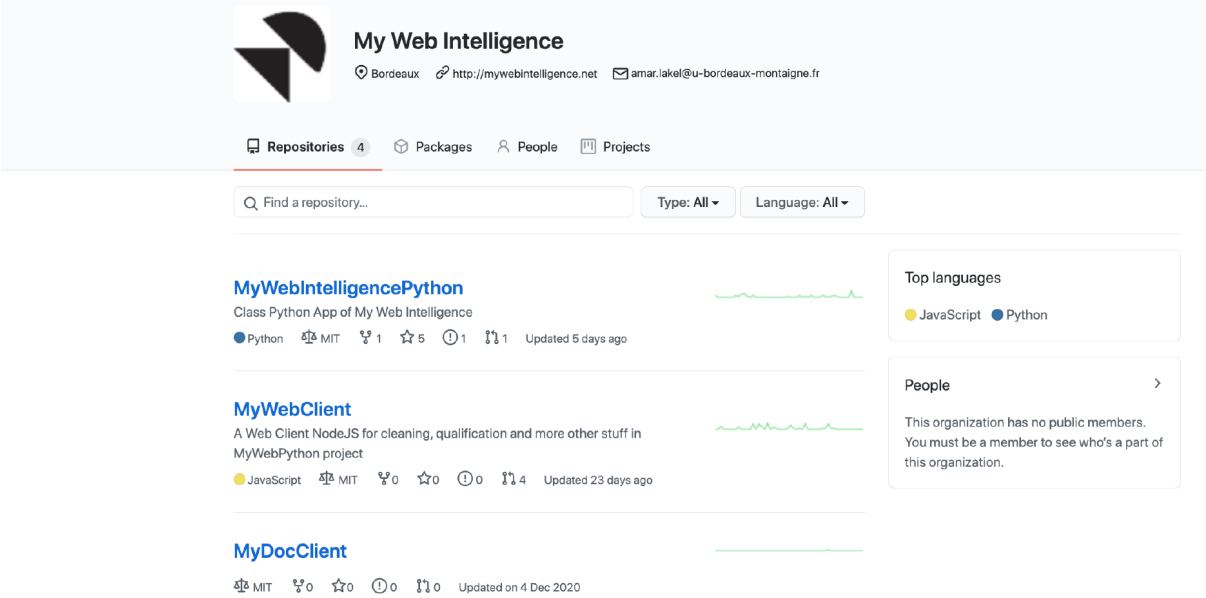
My Web Intelligence: Challenges of Heterogeneous Digital Corpora

My Web Intelligence is a program directed by Amar LAKEL within the E3D team of the MICA Laboratory (MICA) at the University of Bordeaux Montaigne. The program aims to develop a tool for extracting (crawling), archiving, qualifying, and visualizing the Web in the service of digital methods. The objective is to provide, to all experts and researchers who wish to develop studies in digital intelligence and digital humanities, with a device based on the analysis of online speech. In the market sphere, digital intelligence and data marketing have experienced extraordinary scientific and economic development in the last fifteen years. Firstborn of Web Analytics and e-commerce (technology that records user interactions on its own communication supports), digital studies experienced an explosion with Web 2.0. Access to the immensity of digital tweets from populations is still a strategic challenge for understanding masses in real-time. The Cambridge Analytica affair crowns a long series of controversies over the frenzied deployment of mass analysis of online opinions.

But the poor parent of digital intelligence remains the original Web, the one we call 1.0, which in its great immensity is composed of heterogeneous autonomous websites. With complex sources with data formats of great diversity and platforms dispersed in an immense global network, the Web is the chaos that, for a long time, seemed elusive or infinite. Even as it became more and more the space of all public speech every day. In France, the Medialab initiative of Sciences Po Paris highlighted the issues of tooling for digital methods adapted to the analysis of large corpora was born. We owe Franck Ghitalla, a researcher in Information and Communication Sciences at UTC Compiègne, to provide, in France, the first answers to a scientific methodology of web speech that is fully embedded in the sociology of controversies of the École des Mines (Bruno Latour, Michel Callon, and many others). Franck Ghitalla has constituted more than a "hidden college" of web developers (between research and business creation), he initiated a movement that led to the inclusion of the structural sociology of the analysis of social networks (which is over 70 years old) as a key to understanding the web. The Gephi software, released in 2009, is one of the great open-source successes of this college.

