MÉTHODES TECHNIQUES ET OUTILS

Étienne Brunet, Michèle Lénart, Bruno Menon, Patrick Paroubek et Marie-Anne Chabin

A.D.B.S. | « Documentaliste-Sciences de l'Information »

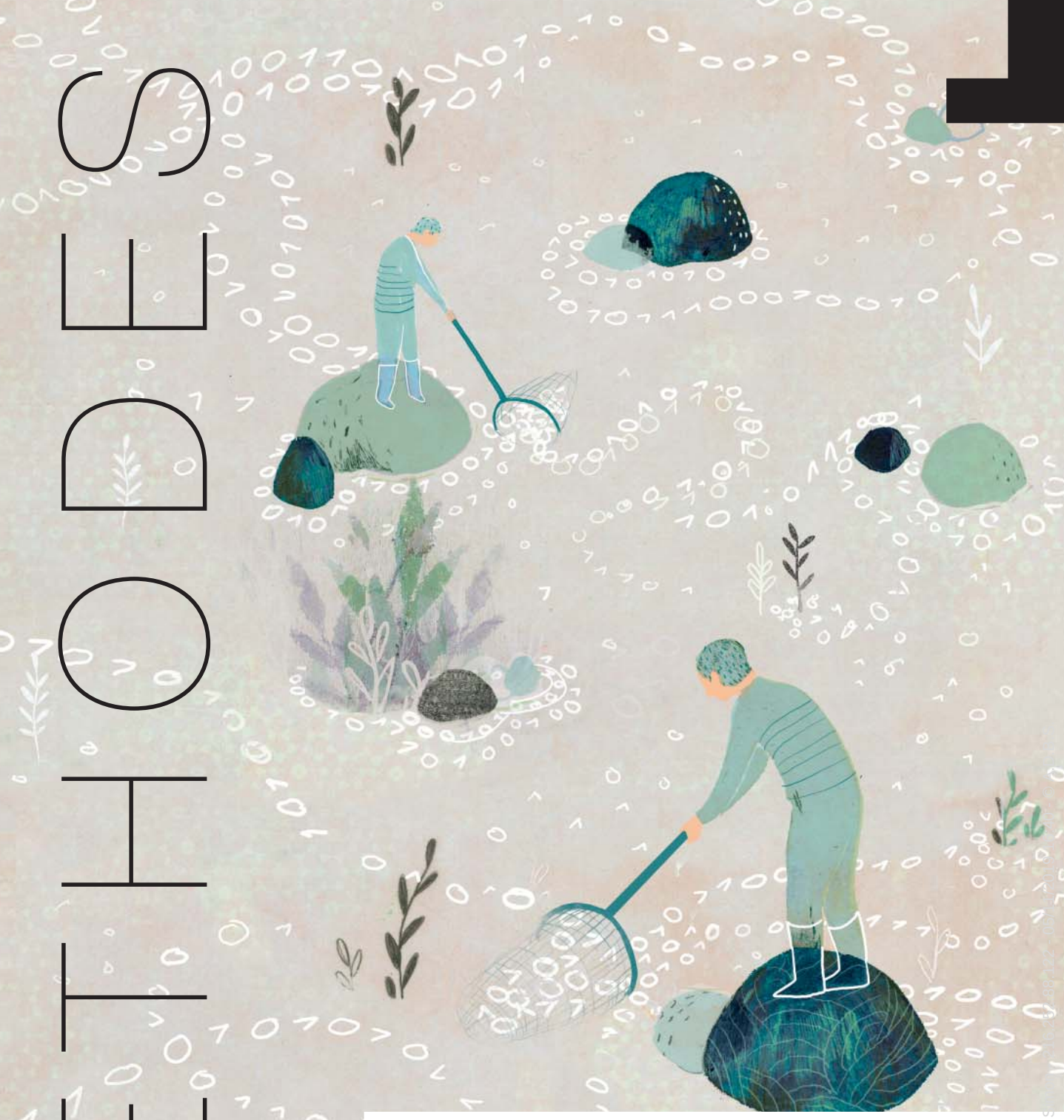
2014/1 Vol. 51 | pages 12 à 19

ISSN 0012-4508

Article disponible en ligne à l'adresse :

https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2014-1-page-12.htm

MÉTHODES



[données] **Les datamasses (ou Big Data) semblent appartenir aux sciences dures, mais les sciences humaines ne sont pas en reste : chaque jour 10 téraoctets s'empilent dans Facebook et 7 téraoctets dans Twitter. La taille de Google Books est du même ordre. Cependant, celui-ci ne travaille pas à l'échelle du jour, mais au rythme des siècles. Et aux contraintes relatives à la taille des données, à leur vitesse d'acquisition, de traitement et de diffusion, s'ajoute, pour les livres, l'obligation de nettoyer le texte.**

