faibles » est convoquée[5].

Dans cette nouvelle configuration, la contribution de l'internet à la prolifération des signes, des annonces, des événements, des débats publics et des mobilisations est massive. Au silence et aux rapports de pouvoirs hiérarchiques de la période antérieure s'oppose une problématique de la prolifération et de l'hétérogénéité des jeux d'acteurs et d'arguments dont les réseaux s'entremêlent à l'infini. Pour s'affranchir des effets

---

[2] Une première description de ce chroniqueur est fournie dans A. Bertrand, F. Chateauraynaud et D. Torny, Expérimentation d'un observatoire informatisé de veille sociologique à partir du cas des pesticides, rapport de la convention GSPR / AFSSET, octobre 2007.

[3] Voir G. Bowker, L. Star, W.A Turner, L. Gasser (eds.), *Social Science, Technical Systems and Cooperative Work: Beyond the Great Divide*, New Jersey: Lawrence Erlbaum Associates, 1997 ; T. Malsch, "Naming the Unnamable: Socionics or the Sociological Turn of/to Distributed Artificial Intelligence" draft sent by the author, september 1999.

[4] B. Adam, U.Beck and J.Van Loon (eds), The Risk Society and Beyond, Sage, 2000.

[5] Voir sur ce point F. Chateauraynaud, «Visionnaires à rebours. Des signaux faibles à la convergence de séries invisibles », Document du GSPR, décembre 2007. En ligne sur http://gspr.ext.free.fr/

produits par l'évolution continue des nœuds et des liens, il faut se doter d'*outils alternatifs*, capables de ramener les séries pertinentes dans un laboratoire afin de les faire parler en rompant le cycle infernal de la navigation sans fin. Il existe toutes sortes d'outils pour traquer et tracer les sources et les liens, et notamment les outils de webcrawling[6]. Mais l'ordre cartographique qui tend à s'imposer avec le web, tout en fournissant de nouveaux outils de totalisation, nous éloigne de la saisie des récits et des arguments qui est au cœur des démarches d'enquête et d'érudition en sciences humaines et sociales, et des efforts de modélisation qui permettent de doter les outils d'un minimum de cohérence et de robustesse. En l'absence d'outil d'évaluation des contextes et des niveaux d'information, comment décider du choix des sites ou des blogs à visiter en priorité ? En fonction de leur position relative dans le jeu des intercitations ou des interconnexions ? On retrouve ici les critiques faites à Google et au système de classement des pages.

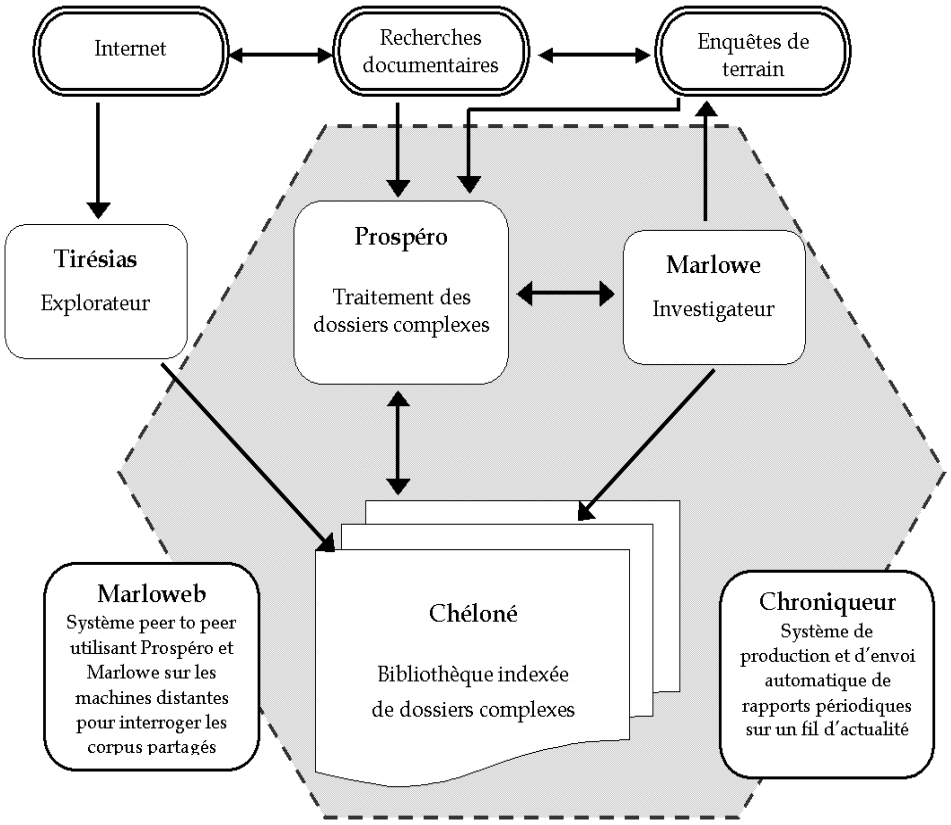
Rééquiper le laboratoire sociologique face à la prolifération des causes