Figure 1. Open-source MIT licensed source code on GitHub

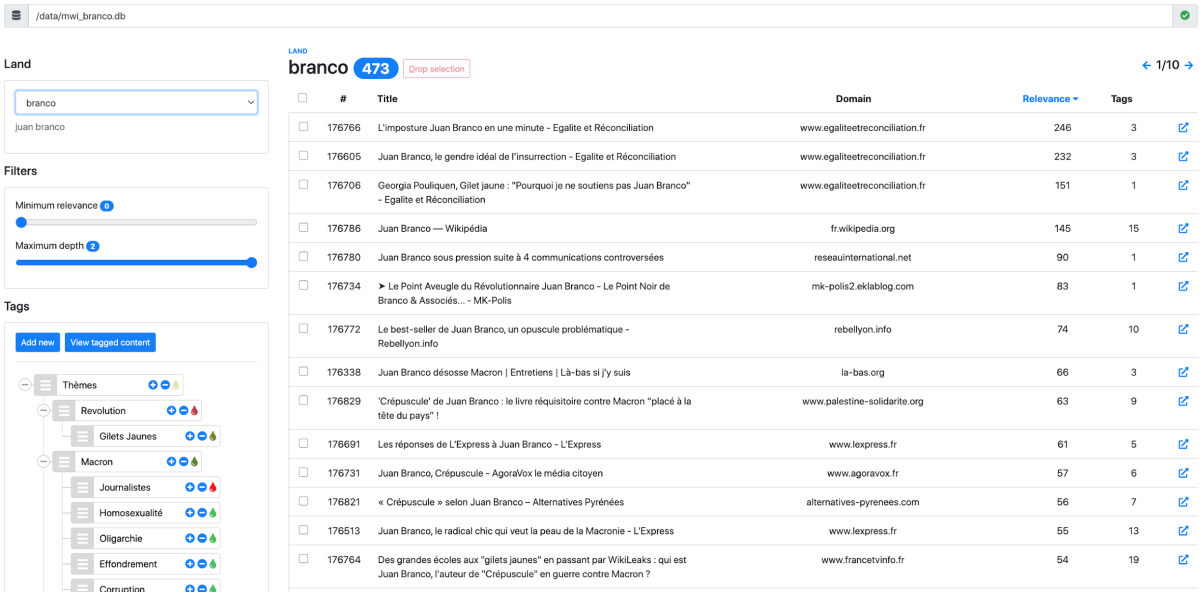


In 2008, I participated in a digital intelligence project for territories called 'Rive Droite Numérique' for the GIP GPV in Bordeaux. We were responsible for the methodology for the digital study of territories. We focused on a methodology for profiling digital speakers and proposed using Linkfluence (the first spin-off from Ghitalla college) for network mapping. However, the solution offered by the company did not meet the transparency, openness, and code access requirements of research. The limitations of private subcontracted research are the true origin of our initiative. With the emergence of the subject of digital humanities, our team refocused its research on the epistemological and methodological dimensions of the analysis of the digital corpus. Graph theories, text mining, and influence theories, it took 4 years to develop a territorialization and analysis tool for digital corpus and mapping of controversies, inspired by the sociology of controversies, in response to a strong program for the analysis of public debates. In 2014, the prototype call for proposals from the Aquitaine Regional Council finally allowed us to finance and develop an application core designed to perform information monitoring capable of absorbing data from public statements made on the Web and making them intelligible: My Web Intelligence (http://mywebintelligence.net).

The Web is a digital library that has long ago exceeded the limits of archiving as it was imagined before the digitalization of communication. The challenges of data processing, the heterogeneity of documents and communication situations, and the difficulty of identifying sources make the web today resemble a huge catch-all filled with commercial flyers, scholarly works, and direct marketing catalogs, with no organization or knowledge management. Nevertheless, this abyss where all the world's spoken words are piled up confronts us with a unique phenomenon in the history of humanity: the possibility, at least for the moment theoretical, for the researcher and the information professional to access the vast production of a new form of discourse that plays an essential role in the construction of the public space. Thus, digital studies are hindered by the possibility of establishing solid Web corpora on transparent scientific bases. The evolution of tools for analyzing large corpora (open source and free) is the only way to provide professional experts in digital

studies with a solid methodology (marketing agencies, intelligence experts, sectoral studies experts, etc.). My Web Intelligence is an open-source crawling tool under the MIT license to allow the greatest possible freedom not only of use but also of access to the code in a concern for transparency of methodological procedures. It is the realization of a fundamental principle for the academic world: open source is the only guarantee of a co-construction of digital methods ensuring the reproducibility of research.

