




My Web Intelligence: Enunciation-Level Web Crawling for Authentic Controversy Mapping

Amar Lakel ¹

¹ MICA Laboratory, Université Bordeaux Montaigne, France

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Open Journals](#) 

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

My Web Intelligence (MWI) is an open-source Python tool that introduces a fundamental methodological shift in digital controversy mapping: extraction at the **enunciation level** rather than the page level. Unlike existing web crawlers that extract all hyperlinks from HTML pages (including navigation, advertisements, widgets, and CMS-generated links), MWI extracts only the links present within the readable content — the actual discourse produced by authors. This distinction operationalizes the difference between technical traces and intentional citation acts, producing authentic cartographies of controversies rather than maps of web infrastructure. MWI further extends this enunciative approach through paragraph-level embeddings and Natural Language Inference, enabling semantic network analysis at the granularity of argumentative units.

Statement of Need

A Two-Decade Methodological Gap

MWI addresses a methodological gap identified in webometrics literature over two decades ago. Henzinger, Motwani, and Silverstein (2002) noted that “there has not been much research on link types, and although such research is needed since it may facilitate distinguishing commercial from editorial links or links to metainformation from links that relate to the actual content of the site.”

Bar-Ilan (2005) subsequently proposed a multi-faceted framework for hyperlink classification that explicitly included “**link area**” as a key analytical dimension — the position of a hyperlink within page structure (content body, navigation, sidebar, footer). This framework acknowledged that the location of a link carries methodological significance: a link placed within argumentative prose represents a different speech act than a link placed in a navigation menu.

Despite this theoretical recognition, **no social science web crawler has operationalized this distinction**. Tools such as Hyphe (Jacomy et al., 2016), IssueCrawler (Rogers, 2010), Navicrawler (Jacomy, 2006), and VOSON (Ackland, 2013) extract all hyperlinks present in the HTML source of a page. This includes:

- Navigation links (menus, headers, footers)
- Advertising and affiliate links
- Social media widgets and share buttons
- CMS-generated “related articles” links
- Sidebar and footer links to unrelated content

When researchers use these tools to map controversies, they inadvertently produce **cartographies of web infrastructure** rather than cartographies of discursive exchange. A controversy, in the sociological sense (Venturini, 2010), consists of intentional argumentative acts between

40 enunciators — citations, refutations, endorsements. These acts occur within the body of texts,
41 not in navigation menus.

42 MWI's Methodological Innovation

43 MWI addresses this fundamental gap through what we term **enunciation-level extraction**. The
44 tool first extracts the “readable” content of each page using boilerplate removal algorithms,
45 isolating the actual text produced by the author from the surrounding technical apparatus.
46 Links are then extracted only from this readable content, capturing intentional citation acts
47 rather than technical URL presence.

48 This approach operationalizes the theoretical framework of “algorithmic hermeneutics” (Lakel,
49 2024) and “augmented enunciative pragmatics” — treating web data as traces of discursive
50 production requiring interpretation, not self-sufficient network data. The distinction between
51 **enunciative links** (intentional citations within discourse) and **page-level links** (all URLs in
52 HTML) constitutes MWI's core methodological contribution.

53 The tool has been deployed for controversy analysis of the French “Gilets Jaunes” movement
54 (30,000 pages, 200,000 paragraphs, October 2018 – November 2019), revealing counter-
55 intuitive patterns invisible to page-level analysis: mainstream media informational dominance
56 despite the movement's purportedly “digital native” character.

57 Historical Development and Prior Art

58 MWI's core methodology — extracting hyperlinks only from readable content rather than full
59 HTML — has been implemented since 2014, as documented in the original prototype funded
60 by the Nouvelle-Aquitaine Regional “Big Data” call for projects. The complete development
61 history is preserved in Software Heritage (Lakel & Bruant, 2014–2016).

62 The conceptual distinction between “relevant links” (*liens pertinents*) and structural page
63 links was explicitly articulated in the 2014 project presentation (Lakel, 2015), which described
64 the platform's goal to “restructure data analyzed by mapping relevant links” (*restructurer les*
65 *données analysées par la cartographie des liens pertinents*). This formulation demonstrates
66 that the enunciative extraction methodology predates both:

- 67 ■ The mainstream adoption of boilerplate removal tools in web mining (trafilatura
68 (Barbaresi, 2021), 2019)
- 69 ■ Systematic attention to content extraction in digital methods literature

70 The methodology was first publicly presented at DHNord 2016 (Lakel & Le Deuff, 2016)
71 alongside a comparative workshop with Hyphe, establishing MWI as a methodological alternative
72 to page-level extraction tools. The article in *Les Cahiers du numérique* (Lakel & Le Deuff, 2017)
73 explicitly noted: “For total page link extraction, refer to Hyphe software” (*Pour l'extraction total*
74 *des liens de la page se reporter au Logiciel Hyphe*) — demonstrating conscious methodological
75 differentiation.