Dans la figure idéale, la sociologie pragmatique suit sur la longue durée l'évolution de dossiers d'alertes, de controverses, d'affaires ou de conflits en examinant systématiquement les transformations des configurations d'acteurs et d'arguments. On raisonne à partir de la confrontation d'un double espace de variations : une série de dossiers (nucléaire, OGM, nanotechnologies, climat, etc.) ; un ensemble de formes de mobilisation et d'argumentation (affaires, débats publics, modes de protestation, procédures d'enquête et d'expertise, modes de « gouvernance », etc.). C'est ce qui se noue et se dénoue au croisement de ces deux plans qui fait l'objet de nos réflexions théoriques. Pourquoi la mobilisation est-elle si radicale sur tel dossier ? Comment la forme instituée du débat public peut-elle modifier des rapports de forces et de légitimités sur tel autre ? On postule que les confrontations d'acteurs et d'arguments sont créatrices d'espaces de possibles à partir desquels s'infléchissent et se réfléchissent les trajectoires des objets. Par exemple, une discussion sur l'énergie nucléaire produit une nouvelle hiérarchisation des jeux d'acteurs et d'arguments et, sans pour autant désigner un vainqueur définitif, permet de caractériser à la fois un rapport de forces et une configuration discursive dominante – dans une période donnée, on parle par exemple beaucoup plus des énergies que des risques sanitaires ou des conditions de sécurité dans les centrales. Lorsque l'on suit de tels processus, la vitesse de reprise, d'influence, de modification et de redistribution des points de vue et des coups que permet le Web, surtout depuis la généralisation du haut débit, change complètement les modalités de l'enquête. Du même coup, la tendance du chercheur est de suspendre la circulation et la prise en compte de multiples sources hétérogènes, pour se concentrer sur des sources plus stables, dotées de références ou de garanties de fiabilité, et généralement placées elles-mêmes en position de centre de tri ou de réduction des informations pertinentes sur un dossier. Ainsi, pour réduire les opérations visant à discerner l'émergence de nouvelles alertes dans le domaine de la santé environnementale, on a été conduit à se concentrer sur le site du Journal de l'Environnement, lequel a rapidement fait apparaître des biais liés aux préoccupations des acteurs porteurs du site. Ces biais ne sont apparus que parce que l'on pouvait croiser systématiquement des sources en transformant leurs informations en corpus comparables.

Pour parvenir à suivre et à analyser en profondeur les multiples corpus évolutifs qui intéressent notre sociologie, on s'est doté d'une architecture dans laquelle opèrent plusieurs entités logicielles. Sans redéployer ici les attendus de ce dispositif socio-informatique, on peut en rappeler le schéma général. Il s'agit de lier quatre fonctions décisives : la structuration des informations (ici des propriétés de grands corpus) ; l'analyse de ces informations (la mise en perspective de jeux d'acteurs et d'arguments à travers des textes et des énoncés, des classes d'objets et des formules) ; la mise à jour et le suivi de sources déterminées sur le Web (exploration de sites et de fils d'informations) ; et enfin l'archivage et la construction d'une mémoire dynamique qui sert de socle de référence pour le repérage et la caractérisation des nouvelles configurations et des processus émergents, ce qui permet de décliner le mot d'ordre selon lequel il n'y a pas

[6] Voir le modèle des « issue crawlers », qui renvoient les cartes de liens entre les sites ou les pages qui traitent des mêmes sujets sur la Toile, mais qui n'ont pas de corpus de référence ni de cadre d'analyse explicite. Voir le « Issue Crawler project » sur http://issuecrawler.net. N. Marres, "Tracing the trajectories of issues, and their democratic deficits, on the Web: The case of the Development Gateway and its doubles", Information Technology & People, 2004, Vol. 17, 2, p. 124 - 149

de veille sans mémoire[7]. L'implémentation de ces fonctions a pris corps dans quatre supports informatiques, formant la suite Prospéro, Marlowe, Tirésias et Chéloné.



Le logiciel Marlowe, qui nous intéresse plus particulièrement ici, est né au cours de l'été 1999. D'abord conçu comme un sous-programme du logiciel Prospéro, destiné à l'activation de fonctions spécialisées, il a pris au fil du temps de plus en plus d'autonomie, notamment depuis l'année 2003, au cours de laquelle a eu

[7] La question de la pérennité des sources est assez cruciale avec les évolutions rapides du Web. On lit des études assez pessimistes sur la capacité du Web à engendrer des archives durables et fiables. Voir W. Koehler, « Web page change and persistence - a four-year longitudinal study», Journal of the American Society for Information Science and Technology, v.53 n.2, p.162-171, January 15, 2002. La question des outils d'archivage était une des préoccupations majeures du dernier sommet de la société de l'information : "As increasing amounts of information find their ways into the Internet's archives, it is vital that we preserve their accessibility, renderability and interpretability. Digital documents often need to be interpreted by specific software packages to be rendered in understandable form. We will need to assure that the bits we preserve on digital media can also be read and understood not only by people but by computers programmed to help us manage this ocean of information. Steps are needed to assure that the information we accumulate today will be usable not merely decades but centuries and even millennia into the future. We need to preserve access to application software, operating systems and perhaps even hardware or simulators so as to retain the ability to make effective use of our digital archives". (Vint Cerf (Chairman, ICANN), "Governance of the Internet: the tasks ahead", Internet Governance Forum ;Athens, Greece, October 30, 2006).

lieu une performance publique mémorable[8]. A partir de 2004, Marlowe (acronyme MRLW) est intégré dans un réseau de recherches en sociologie et, tout en contribuant aux enquêtes, poursuit son « apprentissage ». Plusieurs textes ont déjà évoqué les premiers pas et les présupposés socio-logiques de ce personnage virtuel voué à affronter à la fois d'importantes turbulences dans le champ des sciences sociales contemporaines et des mutations considérables dans les modes de traitement de l'information[9]. En tant qu'enquêteur virtuel ou sociologue électronique Marlowe a besoin d'un réseau de chercheurs et de logiciels avec lesquels il peut développer pleinement une philosophie du dialogisme fondée sur la mise à l'épreuve constante de propositions qui n'ont pas besoin d'être consensuelles pour produire des connaissances[10]. Dans la panoplie des outils de Marlowe, la réalisation de chroniques occupe une place importante et c'est sur cet aspect que nous allons nous concentrer dans la suite du texte.