I - Architecture for collecting data on the web



Navigating the Internet in project mode: intelligence rather than data.

For a digital intelligence platform, such as the one supported by the My Web Intelligence program, the expert must first conduct internet research as a long-term search. Rather than the primitive interface that involves entering a few keywords into a text field and receiving several hundred responses, this involves, in a project logic, conducting research based on issues that will be nourished and enriched as iterations progress. Search engines are based on a usage scenario in a context of limited rationality: finding sufficient information in the shortest possible time. This scenario is the very foundation of their ranking algorithms. In our project, the dynamic (intelligent) processing of queries is one of the keys to the relevance of the results. One of the challenges of My Web Intelligence is to provide the user with the means to conceive their search as a structured and dynamic index that will be nourished throughout their study by selecting the most relevant responses. Here we oppose the practice of study to that of request. In the long run, machine learning logic assists users over the long term in managing indexicality and evaluating digital archives to present only the most relevant documents. The platform aims to make all the features of sustainable project management accessible. The usage logic is actually that of constructing one's digital

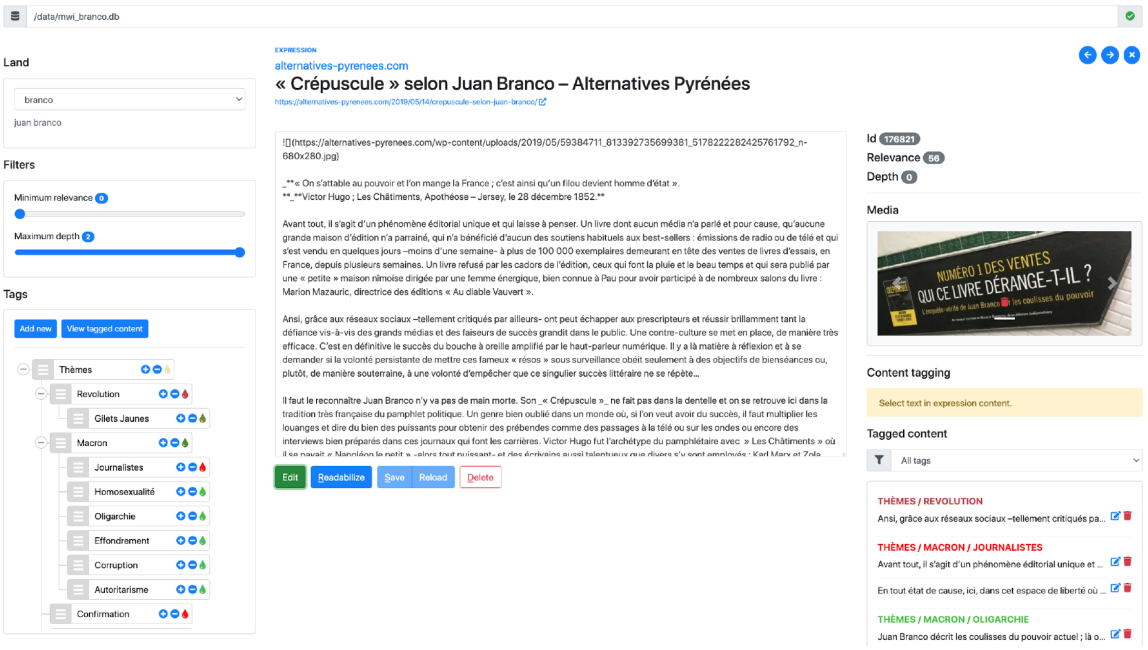
territory, whether alone, collectively, or even assisted by algorithmic agents. But this construction must always be thought of in the long term. No territory can be understood on the basis of a snapshot album; it requires exploration, wandering, and navigation. It also requires annotations and hindsight.

