

My web intelligence : un outil pour l'analyse du web et des réseaux

Amar Lakel

DANS **I2D - INFORMATION, DONNÉES & DOCUMENTS 2021/1 n° 1**, PAGES 96 À 103
ÉDITIONS **A.D.B.S.**

ISSN 2428-2111

ISBN 9782242821111

DOI 10.3917/izd.211.0096

Date de mise en ligne : 24/05/2021

Article disponible en ligne à l'adresse

<https://shs.cairn.info/revue-izd-information-donnees-et-documents-2021-1-page-96?lang=fr>



Découvrir le sommaire de ce numéro, suivre la revue par email, s'abonner...
Scannez ce QR Code pour accéder à la page de ce numéro sur Cairn.info.



Distribution électronique Cairn.info pour A.D.B.S..

Vous avez l'autorisation de reproduire cet article dans les limites des conditions d'utilisation de Cairn.info ou, le cas échéant, des conditions générales de la licence souscrite par votre établissement. Détails et conditions sur cairn.info/copyright.

Sauf dispositions légales contraires, les usages numériques à des fins pédagogiques des présentes ressources sont soumises à l'autorisation de l'Éditeur ou, le cas échéant, de l'organisme de gestion collective habilité à cet effet. Il en est ainsi notamment en France avec le CFC qui est l'organisme agréé en la matière.

Mots-clés: Analyse réseaux, Cartographie web, Corpus web Crawler, Digital Studies, Viralité informationnelle

MY WEB INTELLIGENCE : UN OUTIL POUR L'ANALYSE DU WEB ET DES RÉSEAUX

L'analyse des sources ouvertes nécessite des outils qui soient capables d'effectuer des *crawls* de sites web pour mieux les catégoriser et faciliter leurs analyses sous des formes notamment cartographiques. Basé sur l'analyse des communautés en ligne et des controverses, *My Web Intelligence* est un outil pour les *digital studies* dont l'intérêt dépasse les seuls intérêts de la recherche pour faciliter l'étude et l'analyse des réseaux d'influence et des stratégies de viralité de l'information.



Amar LAKEL

Amar LAKEL est maître de conférences en sciences de l'information et de la communication à l'Université Bordeaux Montaigne. Spécialisé en humanités digitales, en gouvernance politique et en analyse des controverses, il mène ses recherches au sein du laboratoire MICA. Il est le concepteur du logiciel *My Web Intelligence*.

✉ Amar.Lakel@u-bordeaux-montaigne.fr

My Web Intelligence est un programme que je dirige au sein de l'équipe E3D du Laboratoire MICA (MICA) de l'Université Bordeaux Montaigne¹. Le programme vise à développer un outil d'extraction (*crawl*), d'archivage, de qualification et de visualisation du Web au service des *digital methods*. L'objectif est de fournir, à tous les experts et chercheurs qui souhaitent développer des études dans le domaine de l'intelligence numérique et des humanités digitales, un dispositif basé sur l'analyse des prises de parole en ligne.

¹ <https://mywebintelligence.net/en/my-web-intelligence-mapping-web-controversies/>

CRAWLER LE WEB : UN ÉCHANTILLONNAGE DE PROCHE EN PROCHE

My Web Intelligence s'appuie sur les moteurs de recherche web pour obtenir un premier corpus de documents pour démarrer la collecte d'informations. En croisant les sources des différents types d'infomédiaires, on multiplie les rationalités algorithmiques qui nous font entrer dans notre espace public numérique. À partir de ce corpus de premier niveau, l'exploration continue des liens sortants qualifiés permet de s'enfoncer dans les couches profondes du web pour obtenir le territoire numérique le plus complet possible au cœur de nos préoccupations. Ainsi, entre *crawl* profond et évaluation progressive des informations les plus pertinentes au regard de son dictionnaire projet, la plate-forme, travaillant en tâche de fond, finit par constituer un territoire d'informations traitant d'un sujet donné. Le *crawler* est donc une machine à échantillonner sur une méthode de proche en proche. Mais il faut nécessairement l'associer à des algorithmes d'approbation qui se doivent de rejeter le bruit et de classer les documents dans un ordre de priorité. L'extracteur de corpus en charge de la constitution des archives numériques embarque un navigateur web en charge d'absorber les ressources numériques qui est doté de la capacité d'extraire le contenu éditorial de la page (en mode *readable*) et d'isoler les documents multimédias de ce contenu (détections des liens hypertextes, détection des médias, etc.). Si le document est jugé pertinent, les liens hypertextes sont explorés pour récupérer les documents cités. De proche en proche, le *crawler* extrait un échantillon semi-représentatif du web.