Les chroniques du logiciel Marlowe

Comme d'autres formes utilisées par Marlowe, telles que la citation, la définition, l'éphéméride, le proverbe, le résumé, l'anecdote, le récit d'expérience, la biographie ou la nécrologie, la chronique est un genre issu d'une longue histoire lettrée, bien décrite par l'histoire des technologies intellectuelles. Dès la fin du III^{e} millénaire, commence un travail d'archivage des oracles et des événements marquants auxquels ils renvoient. Parmi les érudits, l'idée s'impose alors que l'on peut connaître le passé et que l'avenir est prévisible. Vers 2200, sous le règne de Naram-Sîn d'Akkadé, des scribes composent une chronique royale, la première du genre selon Glassner[11]. La chronique a alors pour but de manipuler le passé afin d'expliquer le présent et de légitimer du même coup le pouvoir du souverain. Le genre de la chronique a ainsi une origine profonde, liée à la systématisation des usages de l'écrit et à l'apparition de classes de lettrés, qui vont progressivement s'attacher à rassembler les savoirs et organiser la mémoire collective, en créant les premières bibliothèques [12]. Evidemment, l'imprimerie jouera par la suite un rôle décisif dans les transformations du statut de l'écrit et du type d'autorité qui sera conféré à la faculté de croiser et de recouper les sources [13]. En Occident, après l'émancipation des lettrés de l'emprise de l'Eglise catholique, des chroniques de plus en plus spécialisées verront ainsi le jour. La naissance des revues savantes est le produit de ce long processus de spécialisation des savoirs et des formes d'érudition. En suivant l'évolution de la relation critique aux technologies de pouvoirs-savoirs, le genre des chroniques s'est ainsi déployé tout au long des siècles. La standardisation du genre semble parachevée avec la professionnalisation de la presse écrite et radiophonique, mais l'exercice du chroniqueur ne cesse de changer de format et d'amplitude, comme le montre l'avènement des blogs, qui lui assure une nouvelle apogée.

La fonction de chroniqueur vient donc se ranger à la suite d'une longue tradition, fortement marquée par le genre critique et polémique[14]. De nos jours, le ton d'une chronique est celui d'un commentaire enlevé jouant de multiples ressorts rhétoriques tout en gardant une prise relativement ferme sur un événement dont le chroniqueur cherche à mettre en variation la véritable signification. De ce point de vue la chronique

---

[8] http://prospero.dyndns.org:9673/prospero/acces_public/02_textes_sur_prospero/id_info_performance_marlowe/id_performance_marlowe

[9] Voir notamment F. Chateauraynaud, « Marlowe - Vers un générateur d'expériences de pensée sur des dossiers complexes », Bulletin de Méthodologie Sociologique, n° 79, juillet 2003.

[10] Sur la philosophie dialogique de la connaissance, voir M. Beller, Quantum Dialogue. The Making of a Revolution, Chicago, The University of Chicago Press, 1999.

[11] Etudiant la chronographie en Mésopotamie antique, Glassner a montré comment l'organisation de l'écriture cunéiforme a pris appui sur une organisation politique, laquelle, pour produire de l'ordre, doit se doter d'interprètes, les devins, capables de lire dans les objets et les signes ce qui a trait au comportement humain - les dieux s'exerçant à distribuer les indices sur toutes sortes de supports. Voir les Chroniques mésopotamiennes, présentées et traduites par J.-J. Glassner (Paris, Les Belles lettres, 1993).

[12] Voir C. Jacob, « Rassembler la mémoire. Réflexions sur l'histoire des bibliothèques », Diogène, 2001/4 - n° 196, p. 53 à 76.

[13] E. Eisenstein, La révolution de l'imprimé à l'aube de l'Europe moderne, Paris, La Découverte, 1991.

[14] Voir M. Angenot, Dialogues de sourds. Traité de rhétorique antilogique, Paris, Mille et une nuits, 2008.

entretient un rapport intime avec la veille : soumis à une contrainte d'actualité, le genre est associé à l'idée d'une sélection percutante de traits ou d'aspects qui permettent une mise en perspective d'un événement ou d'un discours et son insertion dans une série historique, fut-elle rappelée à des fins ironiques. La chronique sert d'outil de veille collective en ce qu'il rend saillants les rapports entre des séries passées, des événements ou des propos présents et des séries à venir dont elle vise à cerner les potentialités. De ce point de vue c'est un dispositif performatif qui pèse sur les représentations que se font ses lecteurs ou ses auditeurs des tendances, des continuités ou des ruptures. En dépit de quelques traits d'humour stylisés et faciles à programmer à partir de mises en variation langagières, le modèle expérimenté dans Marlowe est plus « plat » que les chroniques que l'on trouve dans les medias ou sur les blogs. Il ne s'agit pas à proprement parler d'un point de vue subjectif même si le système dispose d'une large ouverture des possibles par la mise en concurrence d'une pluralité de scripts. A la simulation de chroniques humaines, qui était une des options possibles, on a préféré la mise en œuvre systématique, ou plutôt la recombinaison, de routines déjà incorporées dans le couple Prospéro-Marlowe, de façon à conserver une unité de style : il ne s'agit pas de singer des chroniqueurs humains déterminés mais bien de développer un ensemble de variantes adaptées aux capacités propres de la machine. Si la dimension « exercice de style » est présente, c'est en mode mineur, du moins pour l'instant. Cela dit, la forme chronique est en soi un système de production sous contrainte qui rend possibles d'infinies variations. La réalisation d'une chronique hebdomadaire sur le fil santé-environnement dans le cadre d'une convention avec une agence comme l'AFSSET[15] fournit l'occasion de concrétiser cette fonction de chroniqueur.