Web crawling: a sampling from close to close.

Our tool is more like an archaeological device than a service. My web intelligence relies on the oracles of the web to obtain the first corpus of documents (which can quickly reach several thousand expressions) to start collecting information. By cross-referencing sources, we not only free ourselves from the monopoly of Google (which really starts to be questionable when you know the microscopic part of the web it presents to us) but also from the particular filters that each device has put in place. By cross-referencing the sources of different types of infomediaries, we multiply the algorithmic rationalities that bring us into our digital public space. From this first-level corpus, the continued exploration of qualified outgoing links allows us to delve into the deep layers of the web to obtain the most complete digital territory possible at the heart of our concerns. Thus, between deep crawl and progressive evaluation of the most relevant information with regard to its project dictionary, the platform, working in the background, eventually forms a territory of information dealing with a given subject. It is this "my" web (which will continue to enrich over time) that will form the basis of the digital territory under surveillance or even governance when it comes to acting on it. A crawler is therefore a machine for sampling on a close-to-close method. But it must necessarily be associated with approval algorithms that must reject noise and classify documents in an order of priority. Because if the web that corresponds to my requests is finished, it remains immense. So rather than rejecting relevant documents, prioritization allows each project team to set the limits of its exploration. The corpus extractor responsible for the constitution of digital archives not only includes a web browser responsible for absorbing digital resources, but it is also capable of extracting editorial content from the page (in "readable" mode) and isolating multimedia documents from this content (detection of hyperlinks, detection of media, etc.). If the document is considered relevant, the hyperlinks are explored to retrieve the cited documents. From close to close, the crawler extracts a semi-representative sample of the web. In the sociological exploration of invisible spaces, the so-called snowball sampling is the only one that can be used to date to form a representative corpus.

To constitute a web corpus: a logic of assistance to the researcher

Figure 3. Interface d’annotation des expressions



My Web Intelligence is composed of two software bricks. My Web Intelligence Python, a console-mode software brick developed in python that allows data to be extracted from the web: it is the investigation agent of the project. My Web Client allows you to navigate your research corpus in order to not only clean it up but also to understand it through a web navigation interface. After opening a research project, the professional must feed the project with a dictionary of keywords that will allow the crawl to evaluate the relevance of the pages it collects and to which it will assign a rating, which will later be used for filters and corpus exports. The input and output of data through import/export functions, as part of a project, allows the export of statements in CSV or GEXF format (for network analysis using Gephi), domains in CSV or gexf format (data grouped at the website level), media in CSV list format (images and video) for visual analysis. The CSV, GEXF format, and the use of an SQLite file database ensure interoperability with all market analysis software (R, Iramuteq, etc.). The variables that qualify the page include title*, URL*, relevance* (relevance with respect to the project dictionary), depth* (the depth of extraction with 0 for pages added by the user), domain_id*, and domain_name* (id and name of the domain of expression) and its text content*. The publication date* on Google is manually added, as well as the number of shares*, comments*, and interactions* on Facebook (obtained through access to its API). We must also add the data from the structural analysis of networks that we calculate using the gexf file of the pages and the GEPHI software (indegree*, outdegree*, degree*, weighted in-degree*, weighted outdegree*, weighted degree*, Authority*, Hub*, modularity_class*, pageranks* and eigencentrality*). In addition, the content variable will undergo lexicographic classification in order to not only create a matrix of keywords but also categories of content* through clustering, which will be manually adjusted. There are no fewer than 23 variables

that identify the text and its context of enunciation (inscription in the citation networks of peers and reception by its audience on social networks). A second level of analysis operates by grouping expressions at the level of the domain of expression, which is qualified by humans according to the social nature of the media owner (sectors of activity, level of institutionalization, type of digital media, etc.). To the variables that aim to sociologically inscribe the editorial team, we will add the MOZ indicators (website authority) and the Alexa Rank (audience indicator) but also behavioral data by aggregating the data of the pages engaged in the debate (sum of shares, comments and total reactions on Facebook, number of pages engaged in the debate). It is therefore a question of combining the sociological analysis of the editorial team with the analysis of the reception of the contents by its audience. This double dimension of sociological analysis makes it possible to shed light on the editorial and reception dynamics of the contents of the web.