Yasmine Gateau

Data hygiénisme :

nettoyer les données de Google. Un cas pratique

La mesure de l'opinion à travers les réseaux sociaux ou à travers Internet impose des filtrages pour libérer le texte de la gangue des références, des balises et des hors-textes. Mais il ne s'agit pas à proprement parler de correction. Les fautes d'orthographe de Twitter sont laissées à leurs auteurs et le texte transmis à l'origine sous forme numérique n'a nul besoin d'interprétation. À l'inverse, les livres nés avant l'ère numérique imposent leur matérialité opaque : à moins de recourir à une nouvelle saisie au clavier, le texte n'y est accessible que par le truchement du scanner et de la lecture optique. Or, les progrès dans ce domaine n'ont guère dépassé les débuts prometteurs de Kurzweil[1] car il ne sert à rien d'amplifier la définition de la page numérisée : les défauts de l'encrage ou du papier sont grossis dans la même proportion. Pourtant, les auteurs du site Culturomics[2] qui exploitent les données de Google Books assurent qu'ils atteignent un taux de 95 % de lecture correcte pour l'anglais et de 90 % pour les autres langues[3].

Contrôler les références

On se propose de vérifier cette assertion et de mesurer le crédit à accorder aux données de Google Books. Les millions de livres dépouillés étant repérables et consultables sur le réseau, on pourrait songer à contrôler les renvois au texte. Il n'est certes pas difficile de prendre la base en défaut, surtout si l'on propose une graphie fautive à laquelle il manque une lettre comme dans cet exemple de requête dans Google Books avec le terme tronqué « cependant » qui donne comme résultat un extrait de l'encyclopédie de Diderot : « [...] *l'auteur semble y ivi ir démontré les loix de la catoptrique par des principes plus íxacts & plus >umineux que les auteurs qui l'on: pré- ;édé j cepndant il nc traite que des propriétés des miroirs sphériques , soit joncaves, soit convexes.* » Mais faute de temps et de patience, un tel sondage ne peut s'exercer que sur une frange infinitésimale des données, et aucune conclusion générale ne pourrait être tirée de manquements particuliers.

Outils statistiques

Cette démarche classique dont l'entrée est un mot et la sortie un contexte peut être accompagnée d'une interrogation portant sur les fréquences. Les auteurs de Culturomics ont joué le jeu de la transparence : au lieu de proposer une simple courbe retraçant l'évolution d'un mot, ils livrent les chiffres qui sous-tendent cette courbe. Mieux encore, ils mettent à la disposition des chercheurs leurs dictionnaires de fréquence, où l'emploi de chaque mot est noté année par année. Certes, on ne peut pas remonter le processus jusqu'aux images scannées et il faut se contenter des relevés et comptes qui en ont été tirés. On ne s'effrayera pas du nombre et de la taille des fichiers à transférer, si l'on se satisfait des unigrams (ou mots individuels), même s'il faut procéder à de lourdes opérations de tri, de compactage et de regroupement.

En concentrant les données[4], on aboutit, avec des chiffres seuls et sans aucun texte, à une base de 300 millions d'octets[5], grosse de 1,5 million d'entrées. Le but avoué de cette coûteuse opération n'est pas seulement de contrôler les données, mais surtout d'en permettre une exploitation facile et immédiate, sans les pesanteurs et les lenteurs liées au réseau. ////

Le but avoué de cette coûteuse opération n'est pas seulement de contrôler les données, mais surtout d'en permettre une exploitation facile et immédiate, sans les pesanteurs et les lenteurs liées au réseau.

//// Une discordance initiale nous inquiète déjà : les chiffres que nous relevons dans les données téléchargées ne correspondent pas exactement à ceux qu'on obtient directement par le réseau. Il faut en conclure que les comptes définitifs fixés en 2012 ont été sujets à retouches et que la taille de chaque année du corpus a été calculée après le rejet des rebuts.

Erreurs de lecture

Nous ne pouvons qu'estimer le volume des rebuts, puisque certains (les mots de fréquence inférieure à 30) ont été « caviardés » dans une purge préalable aux relevés disponibles. Mais on en aura une idée à partir d'un extrait qui recense toutes les variétés retrouvées quand l'interrogation porte sur le terme « été ». Avec deux accents dans l'espace de trois lettres, les avatars orthographiques se multiplient et une centaine d'avortons lexicaux sont nés de cette prolifération désordonnée, dont voici les 20 premiers, munis de leur fréquence et de leur code grammatical :
249614 ete , 130 ete_ , 5601 ete_adj , 95 ete_adp , 41892 ete_adv , 62 ete_conj , 16211 ete_det , 99993 ete_noun , 2853 ete_pron , 78 ete_prt , 79049 ete_verb , 3271 ete_x , 126767 eté , 123 eté_ , 111 eté_adp , 20076 eté_adv , 48605 eté_noun , 1493 eté_pron , 54731 eté_verb , 1485 eté_x, etc.