FIGURE 1. CODES SOURCES SUR GITHUB EN LICENCE OPEN-SOURCE MIT

The screenshot shows the GitHub profile for 'My Web Intelligence'. The profile header includes the organization name, location (Bordeaux), website (http://mywebintelligence.net), and email (amar.lakel@bordeaux-montaigne.fr). Below the header, there are tabs for 'Repositories' (4), 'Packages', 'People', and 'Projects'. A search bar and filters for 'Type: All' and 'Language: All' are visible. The repository list shows three items:

- MyWebIntelligencePython**: Class Python App of My Web Intelligence. Language: Python. License: MIT. 1 star, 1 fork, 1 issue. Updated 5 days ago.
- MyWebClient**: A Web Client NodeJS for cleaning, qualification and more other stuff in MyWebPython project. Language: JavaScript. License: MIT. 0 stars, 0 forks, 0 issues. Updated 23 days ago.
- MyDocClient**: Updated on 4 Dec 2020.

Additional sections include 'Top languages' (JavaScript, Python) and 'People' (no public members).



CONSTITUER UN CORPUS WEB : UNE LOGIQUE D'ASSISTANCE AU CHERCHEUR

My Web Intelligence est composé de deux briques logicielles². *My Web Intelligence* Python, une brique logicielle en mode console développée sous python et qui permet d'extraire les données du web : c'est l'agent d'enquête du projet. *My Web Client* qui permet de naviguer dans son corpus de recherche pour non seulement nettoyer, mais appréhender son corpus par une interface de navigation web. Après l'ouverture d'un projet de recherche, le professionnel doit compléter un dictionnaire de mots clés qui permettra au *crawl* d'évaluer la pertinence des pages qu'il collecte et auxquels il attribuera une note qui servira, plus tard, aux filtres et exports de corpus. L'entrée et la sortie de données par des fonctions d'import/export, dans le cadre d'un projet, permettent l'export des énoncés en format csv ou gexf (pour l'analyse réseau sous Gephi), des domaines en csv ou gexf (données regroupées à l'échelle du site web), des médias en liste csv (images et vidéo) pour l'analyse visuelle. Les formats csv, gexf et l'utilisation d'une base de données fichier SQLite assurent l'interopérabilité avec tous les logiciels d'analyse du marché (R, iramuteq, etc.). On trouve parmi les variables qui qualifient la page : le titre*, l'URL*, la *relevance** (pertinence au regard du dictionnaire projet), *depth** (la profondeur d'extraction avec 0 pour les pages ajoutées par l'utilisateur), le *domain_id** et le *domain_name** (id et nom du domaine d'expression) et son contenu *texte**. On ajoute manuellement la date de publication* sur Google et le nombre de partages*, de commentaires*, d'interactions* sur Facebook (obtenus grâce à l'accès à son API). Il faut ajouter les données issues de l'analyse structurale des réseaux que l'on calcule grâce au fichier gexf des pages et le logiciel GEPHI (*indegree**, *outdegree**, etc.)