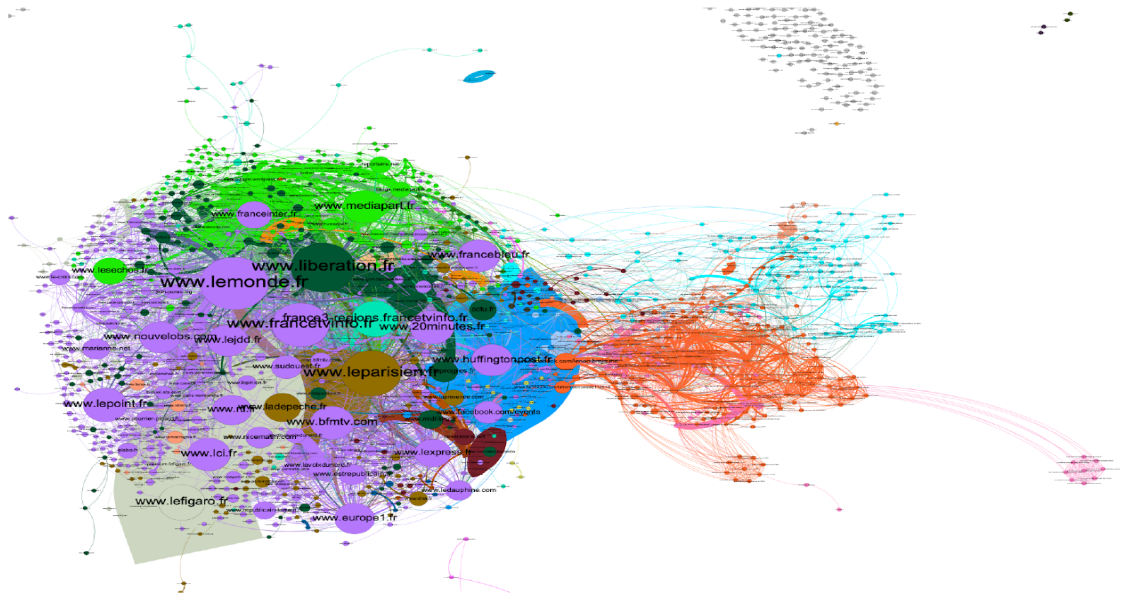
II - Cleaning and enriching data: the challenge of distributed heterogeneity

Cleaning large data: managing collective corpora.

My Web Intelligence has a dashboard for managing large corpora using a number of indicators. This data cleaning and qualification interface allows not only noise control and removal but also thematic qualification of web pages. Data cleaning is an essential step in any research. However, due to the size of the corpora, it cannot be done without the help of algorithms and the mobilization of a group. That is why My Web Intelligence is developed on a web infrastructure that should eventually become an online server. Initially, a document indexing logic should take into account not only the semantic proximity to the query index (relevance index) but also the interaction with users (removal) and the structural position of the document in the corpus (authority index). As a result, the least relevant documents are sent to the end of the index and will only be studied if human resources allow it. The user can then navigate their corpus in list format or page by page to read the cards, find the original pages, modify them or delete them from their corpus. In the long run, a system with a project manager will allow for the creation of work groups by distributing read, modify, or delete resource permissions.

Analyze content: human annotation

Figure 4 : Cartographie médias de la couverture des Gilets Jaunes octobre 2018 à juin 2019



The My Web Client interface adds the ability to humanly annotate the document. However, the platform allows for cumulative performance using the intelligence of the user crowd.

Resources already annotated by others will be offered pre-filled, which will allow for the gradual accumulation of knowledge in a given user community. Thus, each of these documents can be added to a series of annotations that will enrich the content analysis. Thus, the "what is it about?" is managed by a tag tree that represents a series of themes that the researcher can develop as they go. They then select the parts of the text to be themed. Management of themes and identified content allows, subsequently, to work theme by theme on the content analysis either directly on the interface or by exporting the themed content database for later lexicological analysis. Thematic analysis is the basis of content analysis but can also be used in a supervised machine learning model leading to automated interpretation models.