Écartons d'emblée l'effet mécanique du doublage des données, chaque forme pouvant être interrogée sous deux entrées : avec et sans codage grammatical, soit en l'occurrence trois entrées possibles : été_NOUN, été_VERB et été.

Le lecteur optique peut être responsable de la moitié des 60 analyses fautives : chacune des deux voyelles du mot pouvant admettre six interprétations, il y a 36 combinaisons possibles dont presqu'aucune n'a été négligée. Au total, c'est un million d'occurrences qui ont été détournées, ce qui est proportionnellement peu sur une fréquence cumulée de 130 millions.
Les courbes obtenues, portant sur les occurrences, ne sont donc guère affectées par ces erreurs de détail. Mais cela interdit toute statistique sur le nombre des entrées et affaiblit les études qu'on a tentées sur la loi de Zipf[6].

Erreurs de codage

Restent les erreurs imputables aux mauvais choix du *parser* (analyseur syntaxique) dans le processus de lemmatisation. Le linguiste peut s'inquiéter quand il observe la panique du lemmatiseur confronté aux formes inconnues et distribuant les codes à l'aveuglette. Dans le cas du mot « été », on retrouve le jeu complet des codes disponibles, mis à part celui des conjonctions. On peut comprendre le désarroi de l'automate dans les situations confuses : quand un mot est mal saisi ou mal interprété, le lemmatiseur perd ses repères et tombe dans une erreur qui, à son tour, en génère une seconde. Et pour peu que le lecteur optique se trompe de nouveau, on entre dans un lacis inextricable où le fil est perdu. On aura quelque indulgence pour ces flottements qui résultent souvent des erreurs préalables du lecteur optique. Mais, dans d'autres situations, on observe un comportement tout différent du lemmatiseur : un découpage au laser qui distribue les codes sans la moindre hésitation. Faut-il accorder crédit à la bonne foi du premier traitement qui se trompe souvent mais de façon aléatoire, ou à l'autorité péremptoire du second qui peut conduire à l'erreur systémique ?

Pour en décider, portons-nous à la fin de l'alphabet, là où les lexicographes, en fin de chantier, relâchent leur attention. Et observons le mot « ver », avec l'orthographe correcte. L'analyse ne semble pas avoir été supervisée puisqu'on y trouve beaucoup de codes fantaisistes que ne permet pas la grammaire française (pas plus que l'espagnole qui donne un autre sens au même mot) : *758187 ver , 42188 ver_adj , 135 ver_adp , 526 ver_adv , 9446 ver_det , 528603 ver_noun , 1503 ver_pron , 3506 ver_prt , 88034 ver_verb , 83771 ver_x.*

En revanche, le mot « vers » a reçu un traitement expéditif qui verse les 38 millions d'occurrences sur le compte de la préposition en oubliant les poètes qui font des vers et les morts qui

> Au total, c'est un million d'occurrences qui ont été détournées, ce qui est proportionnellement peu sur une fréquence cumulée de 130 millions.

1. http://fr.wikipedia.org/wiki/Raymond_Kurzweil
2. http://en.wikipedia.org/wiki/Culturomics
3. On est loin du seuil de 99,99 % exigé pour le déchiffrage du génome.
4. On a placé la barre inférieure à 100 occurrences, largement au-dessus de celle de Culturomics. La taille du corpus s'en est trouvée réduite à 70 milliards de mots dans le domaine français.
5. Goofre, base téléchargeable sur le site http://logometrie.unice.fr
6. http://fr.wikipedia.org/wiki/Loi_de_Zipf
7. Un autre examen complexe s'impose. Les courbes chronologiques de Culturomics ont souvent une netteté suspecte, où l'évolution semble biaisée par l'hétérogénéité du corpus. Les livres récents de Google Books, transmis sous forme numérique, n'ont guère de fautes. Mais, avec un poids décuplé, ils ont un intérêt moindre : ils représentent le tout-venant de l'édition, alors que les livres du passé - souvent plus littéraires que techniques et utilitaires - ont subi l'épreuve du temps, ce qui est plus difficile pour les ouvrages scientifiques que le progrès dépasse. Il s'ensuit une différence de genre dans la composition du corpus, qui prend indûment le masque de la chronologie.

en font d'autres. On saisit là la preuve d'un double traitement. Le singulier « ver » s'est prêté innocemment à l'automate et s'est trouvé bizarrement tronçonné, avec tout de même un ratio de 2 bonnes réponses sur 3 (528 603 sur 758 187). Le pluriel s'est trouvé assujetti à une décision automatique, sans égard au contexte. Or, de telles décisions arbitraires frappent surtout les mots fréquents. On a observé tous les mots dont la fréquence dépasse le million. Il s'en est trouvé plusieurs milliers qui ont un code unique. Il est impossible que, dans ce lot énorme, un lemmatiseur loyal et fiable n'ait pas repéré - à tort ou à raison - des homographes. L'examen du détail montre qu'on a souvent négligé les formes verbales qui coïncident avec un substantif (aides, amende, arme, armée, armes, attaques, avantage, balance, barre, cause, crainte, charges, chasse, classe, etc.). En d'autres cas, le lemmatiseur a été abusé par les pièges du français, quand un mot très commun en cache un autre auquel on ne songe pas, ce qui est le cas de « vers », mais aussi de « but, bois, bout, car, cours, voies, vins, vit, vives ».