Ce sont en tout pas moins de 23 variables qui viennent identifier le texte et son contexte d'énonciation (inscription dans les réseaux de citation des pairs et réception de son lectorat sur les réseaux sociaux). Un second niveau d'analyse opère par regroupement des expressions au niveau du domaine d'expression que l'on qualifie humainement selon la nature sociale du propriétaire du média (secteurs d'activité, le niveau d'institutionnalisation, type de média numérique, etc.). Aux variables qui visent à inscrire sociologiquement l'équipe éditoriale, on ajoutera les indicateurs MOZ (autorité du site web) et l'*Alexa Rank* (indicateur d'audience), mais aussi les données des pages engagées dans le débat (somme des partages, des commentaires et réactions totales sur Facebook, nombre de pages engagées dans la controverse, date de la première publication).

² Le logiciel est téléchargeable à cette adresse : <https://github.com/MyWebIntelligence>



FIGURE 2. INTERFACE DE NAVIGATION DU CORPUS MY WEB CLIENT

The screenshot displays the 'My Web Client' interface. At the top, the search term 'branco' is entered, resulting in 473 items. The main area is a table of search results with columns for ID, Title, Domain, Relevance, and Tags. The table lists various articles and documents related to Juan Branco, such as 'L'imposture Juan Branco en une minute - Egalité et Réconciliation' and 'Juan Branco sous pression suite à 4 communications controversées'. On the left, there are filter options for 'Minimum relevance' and 'Maximum depth'. Below the filters is a 'Tags' section with a tree view of themes like 'Révolution', 'Gilets Jaunes', and 'Macron'. The right sidebar shows a 'Tags' section with a '+ 1/10' indicator.

#	Titre	Domain	Relevance	Tags
176766	L'imposture Juan Branco en une minute - Egalité et Réconciliation	www.egaliteetconciliation.fr	246	3
176605	Juan Branco, le genre idéal de l'insurrection - Egalité et Réconciliation	www.egaliteetconciliation.fr	232	3
176706	Georgia Proulx, Orlé Jourd - "Pourquoi je ne soutiens pas Juan Branco" - Egalité et Réconciliation	www.egaliteetconciliation.fr	151	1
176786	Juan Branco - Wikipedia	fr.wikipedia.org	146	15
176780	Juan Branco sous pression suite à 4 communications controversées	reseauinternational.net	90	1
176734	Le Point Aveugle du Révolutionnaire Juan Branco - Le Point Noir de Branco & Associés. - MK-Pols	mk-pols2.eklablog.com	83	1
176772	Le best-seller de Juan Branco, un épisode problématique - Rebellynn.info	rebellynn.info	74	10
176338	Juan Branco dévisse Macron Entretien Li-bas et /y suis	li-bas.org	66	3
176829	"Orléans" de Juan Branco: le 3ème réquisitoire contre Macron "placé à la tête du pays"	www.palestine-solidaire.org	63	9
176691	Les réponses de L'Express à Juan Branco - L'Express	www.lexpress.fr	61	5
176731	Juan Branco, Orléans - Agrévoix le média citoyen	www.agrevoix.fr	57	6
176821	« Orléans » selon Juan Branco - Alternatives Pyrénées	alternatives-pyrenees.com	56	7
176813	Juan Branco, le radical chic qui veut la peau de la Macronie - L'Express	www.lexpress.fr	55	13
176764	Des grandes écoles aux "gilets jaunes" en passant par WikiLeaks: qui est Juan Branco, l'auteur de "Orléans" en guerre contre Macron ?	www.franceinfo.fr	54	19

A.D.B.S. | Téléchargé le 09/06/2026 sur https://shs.cairn.info (IP: 216.73.217.92)

NETTOYER LES DONNÉES MASSIVES : LA GESTION COLLECTIVE DES CORPUS

My Web Intelligence est dotée d'un tableau de bord pour gérer les grands corpus à l'aide d'un certain nombre d'indicateurs. Cette interface de nettoyage et de qualification des données permet, non seulement un contrôle et une suppression du bruit, mais aussi une qualification thématique des pages web. Le nettoyage de données est une étape essentielle dans toute recherche. Pour autant face à la taille des corpus, il ne peut se faire sans l'aide d'une part d'agents algorithmiques et d'autre part par la mobilisation de collectif.