Content analysis: the challenges of Natural Language Processing

Indeed, My Web Intelligence is committed to the revolution of natural language processing in order to first allow for automatic extraction and lexicographic management of corpora. This lexicographic management is the basis for algorithms that automatically classify content at both the document and paragraph levels. The challenge of the web, as we have already pointed out, is the immensity of the corpus, that is, the ability to annotate a massive archive. To begin with, it should be noted that real intelligence projects do not work on "the web", but on a particular subset according to a theme directed by the researcher. And the bigger the set is and the more human resources deployed for its study are important. Therefore, the ideal challenge is of course to imagine the most relevant automatic data qualification solution possible. If 100% automation is impossible in terms of intelligence, the automatic analysis of documents has made remarkable progress that can allow for computer-assisted content annotation (Neural Network Clusterisation) in the very short term. However, it is the qualification of the media in its social context that seems difficult. Here again, significant automation progress is possible, but it is especially with the help of crowdsourcing that human annotation offers immense possibilities. It is indeed possible to imagine groups of experts, around given informational sectors, sharing their resources in the qualification and annotation of sectoral subsets of the web. Very quickly, the community of My Web Intelligence users will be able to collaborate in the construction of sectoral qualification indexes of the web in the form of an open data project. In the long term, this cumulative "shared memory" will offer new users relevant sectoral indexing. The ultimate challenge of My Web Intelligence is to provide a research infrastructure for the sharing and open access of enriched research corpora for the benefit of Internet Studies.

III - Classify, categorize and understand the controversy.

