Hyperlink is not dead!

Benjamin Ooghe-Tabanou
Sciences Po, médialab
Paris, France
benjamin.ooghe@sciencespo.fr

Mathieu Jacomy
Sciences Po, médialab
Paris, France
mathieu.jacomy@gmail.com

Paul Girard
Sciences Po, médialab
Paris, France
paul.girard@sciencespo.fr

Guillaume Plique
Sciences Po, médialab
Paris, France
guillaume.plique@sciencespo.fr

ABSTRACT

The emergence and success of web platforms nurtured a trend within social studies: "Hyperlink is dead!". Capturing their users into mobile applications and specialised web interface to propose them a specific user experience (and business model), the platforms indeed created new information silos in the open World Wide Web space. The simplified availability of user behavioural data through these platforms APIs reinforced this idea in academic communities by providing scholars with an easy way to collect rich user centric data for their research. After discussing the methodological aspects of the web divide between platforms and classical websites, we will argue that although it becomes more and more invisible, the hyperlink, modern incarnation of intertextual links between documents, is still a central and structural element of the web. Hyperlinks remain an invaluable resource to turn the web into a research field in spite of the complexity to collect, manipulate and curate them. We will illustrate those methodological challenges by describing the choices we made in designing Hyphe, a tool dedicated to the creation of web corpora tailored for mining hypertexts.

CCS CONCEPTS

• **Information systems → Web mining**; **Web applications**; **Internet communications tools**;

KEYWORDS

Hyperlink, hypertext, web mining, crawler, corpus, curation, network analysis.

ACM Reference Format:
Benjamin Ooghe-Tabanou, Mathieu Jacomy, Paul Girard, and Guillaume Plique. 2018. Hyperlink is not dead!. In *International conference on Web Studies (WS.2 2018), October 3–5, 2018, Paris, France.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3240431.3240434

1 ARE HYPERTEXT STUDIES OUTDATED?

The World Wide Web's original design as a vast open documentary space built around the concept of hypertext made it a fantastic research field to study networks of actors. As a literary technology, hyperlinks concepts are anything but new: "*Links are intrinsic to documents, and have been for millennia* " [27]. According to Bardini [5], two main concepts are embedded in hypertexts: association and connection. Hypertexts allow to create conceptual associations between documents - a powerful function when used in a free and creative context - but they can also be very efficient in connecting documents to communicate ideas (i.e. conceptual links) within a community.

When Tim Berners-Lee proposed his World Wide Web (WWW) project, hypertexts were presented more as a way to connect documents to enhance communication through navigation than as form of conceptual associations: "*The texts are linked together in a way that one can go from one concept to another to find the information one wants. (…) The process of proceeding from node to node is called navigation*" [6]. But following Bardini [5], if associative hyperlinks are created freely by authors for their own use, connective ones which have a value within a community are more likely to be moderated. Although the goal of the WWW is to enhance information flows across communities, connections between documents are not controlled, as each individual website's author is responsible only for the connections from his website to the rest of the WWW.

1.1 Hyperlinks directionality: a bottom-up hierarchy

This directionality of the links reveals asymmetrical associations between the linked documents: the referer knows the referee but not necessarily the other way around. Considering hyperlinks as references provides powerful insights on the distribution of influence on the web. As the study of complex networks has demonstrated, online connections are not randomly distributed across the web. According to the principle known as the "Matthew effect" or "preferential attachment" [4], new web documents tend to cite the already most cited documents, reinforcing the concentration of links to a small fraction of pages. A hierarchy naturally emerges from this pattern across all scales of the web: it can be observed locally (eg. inside Wikipedia) as well as in its general structure. This hierarchy is bottom-up because it emerges spontaneously rather than by design, but also because hyperlinks tend to flow from a metaphorical bottom to a metaphorical top. Actors with high visibility drawing most of citations are a handful compared to the mass of low visibility actors who cite them. The structure emerging from the direction of hyperlinks was famously leveraged by Google's

WS.2 2018, October 3–5, 2018, Paris, France

ACM ISBN 978-1-4503-6438-6/18/10.
https://doi.org/10.1145/3240431.3240434

PageRank algorithm [28]. The efficiency of this idea lies in the pervasiveness of citation asymmetries in the web. The list of search engines results displays a reduced version of the web, its head being massively clicked while its endless tail remains unbrowsed [23].

Hyperlinks progressively turned into a strategic resource for websites in their search for online visibility. The WWW becoming a very popular media and Page Rank-like strategies becoming the only viable approach, search engines settled as inevitable gateway for the rest of the web [21]. This was soon confirmed by the emergence of practices such as "Google Bombing" [3] where individuals or automated robots take advantage of search engine algorithms to try and attract more visibility on specific targets. Another example is the development of Search Engine Optimisation (SEO) techniques, in which websites try either to best conform to the search engines' guidelines, or outright violate them to manipulate the algorithms [24]. The hyperlink being the structural element underpinning the web hierarchy, as well as the main way for actors to get visibility in a SEO era, it settled as the fundamental backbone of the web.