Pour conclure

Culturomics ne permet pas de s'engager sûrement dans l'analyse des parties du discours et de la structure des langues. Mais l'intérêt linguistique n'est pas uniquement lié au codage grammatical. L'évolution des mots (graphies ou lemmes) est en soi une donnée capitale d'où l'on peut extraire l'histoire des idées, des mœurs et des peuples[7]. Vus de si haut, les mouvements de l'histoire apparaissent avec une limpidité fluide où les approximations se fondent et s'évanouissent dans la dérive générale. ■

> **Étienne Brunet**
Linguiste
brunet@unice.fr
http://fr.wikipedia.org/wiki/Étienne_Brunet

Du thésaurus aux référentiels terminologiques

[référentiel] L'intitulé de cette journée d'étude, organisée par l'ADBS le 26 novembre 2013, suggérait une trajectoire et laissait augurer d'une réponse, au moins partielle, aux questions récurrentes sur l'avenir des langages documentaires : une possible convergence entre les besoins en vocabulaires contrôlés locaux à certains systèmes d'information documentaire et les besoins de structuration de l'information dans des contextes plus larges.

En introduction de cette journée, Sylvie Dalbin a situé la notion de référentiel terminologique : dans le temps d'abord, en montrant comment elle s'inscrit dans une longue tradition évolutive des processus d'annotation et d'indexation ; dans l'espace des pratiques professionnelles ensuite, en brossant un tableau général du contexte d'utilisation et des évolutions récentes des langages de référence. Parmi les tendances actuelles, on note le passage progressif de l'indexation humaine à une exploitation des référentiels par les machines, l'évolution de leurs modalités de gestion et le partage par différentes applications. Au-delà de leur statut d'outils de contrôle des index, les référentiels terminologiques deviennent alors sémantiques, c'est-à-dire formatés et formalisés pour être manipulables en machine, potentiellement interconnectables, garantissant des accès fédérés, unifiés et cohérents aux utilisateurs.

Ces évolutions techniques et cette diversification des usages renforcent la nécessité de bonnes pratiques ; en corollaire, la fonction de gouvernance des référentiels évolue pour s'inscrire dans le cadre du *data management*, travail d'administration des données. Il faut, comme toujours, veiller à la qualité sémantique, à la cohérence, à la clarté de la documentation. Mais, avec la généralisation des pratiques de partage des référentiels, une part importante du travail porte désormais sur l'alignement de ressources différentes, en veillant à ne pas produire des tours de Babel, ni à perdre de l'information en se basant sur le plus petit dénominateur commun aux langages à aligner.

À l'évolution des usages correspond une évolution des compétences avec un même enjeu : la reconnaissance de l'intérêt et de la spécificité des référentiels.
Bernard Vatant a entamé son exposé par un panorama historique des réseaux, et en particulier du Web, vu comme un espace d'objets adressables ; avec le web sémantique, on passe à un espace de ressources nommées qui est un véritable système sémiotique : les URI y désignent des représentations variables (grâce à la négociation de contenu qui caractérise aujourd'hui le dialogue client-serveur sur le Web) de toutes sortes de *denotata*. Le web sémantique atteint à présent l'âge de raison et devient une technologie « *mainstream* », par exemple à travers l'exposition en RDF de référentiels de diverses provenances dans des domaines variés : ////

- bibliothèques, avec les initiatives de l'OCLC, de la BnF, de la Bibliothèque du Congrès, ou encore le fichier d'autorités virtuel VIAF[1] ;
- acteurs publics qui ouvrent leurs données et le font selon les principes des données liées propres au web de données (le Gouvernement britannique, la FAO, l'INSEE, l'IGN, etc.) ;
- projets collaboratifs (« *crowdsourcing* ») comme Geonames, DBpedia, Freebase ;
- moteurs de recherche, avec Schema.org, qui propose des moyens aux producteurs de ressources web de les annoter de manière structurée à l'intention des moteurs de recherche, annotations que les *knowledge graphs* de Google, par exemple, mettent à profit.

Partageant le même modèle de représentation (RDF), et s'appuyant souvent sur des ontologies communes (FOAF, SKOS[2]), ces référentiels se consolident mutuellement. Bien sûr, il reste du chemin à faire pour harmoniser les contenus et les pratiques, et pour pérenniser le travail accompli jusqu'ici. Il sera donc indispensable de développer des applications dédiées aux diverses facettes de l'administration et de l'exploitation de ces référentiels.

