

# <sup>1</sup> My Web Intelligence: Enunciation-Level Web Crawling <sup>2</sup> for Authentic Controversy Mapping

<sup>3</sup> **Amar Lakel**  <sup>1</sup>

<sup>4</sup> 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Published: unpublished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 <sup>16</sup> Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).<sup>17</sup>

## <sup>5</sup> Summary

<sup>6</sup> My Web Intelligence (MWI) is an open-source Python tool that introduces a fundamental  
<sup>7</sup> methodological shift in digital controversy mapping: extraction at the **enunciation level** rather  
<sup>8</sup> than the page level. Unlike existing web crawlers that extract all hyperlinks from HTML pages  
<sup>9</sup> (including navigation, advertisements, widgets, and CMS-generated links), MWI extracts only  
<sup>10</sup> the links present within the readable content — the actual discourse produced by authors. This  
<sup>11</sup> distinction operationalizes the difference between technical traces and intentional citation acts,  
<sup>12</sup> producing authentic cartographies of controversies rather than maps of web infrastructure. MWI  
<sup>13</sup> further extends this enunciative approach through paragraph-level embeddings and Natural  
<sup>14</sup> Language Inference, enabling semantic network analysis at the granularity of argumentative  
<sup>15</sup> units.

## Statement of Need

### A Two-Decade Methodological Gap

<sup>18</sup> MWI addresses a methodological gap identified in webometrics literature over two decades ago.  
<sup>19</sup> Henzinger, Motwani, and Silverstein (2002) noted that “there has not been much research  
<sup>20</sup> on link types, and although such research is needed since it may facilitate distinguishing  
<sup>21</sup> commercial from editorial links or links to metainformation from links that relate to the actual  
<sup>22</sup> content of the site.”

<sup>23</sup> Bar-Ilan (2005) subsequently proposed a multi-faceted framework for hyperlink classification  
<sup>24</sup> that explicitly included “link area” as a key analytical dimension — the position of a hyperlink  
<sup>25</sup> within page structure (content body, navigation, sidebar, footer). This framework acknowledged  
<sup>26</sup> that the location of a link carries methodological significance: a link placed within argumentative  
<sup>27</sup> prose represents a different speech act than a link placed in a navigation menu.

<sup>28</sup> Despite this theoretical recognition, **no social science web crawler has operationalized this**  
<sup>29</sup> **distinction**. Tools such as Hyphe (Jacomy et al., 2016), IssueCrawler (Rogers, 2010),  
<sup>30</sup> Navicrawler (Jacomy, 2006), and VOSON (Ackland, 2013) extract all hyperlinks present  
<sup>31</sup> in the HTML source of a page. This includes:

- <sup>32</sup> ▪ Navigation links (menus, headers, footers)
- <sup>33</sup> ▪ Advertising and affiliate links
- <sup>34</sup> ▪ Social media widgets and share buttons
- <sup>35</sup> ▪ CMS-generated “related articles” links
- <sup>36</sup> ▪ Sidebar and footer links to unrelated content

<sup>37</sup> When researchers use these tools to map controversies, they inadvertently produce **cartographies**  
<sup>38</sup> **of web infrastructure** rather than cartographies of discursive exchange. A controversy, in  
<sup>39</sup> the sociological sense (Venturini, 2010), consists of intentional argumentative acts between

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

#### <sup>42</sup> MWI's Methodological Innovation

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

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

<sup>52</sup> The tool has been deployed for controversy analysis of the French “Gilets Jaunes” movement  
<sup>53</sup> (30,000 pages, 200,000 paragraphs, October 2018 – November 2019), revealing counter-  
<sup>54</sup> intuitive patterns invisible to page-level analysis: mainstream media informational dominance  
<sup>55</sup> despite the movement’s purportedly “digital native” character.

#### <sup>57</sup> Historical Development and Prior Art

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

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

- <sup>67</sup> ▪ The mainstream adoption of boilerplate removal tools in web mining (*trafilatura*  
<sup>68</sup> ([Barbaresi, 2021](#)), 2019)
- <sup>69</sup> ▪ Systematic attention to content extraction in digital methods literature

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

| Date | Milestone                               | Documentation   |
|------|---|---|
| 2014 | Regional funding, prototype development | SlideShare presentation ( <a href="#">Lakel, 2015</a> )             |
| 2015 | Public presentation CHU Bordeaux        | SlideShare (5,000+ views)   |
| 2016 | DHNord conference + workshop            | HAL hal-03351672 ( <a href="#">Lakel &amp; Le Deuff, 2016</a> )     |
| 2016 | Code archived                           | Software Heritage ( <a href="#">Lakel &amp; Bruant, 2014–2016</a> ) |

| Date | Milestone                  | Documentation  |
|------|----------------------------|--|
| 2017 | Methodological publication | <i>Les Cahiers du numérique</i> ( <a href="#">Lakel &amp; Le Deuff, 2017</a> ) |
| 2021 | Software paper (French)    | <i>I2D</i> ( <a href="#">Lakel, 2021</a> )                                     |
| 2026 | MWI v2 with embeddings/NLI | This paper, Zenodo ( <a href="#">Lakel, 2026</a> )                             |