1.2 The rise of gated communities

The rapid development of web 2.0 technologies in the mid 2000's, followed by the adoption of smartphones as the main internet access device, pushed the rise of user centric websites and mobile applications and paved the way for yet another sub-field of digital sociology: platforms studies. Social networks incentivised new ways to connect users and share contents through dedicated web applications and drained more and more web traffic inside gated web spaces. They quickly made an art out of user-retaining designs, building their business models around the capture and reselling of audience. Beneficial to platforms when they allow to track user activities or sell advertisements, hyperlinks are detrimental to them when they allow users to escape their gated space. The original project of free connection and access to information is, in this sense, opposite to the increasing "*platformization of the Web*" [15].

Web applications not only change the way people use the web, they also provide new means to gather users' personal information and analyse their behaviours. Redistributing part of these digital traces through Application Programming Interfaces (API), platforms offer an extremely convenient opportunity for research to collect digital traces under the form of structured data with a user centric perspective at a large scale and a low cost [32]. Offering easy to harvest and rich information at an affordable methodological cost, social media APIs' data appeared very appealing to social researchers: more and more papers in the field of Web studies shifted their focus from hypertext analysis to the investigation of specific platforms (Twitter studies, Facebook studies…), picking the right ones depending on their research affordance [35]. Nevertheless, the source of social platforms is progressively drying up: ethical and legal issues on the one hand, and commercial objectives on the other one, are leading platforms to close their APIs or at least restrict considerably the volume and richness of the data they provide [16]. Finally, research based on private APIs is forced to entertain a risky dependency to commercial interests, which can damage protocols [30].

1.3 A mis-valued Lingua Franca

The hyperlink's role has changed and drifted "*from a navigational device into a data-rich analytical device*" [14]. Although platforms create application spaces where users can navigate without using hyperlinks, they still employ hyperlinks to establish a hierarchy among their contents and track user behaviours. What importance the hyperlinks may lose within a platform, they regain as the most central way to point to contents outside of the platform or to be referred from the external world. Hyperlinks are still the unique common way to connect to the rest of the world wide web. As an example of this phenomenon, Gerlitz and Helmond [13] showed how Facebook developed its "*like economy*" by facilitating the integration of contents from other websites through facebook-specific hyperlinks such as embedded "*share buttons*".

Over the years "*hyperlinks have become a familiar and transparent feature to users*" [2]. They were designed to make the consumption of electronic information seamless by proposing a common interface [6]. They are now the backbone of the World Wide Web. Every web information has to be connected through hyperlinks to get known and gain audience. This is true for platforms as well as for independent websites gathering communities of people sharing a hobby, a cause or an organisation. Hyperlinks remains thus a necessary research resource to grasp the diversity of online collective life and debates.

2 QUALI-QUANTITATIVE CHALLENGES OF HYPERTEXT CORPORA

Studying the web through hypertext remains necessary for many research questions. To allow this kind of research, we started designing in 2010 a research instrument allowing to study online phenomena through the prism of hyperlinks. The Hyphe [25] open source software was created to overcome the limitations of existing solutions: many tools existed to crawl the web, including some especially designed for social sciences (such as the IssueCrawler [31]), but none of them was suitable for qualitative corpus building and yet capable to scale up to large website networks. Our ambition was to help scholars crawl the web, finely delimit the web territory of each of the actors they were interested in, curate their web corpus according to their research questions, and analyse web materials through network or content analysis. Web data is both extensive and intensive, massive and detailed: whereas crawling web documents to extract large amounts of links is a quantitative task, curating a corpus, painstakingly choosing which pages to include and exclude, is, an iterative and qualitative task. According to Gabriel Tarde's lesson as modernised by Bruno Latour [22], social researchers should be equipped with the proper instruments to follow collective interactions beyond the divide between qualitative and quantitative methods. But the path to hypertext studies was long and full of obstacles.

2.1 There is no such thing as a website

Defining frontiers between actors and their online territories is a difficult decision which can change from one research to another. Hyperlinks create connections between web pages, but in the study of social phenomena, individual pages are rarely the relevant level to sieve the web as a research field. While the website would seem

like a natural way to aggregate web pages, there is no actual technical definition of a such entity [7]. URLs are used very differently across the web to define where websites start and end. Domain names, for instance, are often bad indicator of a website limits: many blogs or parts of websites representing individual actors appear as subdomains (i.e. medialab.sciencespo.fr) or even subdirectories (i.e. www.sciencespo.fr/bibliotheque/) of the domain name of the organisation to which they belong. Moreover if one scholar wants to study an organisation or individual actors through different web channels (blog, social media…) web pages should not only be aggregated by websites but across them.