Genèse et structure d'une chronique automatique très spécialisée

Dans le cadre de la poursuite d'un projet d'observatoire sociologique informatisé, il est apparu pertinent de se doter d'un outil de veille capable d'adresser, selon une fréquence hebdomadaire, une chronique sanitaire et environnementale, à un réseau de partenaires. Cette chronique hebdomadaire spécialisée a été conçue par adaptation d'une chronique quotidienne généraliste mise en place à la fin de l'année 2004 et en fonctionnement continu depuis. Ce dispositif expérimental n'a pas pour but de produire un outil de communication automatique diffusable tel quel à l'extérieur, mais de s'assurer que l'ensemble des protocoles d'indexation et de codage du contenu des séries textuelles est transposable à des informations en flux non supervisées par des interprètes humains. A un premier niveau, ce dispositif permet de prendre note de la progression des dossiers (objets d'alertes) dans l'actualité sanitaire et environnementale. Si le taux d'échec des procédures analytiques de Marlowe est relativement faible, la difficulté réside dans le passage d'un protocole d'acquisition de connaissances à une boucle d'apprentissage permettant au système de mieux sélectionner et structurer les informations qu'il mobilise[16]. Du point de vue de l'utilisateur, les chroniques permettent d'imaginer et de tester collectivement de nouveaux scripts transposables le logiciel Marlowe.

Comment fonctionne techniquement ce dispositif ? Le chroniqueur automatique repose sur la coopération de plusieurs instances : un module du logiciel Tirésias va sur la Toile chercher les informations de la semaine, tandis qu'un autre les transforme en corpus directement analysable sous Prospéro. Marlowe entre en lice en bout de course : disposant d'une pluralité de scripts et de chemins possibles (voir un état de l'organigramme en annexe), il extrait des informations qu'il juge pertinentes et organise leur « commentaire ». Une fois qu'il a terminé la rédaction de sa chronique, il l'adresse par courriel à un réseau de collaborateurs humains. Dans la foulée, il met à jour des fichiers qui lui servent à établir des palmarès et à repérer des évolutions, à construire pas à pas la structure de ce que l'on peut attendre dans l'actualité, à établir des listes de traits qu'il s'efforce de classer. Il crée également une base d'informations marquantes dont il pourra se servir ultérieurement sans avoir besoin de consulter de nouveau la base complète des dépêches.

---

[15] Cette convention de recherche passée entre l'Agence Française de Sécurité Sanitaire de l'Environnement et du Travail et l'EHESS portée par le GSPR en partenariat avec l'association Doxa a débuté en 2006 et a été renouvelé en 2007 jusqu'en 2010.

[16] On a joint en annexe de larges extraits d'une chronique adressée par Marlowe le 6 octobre 2008. Le « corpus » des chroniques disponibles pour évaluer le fonctionnement du dispositif est d'ores et déjà immense puisque nous disposons de plus de 1400 chroniques quotidiennes et d'une centaine de chroniques hebdomadaires sur le fil santé-environnement.

Soit un exemple de structure élémentaire de données générée par le chroniqueur hebdomadaire : la table des « objets d'alerte » de la semaine – ce genre de table est alimenté depuis fin 2006 (il s'agit ici d'une courte sélection depuis septembre 2008)

1/ 9/2008 - 7/ 9/2008 : ouragan nucléaire inondations CO2 pluies torrentielles cyclone gaz à effet de serre plomb glissements de terrain radon bruit couche d'ozone déforestation changement climatique radiothérapie nosocomiales déchets
8/ 9/2008 - 14/ 9/2008 : ouragan pesticides déchets CO2 changement climatique radiothérapie sécheresse inondations gaz à effet de serre OGM nanoparticules attentat déforestation dioxyde de carbone nucléaire terrorisme vache folle bruit poussière air intérieur pollution de l'air monoxyde de carbone composés organiques volatils champs électromagnétiques
15/ 9/2008 - 21/ 9/2008 : déchets couche d'ozone changement climatique bruit CO2 plomb saturnisme pesticides cyclone amiante incendies sécheresse gaz à effet de serre antennes-relais arsenic mercure incendie hormones tabac nuisances sonores changements climatiques perturbateurs endocriniens grippe aviaire marées noires OGM nitrates marée noire métaux lourds chlore
22/ 9/2008 - 26/ 9/2008 : CO2 déchets paludisme déforestation sida OGM mercure arsenic plomb pesticides gaz à effet de serre sécheresse bruit changement climatique greffes saturnisme canicule alcool tabac eaux de baignade pollution de l'air maladies cardio-vasculaires antennes-relais radioactivité poison nucléaire chlore
29/ 9/2008 - 5/10/2008 : déchets amiante dioxines pesticides CO2 pollution de l'eau bruit changement climatique espèces menacées gaz à effet de serre éthers de glycol nucléaire inondations cigarette benzène formaldéhyde sécheresse radon alcool pollution atmosphérique pollution de l'air SO2 NOx rayonnements ionisants métaux lourds radioactivité mercure plomb
6/10/2008 - 12/10/2008 : amiante déchets changement climatique CO2 nitrates gaz à effet de serre VIH nanoparticules tabac grippe aviaire sida attentats tabagisme pollution de l'air H5N1 fièvre jaune Ebola nucléaire terrorisme incendie cigarette espèces menacées poussière effet de serre pollution atmosphérique épizootie monoxyde de carbone NOx OGM pyralène benzène dioxyde de carbone pesticides
6/10/2008 - 12/10/2008 : amiante déchets changement climatique CO2 nitrates gaz à effet de serre VIH nanoparticules tabac grippe aviaire sida attentats tabagisme pollution de l'air H5N1 fièvre jaune Ebola nucléaire terrorisme incendie cigarette espèces menacées poussière effet de serre pollution atmosphérique épizootie monoxyde de carbone NOx OGM pyralène benzène dioxyde de carbone pesticides
13/10/2008 - 19/10/2008 : déchets pesticides CO2 cigarettes champs électromagnétiques tabac gaz à effet de serre plomb tsunami eau du robinet amiante UV poussière VIH antennes-relais nanoparticules rayons cosmiques pollution de l'air changement climatique hépatites hépatite C hépatite B Plomb OGM éthers de glycol arsenic toluène chlore