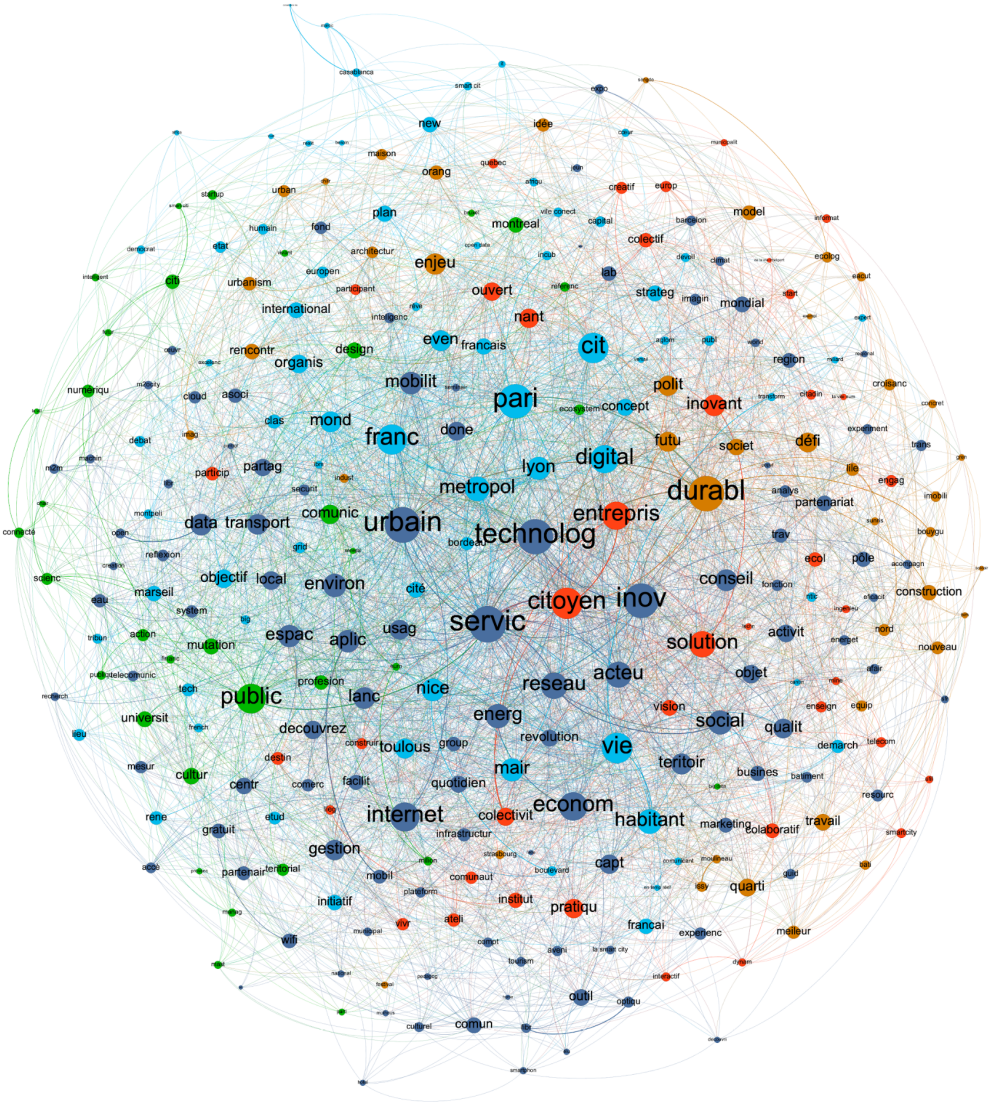


Figure 5 : analyse lexicale des pages web traitant de la smart city en 2020

Graph analysis: the challenge of the structural positions of speakers

Once the extraction and qualification of data from controversy is completed, My web intelligence gives access to a cleaned corpus that allows for a set of data analysis and processing to truly understand the economy of online discussion. The first task is to use graph theory and structural network analysis to generate maps of the media that are at the origin of the controversy. Indeed, behind the words, there are speakers in control of media outlets. Located speakers engaged in a digital public space. It is not only necessary to

qualify these media according to their social nature, and their editorial behaviors, but above all to reveal through the structure of their qualified citations, the context of alliance and adversity that they weave in processes of legitimation but also of opposition. Tell me who you quote, what are your references and I will tell you who you are. A global and structural view of the actors not only reveals the structure of alliances and oppositions, but it also reveals communities of ideological interests and locates each media according to a social role in the debate and within its community (opinion leader, watchman, marginal secant, bridge, etc.). This recontextualization of the speaker at the heart of his "friends" informs us of the social position of the media within a strategic community.

Graph analysis and content analysis: new leads?

Finally, graph analysis can be used to understand latent argumentation structure. Indeed, positions aim to construct reality by punctual intervention. Ultimately, all these impressionistic interventions form an overall picture that aims to construct a reality. The mapping of keywords then reveals a dictionary structure that is the latent product of the construction of reality through speech in given media. Dynamic argument mapping allows for tracing the genesis of argumentative strategies. The use of topological variables in graphs also allows us to understand the role and place of each concept in a global argumentative strategy. In reality, the subjects that take a position in controversy are in their very large majority spokespeople who inhabit preexisting speeches and work on the margin. The controversy rarely sees the innovative creation of arguments and much more often a position on argument trees that are already there in produced statements like the same ones. It mainly allows locating of emergence and innovation, the diffusion or even the virality of certain concepts.

Influence position: a multifaceted approach

We wanted to develop the concept of "influence position" based on a multi-faceted classification that is based on 4 axes: the socio-media identification of the speaker, the classification of produced content, the social resonance of these products, and their authority within structured community networks. By adding the dates of entry and exit into the public debate, the analysis of positions includes the dynamic, even genetic dimension in the construction of speeches. It is a matter of combining the constructivist and behavioral approaches to define the public space in its tension between fields of actors. Indeed, the social nature of these speakers, the subjects addressed, the arguments used, the publication dates and the media outlets allow us to identify profile types that are as many "occupied territories" in the space of ideas. Behind a fairly significant number of messages, it is not uncommon to identify fewer than ten typical positions that share the discursive public space.

Conclusion

We would like to conclude this article with 3 case studies that illustrate our first uses in SIC of the methodology developed around My Web Intelligence. Our first published research was to understand the structuring of the community of digital humanities researchers through a socio-pragmatic analysis of their online intervention (Lakel and Deuff 2017). Through the study of 1800 published pages and 703 areas of expression, we tried to understand the tension between a "hack" type socialization based on a web communication culture (institutional openness of the right to speak, agile association around events and projects, intensive use of independent authors' web 2.0) and a more institutional "yack" type socialization based on more traditional communication from major research institutions. Our second study was dedicated to the study of another form of communication disruption in the field of health information (Lakel 2019). In collaboration with the CIC of the CHU of Bordeaux, we observed the discourses on children's asthma on the Internet. In the context of the deregulation of speech on health issues, health institutions are faced with the rise of anti-scientific criticism that ranges from the rejection of medication to the proposal of alternative, natural or mystical therapies. Through the analysis of 1235 pages and 846 domains, we wanted to discover the lines of fractures between institutional words and the anti-science movement. Our 3rd study, after the scientific field and the health field, relates to the political field (Cormerais and Lakel, 2021). Juan Branco, a political pamphleteer classified in the radical left-wing, managed to create a buzz in the 1st half of 2019 thanks to his book Crépuscule. Initially published in a free self-edition, the book reissued in a paper format becomes N°1 in sales for several weeks. Initially relying on an individual self-promotion strategy, the author and his book become the object of a controversy that opposes, on the one hand, media activists and "pure player" digital press from the new web editorialization, and on the other hand, secular press.