## 76     Functionality

### 77     Enunciation-Level Corpus Constitution (Core Innovation)

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

### 88     Paragraph-Level Semantic Analysis

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

### 98     Network Export and Reproducibility

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

## 103    State of the Field

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

| Tool   | Extraction Level | Link Source    | Methodological Basis |
|--|------------------|----------------|----------------------|
| Hypse<br>( <a href="#">Jacomy et al., 2016</a> ) | Page HTML        | All page links | Web entity curation  |

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

<sup>106</sup> This distinction has significant implications for controversy studies. Existing tools produce  
<sup>107</sup> networks where nodes (pages/domains) are connected by edges that mix intentional citations  
<sup>108</sup> with navigational artifacts. MWI produces networks where edges represent exclusively the  
<sup>109</sup> citation acts performed by authors within their discourse — the actual fabric of controversies.

<sup>110</sup> The distinction operationalizes Bar-Ilan's (2005) theoretical framework, which identified "link  
<sup>111</sup> area" as a key classification dimension but noted that operational tools had not yet implemented  
<sup>112</sup> this distinction. MWI fills this gap, providing researchers with the first tool capable of building  
<sup>113</sup> networks based on discursive intentionality rather than technical HTML structure.

<sup>114</sup> The paragraph-level pseudolinks feature extends this approach, detecting semantic relations  
<sup>115</sup> (entailment, contradiction, neutrality) between argumentative units across the corpus, enabling  
<sup>116</sup> cartography of implicit argumentative structures beyond explicit hyperlink citation.

## <sup>117</sup> Research Applications

<sup>118</sup> MWI has been deployed in peer-reviewed research with publicly archived datasets (Nakala/Huma-  
<sup>119</sup> Num):

- <sup>120</sup> **▪ Health information ecosystem mapping:** enunciative networks of medical authority (Lakel,  
<sup>121</sup> 2017, 2020, 2022)
- <sup>122</sup> **▪ Digital humanities community:** citation practices vs. institutional linking (Lakel, 2016;  
<sup>123</sup> Lakel & Le Deuff, 2016, 2017)
- <sup>124</sup> **▪ Gilets Jaunes controversy:** mainstream media hegemony revealed through enunciative  
<sup>125</sup> analysis (30,000 pages, 200,000 paragraphs) (Lakel, 2019)
- <sup>126</sup> **▪ Intellectual influence networks:** discourse-level rather than page-level citation (Cormerais  
<sup>127</sup> & Lakel, 2023; Lakel, 2023)
- <sup>128</sup> **▪ Automatic classification of digital corpora:** interdisciplinary problematization (Lakel,  
<sup>129</sup> 2024)
- <sup>130</sup> **▪ Communication sciences methodology:** digital methods epistemology (Cormerais et al.,  
<sup>131</sup> 2016; Cormerais & Lakel, 2018)

<sup>132</sup> All research datasets are openly available with DOI identifiers on the Nakala platform (French  
<sup>133</sup> national infrastructure for humanities data).

## <sup>134</sup> Acknowledgements

<sup>135</sup> MWI development was supported by the MICA Laboratory at Université Bordeaux Montaigne  
<sup>136</sup> and the Nouvelle-Aquitaine Region (2014 "Big Data" call for projects). The author thanks  
<sup>137</sup> Franck Cormerais, Olivier Le Deuff, Nathalie Pinede and the E3D research group for theoretical

<sup>138</sup> discussions on enunciative pragmatics, David Bruant for foundational software development  
<sup>139</sup> (2014–2016), and Jean Devalance for contributions to MWI python version.