L'intérêt majeur de cette expérience est de fournir un banc d'essai pour les outils d'indexation, de codage et d'inférence incorporés dans les logiciels et utilisés par ailleurs sur toute une gamme de corpus spécialisés. On peut notamment envisager de lier les données à traiter et les analyses du chroniqueur à des algorithmes de classification et de tri[17].

Au premier stade de l'expérience, on a choisi d'appliquer le chroniqueur à un site d'informations génériques sur l'environnement, consultable en ligne. Les protocoles de Tiresias ont ainsi été adaptés pour construire le socle des articles depuis 2004. On y a ensuite ajouté une sélection hebdomadaire de dépêches d'agences qui, contrairement à la série précédente prise dans son exhaustivité, sont filtrées à l'aide d'un protocole vérifiant la présence d'une liste de thèmes et d'expressions liés à des problématiques sanitaires et environnementales. Un seul de présence et de déploiement (au moins plusieurs éléments de chaque classe d'indices), est utilisé afin d'éliminer les très nombreuses dépêches d'actualité politique et sociale trop générale mais aussi d'actualité sportive ou financière. La seule mention des mots « santé » et « environnement » ne suffit pas à réaliser un filtre sémantique fiable, et on procède donc par ajustements successifs des critères du filtrage : on parle en effet d' « environnement concurrentiel » ou de « santé des entreprises », d' « alertes » et de dangers ou même de « toxicité » dans toutes sortes d'activité qui n'ont rien à voir avec le fil santé-environnement qui nous intéresse

[17] Voir J. Velcin and J.-G. Ganascia "Topic Extraction with AGAPE", Proceedings of the International Conference on Advanced Data Mining and Applications (ADMA), Harbin, China, 2007.

Pour introduire dans la chronique hebdomadaire des variations permettant d'ouvrir différents possibles et d'éviter la monotonie d'un séquençage qui resterait identique de semaine en semaine, tout en assurant à chacun de ces possibles un bon degré de pertinence par rapport aux textes de la semaine, on a organisé les scripts en suivant un « processus aléatoire éclairé », les scripts étant tirés au sort mais la plupart d'entre eux étant assortis de conditions relativement fortes. Concrètement, à chaque niveau où plusieurs scripts sont compossibles, le système en prend un au hasard et réalise des tests pour savoir si les caractéristiques des textes de la semaine lui permettent de l'activer eu égard aux contraintes assignées à ce script. Par exemple, Marlowe porte une attention particulière à la présence éventuelle de thématiques rares mais extrêmement pertinentes comme celle des « faibles doses » ou encore celle des « perturbateurs endocriniens » par exemple. Il croise la recherche d'éléments déjà connus avec des éléments nouveaux ou émergents, et c' nouveaux ou émergents, et c'est dans la rencontre, à chaque fois différente, des deux plans que se fait l'intérêt de la chronique. Lorsqu'une classe d'objets ou une relation est détectée, certaines rubriques sont privilégiées de façon à structurer le fil de la chronique. Mais cette structuration, comme on le voit dans l'exemple fourni en annexe, laisse une large place à l'imprévu et à ce qui n'a pas encore été indexé ou répertorié dans les « ontologies » ou catégories sémantiques du système. Autrement dit, comme dans Prospéro, il s'agit de faire parler doublement les textes : à travers des filtres sémantiques ayant nécessairement incorporé un point de vue et à travers des propriétés émergentes saisies via des grappes, des liens ou des rapprochements inédits. Si aucune des conditions préalables à la structuration sémantique de la chronique n'est réunie, le système tire au sort un autre script du même niveau, vérifie les contraintes, et le cas échéant se replie sur des scripts élémentaires. Diverses possibilités sont prévues à l'intérieur de chaque script, y compris, dans certains cas, l'absence totale des propriétés sur lesquelles il est centré. En effet, il faut s'assurer qu'à chaque niveau il y ait toujours au moins un script possible puisque l'enchaînement des tests dépend du succès de l'étape antérieure au moins pour les deux premiers niveaux.

Une chronique de Marlowe se présente ainsi comme l'enchaînement de requêtes qui lui permettent d'extraire, à partir d'un corpus de dépêches lues quotidiennement sur le Web, des traits suffisamment marquants pour donner lieu à des commentaires, lesquels sont en quelque sorte automatiquement contextualisés. Ces derniers peuvent varier depuis le commentaire d'humeur - « les humains se font encore la guerre ! » - jusqu'au micro-rapport dédié spécifiquement à des programmes de recherche, comme tout ce qui concerne les alertes et les controverses ou les formes de mobilisation autour des enjeux sanitaires, environnementaux ou technologiques – «aujourd'hui, le dossier des OGM a encore rebondi !». Du point de vue procédural, la réalisation des chroniques repose sur l'enchaînement de cinq opérations distinctes : chaque jour, en plusieurs étapes, un module spécialisé de Tirésias, un « crawler », va lire sur la toile des textes librement accessibles ; en fin de journée, un autre module de Tirésias, dénommé le « veilleur », fabrique un corpus à partir des documents retenus – certains fils sont écartés, comme par exemple les résultats sportifs ou les événements culturels, ce qui ne va pas de soi mais procède d'un choix d'orientation du chroniqueur de Marlowe – le sport ou la culture ne surgissant que lorsqu'ils sont liés à des enjeux politiques, des mobilisations ou des affaires, ou encore des innovations technologiques ; un troisième module, appelé le « lanceur », qui sert habituellement à réitérer des questionnements en boucle sur plusieurs corpus, indexe les documents retenus et construit une base de connaissances sous Prospéro ; après quoi, à l'aide d'une centaine de scripts différents, Marlowe analyse les propriétés de ce corpus et en tire un certain nombre d'observations ; enfin, un cinquième module assure la transmission de la chronique au carnet d'adresse du logiciel.