L'interface My Web Client offre la possibilité à l'utilisateur d'annoter humainement le document. Une gestion des thèmes et des contenus qu'il identifie permet, par la suite, de travailler thème par thème sur l'analyse de contenu soit directement sur l'interface soit en exportant la base de contenu thématisée pour une analyse lexicologique postérieure.



FIG. 3. INTERFACE D'ANNOTATION DES EXPRESSIONS

The screenshot displays the My Web Intelligence interface for an article titled « Crépuscule » selon Juan Branco – Alternatives Pyrénées. The interface is divided into several sections:

- Left Panel:**
 - Land:** A dropdown menu showing 'branco' and 'juan branco'.
 - Filters:** Sliders for 'Minimum relevance' and 'Maximum depth'.
 - Tags:** A tree view of categories including 'Thèmes', 'Revolution', 'Gilets Jaunes', 'Macron', 'Journalistes', 'Homosexualité', 'Oligarchie', 'Effondrement', 'Corruption', 'Autoritarisme', and 'Confirmation'.
- Center Panel:**
 - Article title: « Crépuscule » selon Juan Branco – Alternatives Pyrénées.
 - URL: https://alternatives-pyrenees.com/wp-content/uploads/2019/05/59384711_81339273569381_5178222282452619792_n-680x280.jpg
 - Text snippet: «... On s'attache au pouvoir et l'on mange le France, c'est ainsi qu'un filou devient homme d'état...»
 - Text snippet: «... Victor Hugo: Les Châtiments, Apothéose – Jersey, le 28 décembre 1852...»
 - Text snippet: «Avant tout, il s'agit d'un phénomène éditorial unique et qui laisse à penser. Un livre dont aucun média n'a parlé et pour cause, qu'aucune grande maison d'édition n'a osé, qui n'a bénéficié d'aucun des soutiens habituels aux best-sellers...»
 - Text snippet: «Anxi, grâce aux réseaux sociaux –tellement critiqués par ailleurs– ont peut-être échappé aux prescripteurs et réussit brillamment tant la diffusion que le succès du livre. Son, « Crépuscule », ne fait pas dans la dentelle et on se retrouve ici dans la tradition très française du pamphlet politique. Un genre bien oublié dans un monde où, si l'on veut éviter du succès, il faut multiplier les louanges et dire du bien des puissants pour obtenir des prébendes comme des passages à la télé ou sur les ondes ou encore des interviews bien préparés dans ces journaux qui font les carrières. Victor Hugo fut l'archétype du pamphlétaire avec : Les Châtiments » ou « Les Misérables ».
 - Buttons: Edit, Statistiques, Lire, Relais, Déposer.
- Right Panel:**
 - Controls: Id (19831), Relevance (68), Depth (2).
 - Media: A thumbnail image of a book cover titled 'NUMÉRO 1 DES VENTES QUI CE LIVRE DÉRANGE-T-IL ?'.
 - Content tagging: A yellow box with the text 'Select text in expression content.'.
 - Tagged content: A list of tags including 'THÈMES / REVOLUTION', 'THÈMES / MACRON / JOURNALISTES', and 'THÈMES / MACRON / OLIGARCHIE'.