Date	Milestone	Documentation
2014	Regional funding, prototype development	SlideShare presentation (Lakel, 2015)
2015	Public presentation CHU Bordeaux	SlideShare (5,000+ views)
2016	DHNord conference + workshop	HAL hal-03351672 (Lakel & Le Deuff, 2016)
2016	Code archived	Software Heritage (Lakel & Bruant, 2014–2016)

Date	Milestone	Documentation
2017	Methodological publication	<i>Les Cahiers du numérique</i> (Lakel & Le Deuff, 2017)
2021	Software paper (French)	<i>I2D</i> (Lakel, 2021)
2026	MWI v2 with embeddings/NLI	This paper, Zenodo (Lakel, 2026)

Functionality

Enunciation-Level Corpus Constitution (Core Innovation)

- **Readable content extraction:** Boilerplate removal isolates author-produced text from page infrastructure (navigation, ads, widgets, CMS elements)
- **Enunciative link extraction:** Hyperlinks extracted exclusively from readable content, capturing intentional citations rather than technical URL presence
- **Focus crawling on discourse:** Depth crawling follows only enunciative links, building corpora of discursive exchange rather than web topology
- **Search engine bootstrapping:** Corpus seeding via SerpAPI (Google, Bing, DuckDuckGo) with temporal filtering
- **Relevance qualification:** Lemma-based scoring with optional LLM validation (OpenRouter) operating on readable content only

Paragraph-Level Semantic Analysis

- **Paragraph extraction:** Readable content segmented into discrete enunciative units
- **Embeddings generation:** Multi-provider vectorization (OpenAI, Mistral, Gemini, HuggingFace, Ollama) at paragraph granularity
- **Semantic similarity:** Three scalable methods (exact cosine, LSH approximate, FAISS ANN)
- **Natural Language Inference:** Cross-encoder classification (mDeBERTa XNLI multilingual) producing entailment/neutral/contradiction relations between paragraph pairs
- **Pseudolinks:** Semantic connections between paragraphs across documents, extending enunciative analysis beyond explicit citation to implicit argumentative relations

Network Export and Reproducibility

- **Multi-level aggregation:** Paragraph pairs, expression (page), and domain-level projections
- **Export formats:** CSV, GEXF (Gephi-compatible), raw corpus with full audit trail
- **Docker infrastructure:** One-command reproducible deployment
- **Database migrations:** Schema versioning for longitudinal studies

State of the Field

MWI is, to our knowledge, the **first and only robust open-source web crawling tool** that distinguishes enunciative links from page-level links for social science research:

Tool	Extraction Level	Link Source	Methodological Basis
Hyphe (Jacomy et al., 2016)	Page HTML	All page links	Web entity curation

Tool	Extraction Level	Link Source	Methodological Basis
IssueCrawlerPage HTML (Rogers, 2010)		All page links	Co-link analysis
Navicrawler Page HTML (Jacomy, 2006)		All page links	Manual navigation
VOSON Page HTML (Ackland, 2013)		All page links	Hyperlink networks
MWI	Readable content	Enunciative links only	Discursive exchange

This distinction has significant implications for controversy studies. Existing tools produce networks where nodes (pages/domains) are connected by edges that mix intentional citations with navigational artifacts. MWI produces networks where edges represent exclusively the citation acts performed by authors within their discourse — the actual fabric of controversies.

The distinction operationalizes Bar-Ilan's (2005) theoretical framework, which identified “link area” as a key classification dimension but noted that operational tools had not yet implemented this distinction. MWI fills this gap, providing researchers with the first tool capable of building networks based on discursive intentionality rather than technical HTML structure.

The paragraph-level pseudolinks feature extends this approach, detecting semantic relations (entailment, contradiction, neutrality) between argumentative units across the corpus, enabling cartography of implicit argumentative structures beyond explicit hyperlink citation.

Research Applications

MWI has been deployed in peer-reviewed research with publicly archived datasets (Nakala/Humanum):

- **Health information ecosystem mapping:** enunciative networks of medical authority (Lakel, 2017, 2020, 2022)
- **Digital humanities community:** citation practices vs. institutional linking (Lakel, 2016; Lakel & Le Deuff, 2016, 2017)
- **Gilets Jaunes controversy:** mainstream media hegemony revealed through enunciative analysis (30,000 pages, 200,000 paragraphs) (Lakel, 2019)
- **Intellectual influence networks:** discourse-level rather than page-level citation (Cormerais & Lakel, 2023; Lakel, 2023)
- **Automatic classification of digital corpora:** interdisciplinary problematization (Lakel, 2024)
- **Communication sciences methodology:** digital methods epistemology (Cormerais et al., 2016; Cormerais & Lakel, 2018)

All research datasets are openly available with DOI identifiers on the Nakala platform (French national infrastructure for humanities data).