Le cahier des charges de ce chroniqueur s'est fixé graduellement et, depuis le début de l'année 2005, Marlowe produit quotidiennement un texte qu'il adresse à un réseau d'interlocuteurs proches qui contribuent en retour à enrichir ses « scripts ». Peu coûteuse en terme de développement stricto sensu, la réalisation de la chronique quotidienne a permis d'avancer sur plusieurs lignes de front :

- En premier lieu, il s'agissait d'assembler des modules et des fonctionnalités déjà réalisés pour créer un dispositif autonome. L'autonomie est entendue ici au sens où, pour effectuer toutes les opérations qui mènent de la capture d'informations sur le Web à l'envoi d'un compte rendu rédigé à l'intention d'un réseau de lecteurs, ce dispositif ne suppose aucune intervention humaine. D'un point de vue formel, le procédé est

assez simple puisqu'il s'agit de projeter des éléments extraits de fils d'actualité, triés à l'aide d'indices élémentaires, dans des figures de style d'inspiration plus littéraire, en réalisant la jonction entre deux techniques d'écriture. La réalisation de ce dispositif, d'abord conçu sur un mode généraliste, a permis de concevoir des variantes tournées vers des problématiques ou des sources plus spécifiques. Il s'est avéré utile, par exemple, de développer un chroniqueur hebdomadaire spécialisé en matière de relations entre la santé et l'environnement.

- En deuxième lieu, ce processus rend clairement visibles, et partant critiquables, les procédés syntaxiques et sémantiques sur lesquels reposent les routines du système. Le fait de laisser une machine produire une interprétation de séries de documents au format relativement stable mais au contenu ouvert sur toutes sortes d'événements et d'affaires publiques fournit en effet un contre-test précieux face au travail plus classique d'ancrage de jeux de descripteurs et de catégories d'analyse sur des dossiers pris un à un et soumis à l'examen continu d'un expert humain. Les points de décrochage ou de dérivation sont plus clairement lisibles et permettent de poser la question du domaine d'extension d'un langage de description, dont l'analyse de multiples dossiers montre qu'il est soit trop général, soit trop spécialisé, et qu'une des qualités majeures du raisonnement humain, même sous forme de routines, est de négocier constamment les changements d'échelles ou de niveaux logiques.

- Enfin, le fait de disposer d'un agent capable de suivre continûment un flux d'informations permet de donner corps à l'idée de délégation socio informatique et d'assurer une mémoire transversale, fort utile pour recontextualiser des alertes ou des controverses et toutes sortes d'événements dont le caractère marquant n'est aperçu qu'après coup. Par exemple, en suivant quotidiennement l'actualité politique, en relevant systématiquement les thèmes liés au domaine des risques ou en repérant les affaires dont on parle le plus, le système peut générer une base d'événements marquants utilisables comme points d'appui pour contextualiser ou interpréter des transformations observées dans des séries particulières, coupées artificiellement du fonds avec lequel elles communiquent. On peut penser à ce que les politistes appellent, sans trop s'interroger sur la dimension fonctionnaliste de cette appellation, l' « agenda politique » : si des mobilisations ou des interventions ont lieu sur un dossier comme le nucléaire, il peut être utile de savoir si c'est, ou non, en période de campagne électorale ou lors de l'examen d'un texte de loi concernant l'énergie. Tout peut servir de contexte à tout et c'est bien la difficulté de la modélisation de l'interprétation de séries textuelles sur laquelle ont buté de multiples approches. Sans chercher à tout résoudre, on s'est donné pour objectif d'accroître la modularité du dispositif en développant pleinement l'idée d'une pluralité de technologies littéraires.

Forger des alliances interdisciplinaires pour explorer les différentes formes d'apprentissage du système

Le dispositif présenté ici peut-il relever de l'apprentissage ? La notion d'« apprentissage » a déjà une lourde histoire dans le champ de l'IA : dans la version radicale, apprendre à des machines à apprendre, ce n'est pas seulement les doter d'outils de « data mining », « d'indexation », ou encore de « clustering », c'est littéralement créer des calculateurs autonomes capables de revoir leur propre programme ; dans la version plus académique, c'est avant tout l'explicitation des boucles logiques et sémantiques que peut prendre en charge un programme[18]. Le nombre d'expériences est immense mais la controverse est toujours ouverte : la part d'ajout humain dans l'augmentation des capacités cognitives des programmes informatiques est toujours plus grande, de sorte que la question de leur faculté d'apprentissage est de plus en plus difficile à évaluer sans métaphore, comme dans le cas des « consciences artificielles » réputées « émergentes » selon Rey Kurzweil ou « constructibles » selon Alain Cardon[19]. L'idée que l'on est entouré d' « agents intelligents » de plus en plus autonomes circule sans que l'on puisse sérieusement mettre à l'épreuve les protocoles[20]. Pour ce