Documentariser les ressources

À la suite de ces interventions « en surplomb », la journée proposait des réflexions, des projets et des retours d'expérience dans un large éventail de domaines. Qu'il soit question de ressources pédagogiques ou culturelles, de livres, d'informations sur les entreprises, de programmes de concerts ou d'émissions de radio, les problématiques sont parentes : décrire, représenter, mettre en mémoire et à disposition, bref « documentariser » ces ressources. Pour cela, il faut disposer de systèmes d'organisation des connaissances, de vocabulaires au sens le plus large, servant de référentiels de valeurs pour articuler les descriptions, les mettre en cohérence, les lier à d'autres, et offrir des fonctions de recherche satisfaisantes.

Le projet ScoLOMFR-VocabNomen[3] est aujourd'hui en phase de production. Son propos était double : avec ScoLOMFR, définir un profil d'application du schéma standard de métadonnées LOM pour décrire les ressources pédagogiques de l'enseignement scolaire ; avec VocabNomen, fournir des référentiels de valeurs pour renseigner de façon pertinente et cohérente les différentes métadonnées du profil. C'est une véritable chaîne de production de référentiels qui a été mise en place, avec son équipe, son *workflow* et son outil de gestion des vocabulaires[4]. En juin 2013, VocabNomen avait déjà à son actif la publication de 32 vocabulaires, dont le référentiel des diplômes, celui des publics cible, celui des modalités pédagogiques.

Les référentiels, valeur ajoutée des professionnels de l'information

La société Électre regroupe des services et activités destinés aux professionnels du livre : la base de données bibliographiques electre.com, le magazine *Livres Hebdo* et les Éditions du Cercle de la Librairie. Les référentiels - les principaux étant ceux des auteurs, des œuvres, et des sociétés (éditeurs en particulier) - y apparaissent comme le ciment du système d'information (SI). Outre leur usage pour l'indexation, ils offrent des possibilités d'enrichissement dans la recherche d'informations (rebond) et facilitent la mise en relation avec les données d'autres acteurs du monde du livre, comme la BnF. Adopter les standards du web de données est une évidence pour l'interopérabilité de ces référentiels, mais les principes des *linked open data* (LOD) posent un problème de modèle économique à une société commerciale (même si le profit n'est pas son objectif majeur) ; la solution d'un entrepôt RDF au cœur du SI de l'entreprise (*linked enterprise data*), expérimentée avec la société Antidot, s'avère néanmoins avantageuse pour répondre aux besoins internes et joue un rôle important dans plusieurs des projets de l'entreprise. Il ne s'agit pas d'une solution miracle pour autant : par exemple, l'harmonisation entre le référentiel sociétés Électre et l'annuaire professionnel géré à *Livres Hebdo* n'est guère aisée, du fait des perspectives différentes ayant présidé à l'élaboration de ces outils (approche journalistique *versus* base de données).

Le projet HDA-Lab[5] se présente comme une interface de recherche sur une version « sémantisée » des ressources du portail Histoire des arts. La préparation des données est assurée à travers un module de « tagging sémantique » qui permet d'enrichir les indexations des 5 000 notices par des liens vers d'autres ressources, *via* les données pivot d'un référentiel. C'est ici le duo Wikipedia/DBpedia (extraction en RDF de données depuis Wikipedia) qui fait office de référentiel, (ce qui ne manquera pas de surprendre nombre de professionnels). Cette solution semble bien répondre aux besoins de sémantisation des index et ouvre des perspectives passionnantes : multilinguisme, désambigüisation, localisation, illustration, etc. Une preuve de concept qui se prolongera par une expérience sur la base Joconde (près de 500 000 notices). Selon Bertrand Sajus, les référentiels vont jouer un rôle de plus en plus grand et « *une grosse part de la valeur ajoutée des professionnels va se porter davantage sur les référentiels dans l'écosystème des LOD que vers les thésaurus maison* ».

La culture des outils sémantiques

Chez Kompass, la question de l'ouverture des données n'est guère à l'ordre du jour : le référentiel dont il s'agit, la classification ou nomenclature de produits et de services, disponible en 25 langues, est l'épine dorsale de l'activité de la société et, comme

1. http://viaf.org
2. FOAF et SKOS sont des schémas RDF pour la description, respectivement, des personnes avec leurs activités et leurs relations, et des vocabulaires contrôlés.
3. Initié par la Direction générale de l'enseignement scolaire (DGESCO) et le Centre national de documentation pédagogique (CNDP)
4. ITM T3 de la société Mondeca
5. http://hdalab.iri-research.org/hdalab/

telle, n'a pas vocation à être exposée sur le web de données. Sa refonte en 2012/2013, dans le cadre d'un projet fortement structuré, avait pour objectifs simplification et modernisation.