## <sup>140</sup> References

- <sup>141</sup> Ackland, R. (2013). Web social science: Concepts, data and tools for social scientists in the  
<sup>142</sup> digital age. *SAGE Publications*.
- <sup>143</sup> Barbaresi, A. (2021). Trafilatura: A web scraping library and command-line tool for text  
<sup>144</sup> discovery and extraction. *Proceedings of the 59th Annual Meeting of the Association for*  
<sup>145</sup> *Computational Linguistics and the 11th International Joint Conference on Natural Language*  
<sup>146</sup> *Processing: System Demonstrations*, 122–131. <https://doi.org/10.18653/v1/2021.acl-demo.15>
- <sup>148</sup> Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying  
<sup>149</sup> links in academic environments. *Information Processing & Management*, 41(4), 973–986.  
<sup>150</sup> <https://doi.org/10.1016/j.ipm.2004.02.005>
- <sup>151</sup> Cormerais, F., & Lakel, A. (2018). Recherches digitales et production des données,  
<sup>152</sup> bouleversement des agencements pour le chercheur en SIC. *Études Digitales*, 6, 155–179.
- <sup>153</sup> Cormerais, F., & Lakel, A. (2023). Juan branco, influenceur éphémère ou figure d'un nouvel «  
<sup>154</sup> intellectuel numérique » ? *Quaderni*, 109, 39–58. <https://doi.org/10.4000/quaderni.2731>
- <sup>155</sup> Cormerais, F., Le Deuff, O., Lakel, A., & Pucheu, D. (2016). Les SIC à l'épreuve du digital et  
<sup>156</sup> des humanités : Des origines, des concepts, des méthodes et des outils. *Revue Française Des*  
<sup>157</sup> *Sciences de l'information Et de La Communication*, 8. <https://doi.org/10.4000/rfsic.1820>
- <sup>158</sup> Henzinger, M. R., Motwani, R., & Silverstein, C. (2002). Challenges in web search engines.  
<sup>159</sup> *ACM SIGIR Forum*, 36(2), 11–22.
- <sup>160</sup> Jacomy, M. (2006). *Navicrawler*. Sciences Po médialab.
- <sup>161</sup> Jacomy, M., Girard, P., Ooghe-Tabanou, B., & Venturini, T. (2016). Hyphe, a curation-  
<sup>162</sup> oriented approach to web crawling for the social sciences. *Proceedings of the International*  
<sup>163</sup> *AAAI Conference on Web and Social Media*, 10(1), 595–598.
- <sup>164</sup> Lakel, A. (2015). *My web intelligence : Une plateforme open source au service des*  
<sup>165</sup> *humanités digitales*. SlideShare. <https://www.slideshare.net/alakel/my-web-intelligence-une-plateforme-open-source-au-service-des-humanites-digitales>
- <sup>167</sup> Lakel, A. (2016). *French digital humanities web communities dataset*. Nakala (Huma-Num).  
<sup>168</sup> <https://doi.org/10.34847/nkl.f43by03n>
- <sup>169</sup> Lakel, A. (2017). *Health information ecosystem: Childhood asthma dataset*. Nakala (Huma-  
<sup>170</sup> Num). <https://doi.org/10.34847/nkl.0f3a97l0>
- <sup>171</sup> Lakel, A. (2019). *Yellow vests online controversy dataset (nov 2018 - nov 2019)*. Nakala  
<sup>172</sup> (Huma-Num). <https://doi.org/10.34847/nkl.0bfeq252>
- <sup>173</sup> Lakel, A. (2020). Prises de positions et influences sur le web : Le cas de l'information  
<sup>174</sup> de santé. *Revue Française Des Sciences de l'information Et de La Communication*, 18.  
<sup>175</sup> <https://doi.org/10.4000/rfsic.8376>
- <sup>176</sup> Lakel, A. (2021). My web intelligence : Un outil pour l'analyse du web et des réseaux. *I2D –*  
<sup>177</sup> *Information, Données & Documents*, 2021/1(1), 96–103. <https://doi.org/10.3917/i2d.211.0096>
- <sup>179</sup> Lakel, A. (2022). Health literacy in complex digital information environments. In *Health*  
<sup>180</sup> *information science*.
- <sup>181</sup> 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    Lakel, A., & Bruant, D. (2014--2016). *MyWebIntelligence v1.0*. Software Heritage. <https://archive.softwareheritage.org/swh:1:snp:a3f4ddb6a7e689c582811e36a87c3e3950ec0857>
- 187    Lakel, A., & Le Deuff, O. (2016). Cartographie web de la communauté francophone  
188         des humanités numériques et développement d'une méthode critique avec l'outil  
189         MyWebIntelligence. *DHNord 2016 : Humanités Numériques : Théories, Débats, Approches*  
190         *Critiques*.
- 191    Lakel, A., & Le Deuff, O. (2017). À quoi peut bien servir l'analyse du web ? Les communautés  
192         de sites des humanités numériques sur internet. *Les Cahiers Du Numérique*, 13(3-4),  
193         107–138. <https://doi.org/10.3166/lcn.13.3-4.107-138>
- 194    Rogers, R. (2010). Mapping public web space with the Issuecrawler. In *Digital cognitive*  
195         *technologies: Epistemology and knowledge society* (pp. 89–99). Wiley.
- 196    Venturini, T. (2010). Diving in magma: How to explore controversies with actor-network  
197         theory. *Public Understanding of Science*, 19(3), 258–273. <https://doi.org/10.1177/0963662509102694>
- 198    199    200