[18] Y. Kodratoff, Leçons d'Apprentissage Symbolique Automatique, Toulouse, Cepadues, 1988.
[19] R. Kurzweil, The Age of Spiritual Machine - When Computers exceed Human Intelligence, Penguin Books, New York, 1999 ; A. Cardon, Conscience artificielle et systèmes adaptatifs, Paris, Eyrolles, 1999.
[20] Voir les fils d'informations et les chroniques du site automatesintelligents.com

qui nous concerne, à savoir l'intelligence sociologique des dossiers complexes, on est conduit à rester beaucoup plus modestes : l'intelligence est d'abord dans les discours et les textes étudiés. Et la révision des connaissances et la dynamique inférentielle se produisent dans les textes eux-mêmes, et la première tâche consiste à apprendre à suivre les raisonnements des acteurs-auteurs. Mais à partir des cadres conceptuels et des premières procédures mis en place, il est possible d'interconnecter la suite logicielle Prospéro-marlowe avec des approches de clustering automatique et des protocoles d'évaluation des informations et de leurs changements graduels de statut, comme lorsque des micro-configurations basculent du mode mineur –on en cause dans les coins ou toujours à propos d'un domaine précis – au mode majeur – émergence d'un nouveau prototype ou d'un paradigme, comme ce fut le cas avec le principe de précaution, et aujourd'hui avec le modèle d'expertise du GIEC sur le climat transformé en modèle (cf. le Grenelle de l'environnement ..).

Un des antidotes à la dispersion, la redondance et la prolifération des noeuds et des liens qui caractérisent la mise en réseau non seulement des informations mais aussi des analyses de ces informations, consiste à se doter d'instruments alternatifs permettant de sérier les problèmes en prenant à la lettre la signification du verbe « sérier » : construire des séries raisonnées, c'est-à-dire dont la raison est explicitée et dont on peut fournir le principe de mise en équivalence ou d'homogénéité, ou encore de congruence. En concevant une collection de dossiers sous la forme de séries de corpus distribués entre de multiples machines et accessibles à un réseau d'enquêteurs capables non seulement d'enrichir les corpus, mais aussi de développer des outils d'analyse (descripteurs, catégories, formules, scripts et protocoles), on tente de retourner le rapport de force en remettant des contraintes cognitives fortes sur le traitement des séries et leurs comparaisons. Il s'agit ainsi d'assurer une explicabilité et une lisibilité des ressorts interprétatifs en vertu desquels des modèles, des concepts, des énoncés ou des théories réussissent mieux que d'autres à rendre compte d'un phénomène ou d'un processus. Ce déplacement ne rompt pas les liens avec l'exigence d'ouverture impliquée dans l'idée d'une connaissance ouverte et d'un Web coopératif par nature, mais oblige à sortir quelques temps de la boucle interprétative pour mettre à distance certaines classes d'événements ou d'énoncés sur ces événements, et en travailler la description, la compréhension et l'analyse avec d'autres outils que ceux que partagent les protagonistes eux-mêmes.

Annexe : Extrait d'un chronique santé-environnement exemplaire de Marlowe, le 6 octobre 2008

Le texte qui suit est le contenu assez typique d'une chronique hebdomadaire adressée automatiquement par Marlowe à une série de correspondants dans le but d'aiguillonner la veille sociologique collective sur le fil santé-environnement et de permettre la réinterrogation des corpus qui sous-tendent le dispositif et qui sont régulièrement mis à jour. Marlowe simule une sorte d'interrogatoire en s'adressant à lui-même des requêtes. Pour faire court et pour limiter le matériel au strict nécessaire, on a retiré des références littéraires utilisées notamment dans la phase d'amorce de la chronique. Pour « gratuites » qui puissent paraître ces références, elles ont pour fonction d'ancrer le dispositif dans une solide culture en sciences humaines et sociales …

[…]

Marlowe: lance le chroniqueur santé environnement

Marlowe : On a tous nos petites manies. En ce moment je scrute la moindre apparition de la problématique des "faibles doses". Cette semaine, je n'ai pas trouvé grand chose sauf dans le texte intitulé " Cancers et environnement: forts soupçons sur les pesticides " (JDLE 3/10/2008) :

Les experts de l ' Inserm ont insisté sur l ' importance de poursuivre les efforts de recherche pour améliorer les connaissances sur l ' effet à long terme des faibles doses de polluants, en jeu notamment pour les dioxines et les rayonnements ionisants, ainsi que l ' impact des " mélanges " de plusieurs facteurs ainsi que la susceptibilité individuelle face au risque.
Auteur :JDLE Date :03/10/2008 Titre :Cancers et environnement: forts soupçons sur les pesticides

Certes, ce n'est pas ça qui va faire faire un grand bond en avant à nos recherches sur la question des faibles doses mais je tenais à signaler que je suis vigilant !

Marlowe: approfondis

Marlowe : Si on regarde les objets d'alertes de la semaine, on obtient une liste ordonnée qui a la forme suivante :
déchets, amiante, dioxines, pesticides, CO2, pollution de l'eau, bruit, changement climatique, espèces menacées, gaz à effet de serre, éthers de glycol, nucléaire, inondations, cigarette, benzène, formaldéhyde, sécheresse, radon, alcool, pollution atmosphérique, pollution de l'air, SO2, NOx, rayonnements ionisants, métaux lourds, radioactivité, mercure, plomb …

J'ai entr'aperçu soigneusement cette liste et après avoir comparé consciencieusement, pour chaque objet d'alerte, les thèmes qui lui sont associés cette semaine à ceux qui forment son réseau typique dans l'ensemble du « corpus socle », j'ai retenu « cigarette ». En effet, la configuration engagée s'écarte du noyau habituel : décès, charbon, bois, chauffage, conjoint, usage, responsabilité, cause, poumon, cancers, mortalité, chroniques, respiratoires, maladies, Chine, renoncement, personnes, nombre, Etats-Unis, Harvard, santé publique, Ecole, Ezzati, Majid, Lin, Hsien-Ho, équipe, chercheurs …