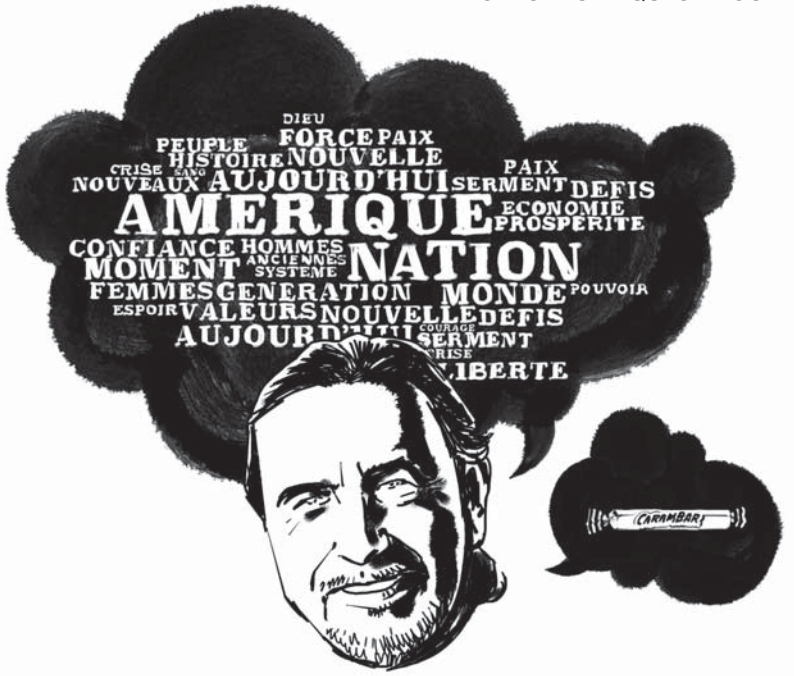
A Radio France enfin, le besoin de référentiels apparaît criant : dans les services de documentation, dans les diverses branches métier et sites de l'entreprise, ou pour la description des programmes diffusés ou de l'offre web, les mêmes informations peuvent être produites et décrites à plusieurs reprises. Ce qui génère coûts, redondances et incohérences. Du fait de leur expertise, c'est aux services documentaires que devrait revenir l'élaboration des futurs référentiels (personnes, œuvres, genres, événements, lieux, etc.), dans une triple perspective d'amélioration de la productivité, de valorisation et de réutilisation des connaissances. De ce dernier point de vue, c'est vers les modèles du web sémantique que s'oriente Radio France.

Des thésaurus mentionnés par le titre, il ne fut donc pas question. Ils n'ont pas pour autant disparu : chaque année, de nouveaux ou de nouvelles versions de ceux existants apparaissent. Mais les différentes interventions ont montré que les référentiels, qui représentent un élargissement et une généralisation du modèle des systèmes d'organisation des connaissances, sont le point focal des horizons informationnels du XXIe siècle, et notamment du web de données : dans ce paysage, la culture des outils sémantiques commune aux professionnels de l'information et du document est plus que jamais indispensable. ■

> **Michèle Lénart**
Enseignante à l'EBD
michelelenart@free.fr

> **Bruno Menon**
Maître de conférences en sciences de l'information
bruno.menon@univ-paris8.fr

Programme de la journée : www.adbs.fr > Manifestations ADBS > Journées d'étude



Eric Nosal

Jean Véronis,

une odyssée du 3e millénaire à la croisée de l'informatique et de la linguistique

[communication] Retour sur le parcours de Jean Véronis, auteur du nébusloscope, du présidographe, du chronologue, et autres travaux ludico-informatiques, connu du grand public pour la représentation par des mots clés des discours politiques de Nicolas Sarkozy et pour son blog sur les technologies du langage.

Le 8 septembre 2013, nous apprenions avec stupeur le décès accidentel de Jean Véronis, professeur d'informatique à la Faculté des lettres de l'Université d'Aix-Marseille, à l'âge de 58 ans. Jean Véronis était informaticien, chercheur en traitement automatique des langues (TAL), linguiste, consultant pour les grands comptes, entrepreneur... Il s'était fait connaître du grand public pour son blog sur le TAL[1] et ses analyses pointues des propos des politiciens[2], qui avaient attiré l'attention des médias.