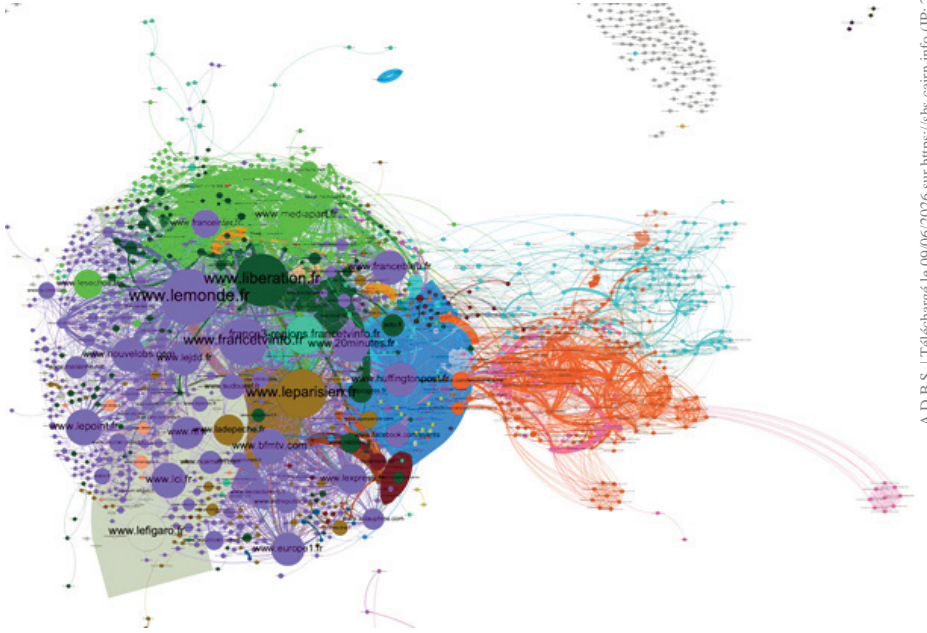
CLASSER, CATÉGORISER ET COMPRENDRE

Une fois l'extraction et la qualification des données d'une controverse achevées, My Web Intelligence donne accès à un corpus nettoyé qui permet de pouvoir mettre en place un ensemble de traitements d'analyse et de traitements des données pour tirer véritablement une compréhension de l'économie de la discussion en ligne. Le premier travail est d'utiliser la théorie des graphes et l'analyse structurale des réseaux pour générer des cartographies des médias qui sont à l'origine de la controverse. En effet, derrière les mots, il y a des locuteurs aux commandes de supports médiatiques. Des locuteurs situés et engagés dans un espace public numérique. Il faut non seulement pouvoir qualifier ces médias selon leur nature sociale, leurs comportements éditoriaux, mais il faut avant tout révéler à travers la structure de leurs citations qualifiées, le contexte d'alliance et d'adversité qu'ils tissent dans les processus de légitimation, mais aussi d'opposition. « Dis-moi qui tu cites, quelles sont tes références et je te dirai qui tu es. »

Une vision globale et structurale des acteurs révèle non seulement la structure des alliances et des oppositions, mais elle révèle les communautés d'intérêts idéologiques et situe chaque média selon un rôle social dans le débat et au sein de sa communauté (leader d'opinion, vigie, marginal sécant, bridge, etc.). Cette recontextualisation du locuteur au cœur de ses « amis » nous informe sur la position sociale du média au sein d'une communauté stratégique.



FIG. 4. CARTOGRAPHIE MÉDIAS DE LA COUVERTURE DES GILETS JAUNES DE OCT. 2018 À JUIN 2019



et la place de chaque concept dans une stratégie argumentaire globale. En réalité les sujets qui prennent position dans une controverse sont dans leur très grande majorité des porte-paroles qui habitent des discours qui leur pré-existent et qu'ils travaillent à la marge. La controverse voit rarement la création innovante d'arguments et bien plus souvent une prise de position sur des arbres argumentaires toujours déjà là dans des énoncés produits comme des mêmes. Elle permet surtout de repérer les émergences et les innovations, la diffusion voire la viralité de certains concepts.

BIBLIOGRAPHIE INDICATIVE

LAKEL, A. 2019. « Prises de positions et influences sur le web : le cas de l'information de santé ». *Revue française des sciences de l'information et de la communication* (18). doi: [10.4000/rfsic.8376](https://doi.org/10.4000/rfsic.8376).

LAKEL, A., ET LE DEUFF, O. 2017. « À quoi peut bien servir l'analyse du web ? » *Les Cahiers du numérique*, 13(3):39-62.

Des vidéos de formation au niveau de la démarche et de la prise en main de l'outil sont disponibles en vidéo : *My Web Intelligence - Formations*
<https://www.youtube.com/playlist?list=PLbCMGWVe0ggGjHwqSwz9TT5nhTFWpthQZ>