Parmi ces éléments, 27 sont assez inhabituels : décès, charbon, bois, chauffage, conjoint, usage, responsabilité, cause, poumon, cancers, mortalité, chroniques, respiratoires, Chine, renoncement, personnes, nombre, Etats-Unis, Harvard, santé publique, Ecole, Ezzati, Majid, Lin, Hsien-Ho, équipe, chercheurs

Tant que j'y suis, je fournis la liste ordonnée des thèmes les plus fortement associés de manière générale (dans le socle) à un objet comme « cigarette » : alcool, risques, euros, dangers, tabac, industrie, réchauffement climatique, désinformation, campagne, ExxonMobil, pétrolier, stress, drogue, violence, BIT, sida, mégots, plastiques, matières, voitures, carcasses, fumée, perturbateurs endocriniens, monoxyde de carbone, bactéries, amiante, santé, effets, maladies, composés organiques volatils, ventilation, bâtiments, âge, activités, chant, facteurs aggravants, scientists, Union, association, bénévoles, Australie, Nettoyons, la journée, année, succès, ambition, international, bâtiment, syndrome, substance …

Un peu de verbatim rendra tout ceci plus cristallin :

Les chercheurs, dont l ' équipe était conduite par Hsien-Ho Lin et Majid Ezzati de l ' Ecole de santé publique d ' Harvard (Etats-Unis), ont estimé le nombre de personnes qui devraient mourir d ' ici 2033 en Chine de maladies respiratoires chroniques (2e cause de mortalité) ou de cancers du poumon (6e cause), et la responsabilité dans ces décès de l ' usage conjoint de la cigarette et du charbon ou du bois de chauffage. Auteur :AFP Date :04/10/2008 Titre :Chine: tabac et chauffage domestique feront 32 millions de morts d'ici 2033

Ils ont pu établir ainsi que le renoncement à la cigarette et en même temps au chauffage au charbon ou au bois devrait éviter 32,3 millions de décès.
Auteur :AFP Date :04/10/2008 Titre :Chine: tabac et chauffage domestique feront 32 millions de morts d'ici 2033

[…]

L'aventure de cette chronique environnementale et sanitaire (et vice versa) ne faisant que débuter, je pioche un peu dans des répertoires d'objets convenus. Ainsi, je note la présence de thèmes comme PNSE, expertise, exposition professionnelle, santé publique, développement durable, principe de précaution ... Voici de quoi justifier (éventuellement) cette sélection :

PNSE : *L ' Afsset utilisera également ce rapport dans le cadre de sa contribution au plan national Santé-environnement (PNSE 2), au plan Cancer 2 et au plan Santé au travail 2.* Auteur :JDLE Date :03/10/2008 Titre :Cancers et environnement: forts soupçons sur les pesticides

expertise : *(1) Cancer et environnement - expertise collective - Inserm, Afsset (octobre 2008)* Auteur :JDLE Date :03/10/2008 Titre :Cancers et environnement: forts soupçons sur les pesticides

exposition professionnelle : *L ' exposition professionnelle aux éthers de glycol et aux autres facteurs de risque a été évaluée à partir de questionnaires, d ' entretiens individuels et d ' une expertise par des spécialistes de l ' hygiène au travail.* Auteur :JDLE Date :30/09/2008 Titre :Ethers de glycol: facteur de risque d'infertilité pour l'homme

[…]

Marlowe : Depuis Duval R., Temps et vigilance, Paris, Vrin, 1990., on sait que les formes d'ouverture d'avenir se déclinent de plusieurs manières (le projet, le délai, l'attente, l'anticipation, l'agenda - j'abrège pour ne pas saturer la fenêtre -). Quod erat demonstrandum...

2010 : *La taxe poids lourd, toujours en discussion selon le secrétaire d'Etat au transport Dominique Bussereau, pourrait entrer en vigueur en 2011, et en 2010 en Alsace.* Auteur :JDLE Date :29/09/2008 Titre :Projet de loi de finances 2009: "un verdissement de la fiscalité"

D'après le projet de loi, la réduction de la taxe intérieure sur les produits pétroliers (TIPP), de 0,22 euro par litre actuellement pour le biodiesel, passera à 0,135 ¤ en 2009, à 0,10 en 2010, 0,06 en 2011 et sera abandonnée en 2012. Auteur :JDLE Date :02/10/2008 Titre :Biocarburants : suppression des aides fiscales d'ici 2012

Pour l ' éthanol, la réduction de TIPP (0,27 ¤ par litre actuellement) sera ramenée à 0,17 ¤ dès l'an prochain, puis 0,15 en 2010, 0,11 en 2011 et enfin zéro en 2012. Auteur :JDLE Date :02/10/2008 Titre :Biocarburants : suppression des aides fiscales d'ici 2012

2015 :
Le projet de loi de finances 2009 crée une taxe incinération et double, d ' ici 2015, la taxe décharge. Auteur :JDLE Date :01/10/2008 Titre :Evolution de la réglementation sur les déchets: des réactions mitigées

Boom du numérique aidant, la consommation européenne de ces équipements devrait passer à 14 TWh/an en 2015. Auteur :JDLE Date :01/10/2008 Titre :Eco-conception: les Etats membres approuvent la Commission

Réduction de 25 kg de la production de déchets en 5 ans ; recyclage de 35% des déchets ménagers et assimilés en 2012, puis 45% en 2015 (contre 24% en 2004); baisse de 15% des quantités incinérées ou mises en décharge en 2012. Auteur :JDLE Date :01/10/2008 Titre :Evolution de la réglementation sur les déchets: des réactions mitigées

[…]