Le parcours de Jean Véronis est à la fois exemplaire et emblématique de la révolution en cours pour la communication humaine depuis notre entrée dans le troisième millénaire. Cette révolution est comparable, par son impact sur la société, à celle qu'eut en son temps l'invention de l'imprimerie moderne par Johannes Gutenberg. En rendant accessible l'écrit au plus grand nombre, ce dernier avait été ////

1. Aixtal http://blog.veronis.fr
2. L.-J. Calvet, J. Véronis. *Combat pour l'Elysée: Paroles de prétendants*, Seuil, 2006. Et L.-J Calvet, J. Véronis. *Les mots de Nicolas Sarkozy*, Seuil, 2008
3. E. Chételat-Pelé, A. Braffort. & J. Véronis. « Description des mouvements des sourcils pour la génération automatique ». Traitement Automatique des Langues des Signes (TALS 2008), atelier de Traitement Automatique des Langues Naturelles 2008 (TALN 2008), www.afcp-parole.org/doc/Archives_JEP/2008_XXVIIe_JEP_Avignon/PDF/session/session_TALS_02.html

//// à l'origine d'une technologie qui a prévalu pendant presque 600 ans. Désormais, les moyens d'édition et de communication électroniques permettent à tous de publier en temps réel leurs émotions, sentiments ou réflexions, et de les rendre accessibles à l'ensemble des internautes. Ces nouveaux moyens révolutionnent non seulement la manière de communiquer, les documents, le langage mais aussi la société, en particulier dans ce qu'elle a de plus incontournable : la vie politique. Comme pour l'impression papier, ces changements s'accompagnent d'une révolution industrielle qui ne touche pas simplement les moyens de communication (papier, ordinateurs, réseaux de communication, médias électroniques) mais la manière dont nous appréhendons l'information car, au-delà de l'information factuelle véhiculée par un message, c'est la partie affective, subjective, émotionnelle qui devient analysable dans l'instant, sur une échelle planétaire.

Cela, Jean Véronis l'avait compris très tôt, car il a toujours situé son activité dans les domaines émergents du traitement automatique des langues, s'intéressant aussi bien à l'écrit, à la parole ou à la langue des signes[3]. Dans son parcours on trouve d'abord les didacticiels et l'étude des systèmes d'écriture en articulation avec la phonologie[4], sa thèse en 1988 sur l'étude des erreurs dans les dialogues homme-machine, les dictionnaires[5], puis une période consacrée aux corpus multilingues dans le cadre de projets européens comme Multext[6], la représentation des corpus avec la Text Encoding Initiative[7]. Vint ensuite l'évaluation des technologies du langage, qui commençait à reconquérir au sein de la communauté TAL les lettres de noblesses perdues suite à la publication du rapport Alpac[8], avec les campagnes d'évaluation Arcade[9] et Romanseval qu'il organisa.

...car, au-delà de l'information factuelle véhiculée par un message, c'est la partie affective, subjective, émotionnelle qui devient analysable dans l'instant, sur une échelle planétaire.

Pendant sa présidence de l'Association pour le traitement automatique des langues (Atala), de 2000 à 2008, Jean Véronis transforma la revue de l'association *Traitement automatique des langues*, alors publiée par Hermès, en une revue électronique diffusée de façon libre et gratuite[10]. Dans un même temps, il acquérait la notoriété médiatique que nous lui connaissons avec son blog et ses ouvrages sur le discours politique, tout en recentrant ses activités professionnelles vers l'industrie et la création d'entreprises autour du TAL et de l'analyse d'opinion, qui aboutit en janvier 2013 à la création avec Benoît Raphaël de la *start-up* Trendsboard[11].

C'est au moment où il préparait son retour dans le monde académique que Jean Véronis nous a quittés. Mais cet Ulysse des temps modernes a laissé un sillage qui restera gravé dans nos mémoires comme emblématique de l'évolution du TAL à l'aube d'une révolution qui a commencé de toucher le quotidien de chacun. ■

> **Patrick Paroubek**
LIMSI-CNRS, Président de l'Atala
pap@limsi.fr

4. J. Véronis. *TASKIL : un logiciel d'apprentissage de la voyellation en arabe.* (Rapport technique). Aix-en-Provence : Centre de Recherches et d'Etudes Linguistiques, 1982

5. J. Véronis & N. Ide. « A feature-based model for lexical databases », 14th International Conference on Computational Linguistics (COLING'92), Nantes, 1992, http://aclweb.org/anthology//C/C92/C92-2089.pdf

6. J. Véronis & L. Khouri. « Etiquetage grammatical multilingue: le projet Multext ». *Traitement Automatique des Langues*, 1995, 36(1-2), p. 233-248.