Acknowledgements

MWI development was supported by the MICA Laboratory at Université Bordeaux Montaigne and the Nouvelle-Aquitaine Region (2014 “Big Data” call for projects). The author thanks Franck Cormerais, Olivier Le Deuff, Nathalie Pinede and the E3D research group for theoretical

discussions on enunciative pragmatics, David Bruant for foundational software development (2014-2016), and Jean Devalance for contributions to MWI python version.

References

- Ackland, R. (2013). Web social science: Concepts, data and tools for social scientists in the digital age. *SAGE Publications*.
- Barbaresi, A. (2021). Trafilatura: A web scraping library and command-line tool for text discovery and extraction. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 122–131. <https://doi.org/10.18653/v1/2021.acl-demo.15>
- Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing & Management*, 41(4), 973–986. <https://doi.org/10.1016/j.ipm.2004.02.005>
- Cormerais, F., & Lakel, A. (2018). Recherches digitales et production des données, bouleversement des agencements pour le chercheur en SIC. *Études Digitales*, 6, 155–179.
- Cormerais, F., & Lakel, A. (2023). Juan branco, influenceur éphémère ou figure d'un nouvel « intellectuel numérique » ? *Quaderni*, 109, 39–58. <https://doi.org/10.4000/quaderni.2731>
- Cormerais, F., Le Deuff, O., Lakel, A., & Pucheu, D. (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. <https://doi.org/10.4000/rfsic.1820>
- Henzinger, M. R., Motwani, R., & Silverstein, C. (2002). Challenges in web search engines. *ACM SIGIR Forum*, 36(2), 11–22.
- Jacomy, M. (2006). *Navicrawler*. Sciences Po médialab.
- Jacomy, M., Girard, P., Ooghe-Tabanou, B., & Venturini, T. (2016). Hyphe, a curation-oriented approach to web crawling for the social sciences. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1), 595–598.
- Lakel, A. (2015). *My web intelligence : Une plateforme open source au service des humanités digitales*. SlideShare. <https://www.slideshare.net/alakel/my-web-intelligence-une-plateforme-open-source-au-service-des-humanits-digitales>
- Lakel, A. (2016). *French digital humanities web communities dataset*. Nakala (Huma-Num). <https://doi.org/10.34847/nkl.f43by03n>
- Lakel, A. (2017). *Health information ecosystem: Childhood asthma dataset*. Nakala (Huma-Num). <https://doi.org/10.34847/nkl.0f3a9710>
- Lakel, A. (2019). *Yellow vests online controversy dataset (nov 2018 - nov 2019)*. Nakala (Huma-Num). <https://doi.org/10.34847/nkl.0bfeq252>
- Lakel, A. (2020). 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. <https://doi.org/10.4000/rfsic.8376>
- Lakel, A. (2021). My web intelligence : Un outil pour l'analyse du web et des réseaux. *I2D – Information, Données & Documents*, 2021/1(1), 96–103. <https://doi.org/10.3917/i2d.211.0096>
- Lakel, A. (2022). Health literacy in complex digital information environments. In *Health information science*.
- Lakel, A. (2023). *Juan branco: Digital influencer analysis dataset*. Nakala (Huma-Num).

- 182 <https://doi.org/10.34847/nkl.c4fc83mv>
- 183 Lakel, A. (2024). Classification automatique des grands corpus numériques : Une
184 problématisation interdisciplinaire. *Essais*, 21. <https://doi.org/10.4000/essais.12989>
- 185 Lakel, A. (2026). *My Web Intelligence* (Version 1.0.0). Zenodo. <https://doi.org/10.5281/zenodo.18376429>
- 186
- 187 Lakel, A., & Bruant, D. (2014--2016). *MyWebIntelligence v1.0*. Software Heritage. <https://archive.softwareheritage.org/swh:1:snp:a3f4ddb6a7e689c582811e36a87c3e3950ec0857>
- 188
- 189 Lakel, A., & Le Deuff, O. (2016). Cartographie web de la communauté francophone
190 des humanités numériques et développement d'une méthode critique avec l'outil
191 MyWebIntelligence. *DHNord 2016 : Humanités Numériques : Théories, Débats, Approches Critiques*.
192
- 193 Lakel, A., & Le Deuff, O. (2017). À quoi peut bien servir l'analyse du web ? Les communautés
194 de sites des humanités numériques sur internet. *Les Cahiers Du Numérique*, 13(3-4),
195 107–138. <https://doi.org/10.3166/lcn.13.3-4.107-138>
- 196 Rogers, R. (2010). Mapping public web space with the Issuecrawler. In *Digital cognitive*
197 *technologies: Epistemology and knowledge society* (pp. 89–99). Wiley.
- 198 Venturini, T. (2010). Diving in magma: How to explore controversies with actor-network
199 theory. *Public Understanding of Science*, 19(3), 258–273. <https://doi.org/10.1177/0963662509102694>
- 200