The 3 studies first have a methodological unity: operationalizing the socio-linguistic concept of "position" which aims to correlate sociological and institutional factors, pragmatic factors, and content produced in a controversial dynamic. They also revolve around a common problem: understanding how the Internet as a new media transforms the conditions of possibility for intervention in the public space. The sociology of the web studies the socio-pragmatic upheavals that occur in the construction of the ideological horizons that underlie living together. My Web Intelligence is intended to be a methodological proposition but also a field of study that contributes to the understanding of this major societal change. At the end of the prototyping project, My Web Intelligence will have addressed most of the initial technical and scientific challenges. Nonetheless, the transition from prototype to stable and professional service entails its own challenges: blocking algorithmic squares, Big data management, multi-user support, etc. The prototype is today more of a call for a research program than a finished and self-contained offer.

Indicative bibliography

Barabasi, Albert-Laszlo. 2005. « The Architecture of Complexity: The Structure and the Dynamics of Networks, from the Web to the Cell ». P. 3‑3 in Proceedings of the

Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD ’05. New York, NY, USA: ACM.

Björneborn, Lennart. 2004. « Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach ».

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, et Etienne Lefebvre. 2008. « Fast Unfolding of Communities in Large Networks ». Journal of Statistical Mechanics: Theory and Experiment 2008(10):P10008. doi: 10.1088/1742-5468/2008/10/P10008.

Cormerais, Franck, et Amar Lakel. 2019. « Recherches digitales et production des données, bouleversement des agencements pour le chercheur en SIC ». Etudes Digitales 3(6):94‑111.

Cormerais, Franck, Olivier Le Deuff, Amar Lakel, et David Pucheu. 2016. « Les SIC à l’épreuve du digital et des Humanités : des origines, des concepts, des méthodes et des outils ». Revue française des sciences de l’information et de la communication (8). doi: 10.4000/rfsic.1820.

Cormerais, Franck, Olivier Le Deuff, Amar Lakel, et David Pucheu. 2017. « L’école et l’avenir de la culture digitale ». Hermès, La Revue 78(2):87‑95.

Le Deuff, Olivier. 2012. « Littératies informationnelles, médiatiques et numériques : de la concurrence à la convergence ? » Études de communication. langages, information, médiations (38):131‑47. doi: 10.4000/edc.3411.

Le Deuff, Olivier. 2015. Les humanités digitales précèdent-elle le numérique ? Iste éditions.

Kleinberg, Jon M. 1999. « Authoritative Sources in a Hyperlinked Environment ». J. ACM 46(5):604‑32. doi: 10.1145/324133.324140.

Lakel, Amar. 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.

Lakel, Amar, et Olivier Le Deuff. 2017. « À quoi peut bien servir l’analyse du web ? » Les Cahiers du numérique 13(3):39‑62.

Poibeau, Thierry. 2014. « Le traitement automatique des langues pour les sciences sociales, Automatic language processing for the social sciences ». Réseaux (188):25‑51. doi: 10.3917/res.188.0025.

Scott, John. 2017. Social Network Analysis. 4 edition. Thousand Oaks, CA: SAGE Publications Ltd.

de Surmont, Jean-N. 2005. « La théorie des jeux et la pensée en réseau. Dynamique hypertextuelle et réticulaire ». Communication. Information médias théories pratiques (Vol. 24/1). doi: 10.4000/communication.3326.

Wasserman, Stanley, et Katherine Faust. 1994. Social Network Analysis: Methods and Applications. 1re éd. Cambridge University Press.