7. N. Ide & J. Véronis. « Présentation de la TEI: Text Encoding Initiative ». *Cahiers GUTenberg*, 1996, 24, p. 4-10

8. M. King. « When is the next Alpac report due », *in : Proceedings of the 22nd meeting of ACL*, 1984. http://acl.ldc.upenn.edu/P/P84/P84-1072.pdf

9. Y.-C. Chiao, O. Kraif, D. Laurent, T. M. H. Nguyen, N. Semmar, F. Stuck, J. Véronis, W. Zaghouani. « Evaluation of multilingual text alignment systems: the ARCADE II project ». *Proceedings of LREC 2006,* Genoa, 22-26 may 2006, http://hal.inria.fr/inria-00115670

10. www.atala.org/-Numeros-en-ligne-

11. www.trendsboard.com

Les ouvrages de Jean Véronis

- *TEI : Background and Context*, en collaboration avec Nancy Ide, Kluwer Academic Publishers, 1995
- *Parallel Text Processing : Alignment and use of translation corpora*, Kluwer Academic Publishers, 2000
- *Le Traitement automatique des corpus oraux*, Hermès Science, 2004
- *Combat pour l'Élysée : paroles de prétendants*, avec Louis-Jean Calvet, Seuil, 2006
- *Les politiques mis au Net : l'aventure du PoliTIC'Show*, avec Estelle Véronis et Nicolas Voisin, Éditions Max Milo, 2007
- *François Bayrou : confidences*, avec Estelle Véronis et Nicolas Voisin, Éditions Max Milo, 2007
- *Les mots de Nicolas Sarkozy*, avec Louis-Jean Calvet, Seuil, 2008

Vous avez dit intégrité ?

[document] Le besoin d'intégrité fait l'unanimité, au moins dans les normes sur l'archivage et la conservation.

Suliko

Pour ISO 15489 et les normes qui s'en réclament telles que SO 3030X, MoReq, ICA-Req, l'intégrité renvoie au caractère authentique (le document, au travers notamment de sa signature et de sa date, montre qu'il est bien ce qu'il prétend être) et non altéré dans le temps des documents archivés. L'Arma (Association des professionnels du records management) fait de l'intégrité le principe d'archivage n° 2 sur ses huit *Generally accepted recordkeeping principles*.

La norme de conservation numérique NF Z42-013 (et ISO14641-1 qui en découle) ne parle pas d'authenticité mais définit l'intégrité comme la caractéristique d'une information qui n'a subi aucune destruction, altération ou modification intentionnelle ou accidentelle. Mais comment faire une migration de format ou de support (nécessaire tous les 5 ans environ dans le monde numérique) sans modifier les trains de bits qui constituent les documents ?

La norme OAIS (Open Archival Information System, ISO 14721) répond à cette question dans sa version 2010 avec l'« information d'intégrité » (*fixity Information*) et surtout l'« autorisation de modification de l'information de contenu » qui vise une documentation rigoureuse des nécessaires modifications des fichiers numériques en cas de migration technologique.

L'intégrité n'est pas une trouvaille des technologies numériques. Le concept est très ancien, qu'il s'agisse de l'intégrité du territoire, du corps, de la morale ou d'un document (un acte sur parchemin, gratté et falsifié, n'est pas un document intègre). Mais le numérique booste l'intégrité.

Le Code civil français (article 1316-1, loi du 13 mars 2000) reconnaît à l'écrit sous forme électronique la même force qu'à l'écrit sur support papier, à deux conditions : l'identification de la personne dont émane l'écrit ; la conservation du document « *dans des conditions de nature à en garantir l'intégrité* ». Les deux notions sont liées, ce qui montre bien l'importance du *terminus a quo* : à partir de quand faut-il contrôler l'intégrité ?

Dire qu'un outil garantit la valeur probante d'un document numérique est un raccourci qui frise l'amalgame car c'est le juge et non l'outil qui en décide. En revanche, oui, le dispositif d'archivage et de conservation peut (*via* des techniques tels que l'horodatage, les empreintes, la piste d'audit, les logs, etc.) garantir l'intégrité des documents depuis leur entrée dans le système jusqu'au jour de leur consultation.

Mais se limiter à cela est ignorer que l'intégrité de l'acte ne recouvre pas systématiquement l'intégrité physique. L'intégrité de l'acte part du moment où l'écrit a été achevé (du point de vue de l'auteur) ou reçu (du point de vue du destinataire).

Un exemple pour résumer la question : la Vénus de Milo est-elle intègre ? Du point de vue de son sculpteur (Praxitèle ou Alexandre d'Antioche), la réponse est non : il lui manque le bras gauche. Mais pour le Louvre, oui, car elle est entrée dans cet état dans les collections nationales en 1821. En revanche, si la Vénus venait à perdre son sein droit… ■

MARIE-ANNE CHABIN

Expert indépendant (cabinet Archive 17) et concepteur de la méthode Arcateg™, Marie-Anne Chabin accompagne les entreprises dans leurs projets de gestion documentaire et d'archivage *(document and records management)*. Elle est membre fondateur et secrétaire général du CR2PA, le Club de l'archivage managérial. Passionnée de diplomatique numérique, elle tient un blog critique et décalé sur l'information numérique dans la société (www.marieannechabin.fr), et un autre plus technique à l'attention des professionnels des archives, (www.transarchivistique.fr).

marie-anne.chabin@archive17.fr