This difficulty to map issues or actors on Web pages encouraged us to define a flexible web pages aggregation system to allow users creating bundles of web pages according to their research interests: we call these bundles "*web entities*". Such entities are defined as a combination of prefixes of LRUs - URLs in which the domain terms are reversed to obtain a list of stems from the most generic to the most specific (i.e. "https://medialab.sciencespo.fr/people/" becomes "http|fr|sciencespo|medialab|people") [18]. By defining web entities as groups of pages respecting the same prefix patterns, Hyphe users define their actors by drawing frontiers that let them decide how the web pages will be aggregated to best support their research questions.

2.2 Web curation: creating a corpus iteratively

To use the web as a research field, scholars need a method to select which groupings of web pages are relevant for their study. Web crawling is a method which relies on hyperlinks to discover and collect documents. Using the straightforward crawl strategy of following every links encountered (or "*snowballing*" [31]) raises two issues: "*top level magnets*" (the hyperlinks bottom-up hierarchy attracts all crawls to the same top level websites such as Google or Wikipedia, highly generic and irrelevant for many studies) and "*topic drifts*" (the very high density of links in the WWW implies that a random walk ends up exploring a very heterogeneous set of websites [10]). Both problems can be avoided by using a qualitative approach, in which scholars curate the discovered web pages by deciding to include them or not in their quantitative collection. There where automatic data harvesting fails, manual corpus building can succeed.

Hyphe proposes to this end a manual-yet-assisted data collection through a prospection process. The web crawl module only downloads and indexes the web pages belonging to a predefined web entity. External links are not followed automatically. Instead, the discovered web entities are proposed from most to least cited and the user is asked to manually prospect them and decide whether or not to include them in the corpus, turning them into web entities to be crawled. This method forms an iterative loop between web crawling and human curation, allowing to exclude top level magnets and progressively build the corpus by extending sources though hyperlink prospection, while taking care to avoid topic drifts [18].

2.3 Which data structure to manage hypertext corpora?

The World Wide Web is a vast documentary space. Building a web corpus iteratively requires a technical infrastructure able to efficiently store and retrieve data from a complex and large dataset in order to be reactive to both web crawls and human curation.

Analysing the network of hyperlinks between the actors defined as web entities requires to compute how gathered web pages assemble in web entities and how their hyperlinks aggregate into web entity links, whereas researchers might constantly change the web entities boundaries. The typical approach of indexing each page's association to a web entity, and using it to aggregate page-to-page hyperlinks into entity-to-entity links is efficient but has a drawback: redefining the boundaries of a web entity requires to reindex all concerned pages and links, which can become prohibitively long as the number of pages grows. Putting the users at the center of the data collection to let them curate their corpora demands a dynamic index which can constantly rebuild at a minimum cost the links between web entities from the data recently crawled and the latest entities definition.

Hyphe's dynamic index, "*Traph*", leverages the natural tree structure of URLs to store pages and web entities as well as the hyperlinks graph [29]. It allows to store web entities as flags placed freely on branches of the URLs tree. When a user changes the boundaries of a web entity, a flag is just moved to a different place without reindexing. Drawing inspiration from the structure known as a "*trie*" or "*prefix tree*", Traph answers efficiently heavy queries such as getting all the pages or neighbors of a web entity, getting a page's entity, or building all the entity-to-entity links. From user observations we reflected on how to implement the desired method and from there we engineered a new memory structure for such user-controlled crawlers.

2.4 Enabling digital field work by design and engineering

As in all other form of fieldwork, iteration is key in hyperlinks ethnography to sustain the researcher's empirical engagement with the web. This raises methodological and technical implications. An accessible user-interface is more than a commodity: it enables iterative curation. The quality of the results depends directly on the users' ability to monitor the outcome of automatic tasks and curate them quickly. It requires monitoring, visualisation and interactions in multiple places: selecting and setting the boundaries of web entities, categorising, etc. Hyphe's user interface was developed to this end using the most accessible option available: web technologies. Multiplatform, supporting modern interaction design, well documented and reliable, HTML5 is ideally suited to disseminate scientific methods to non specialists (as illustrated by the rise of data science notebooks such as Jupyter [20]). This investment in interaction design and engineering makes Hyphe accessible to a variety of users, including (but not exclusively) social science researchers, students and data journalists.

Hyphe also comes with an alternative rich user interface called *Hyphe Browser*, allowing users to build their corpora while browsing the web. Web technologies have a major loophole despite their benefits: web pages can refuse to be displayed inside other web

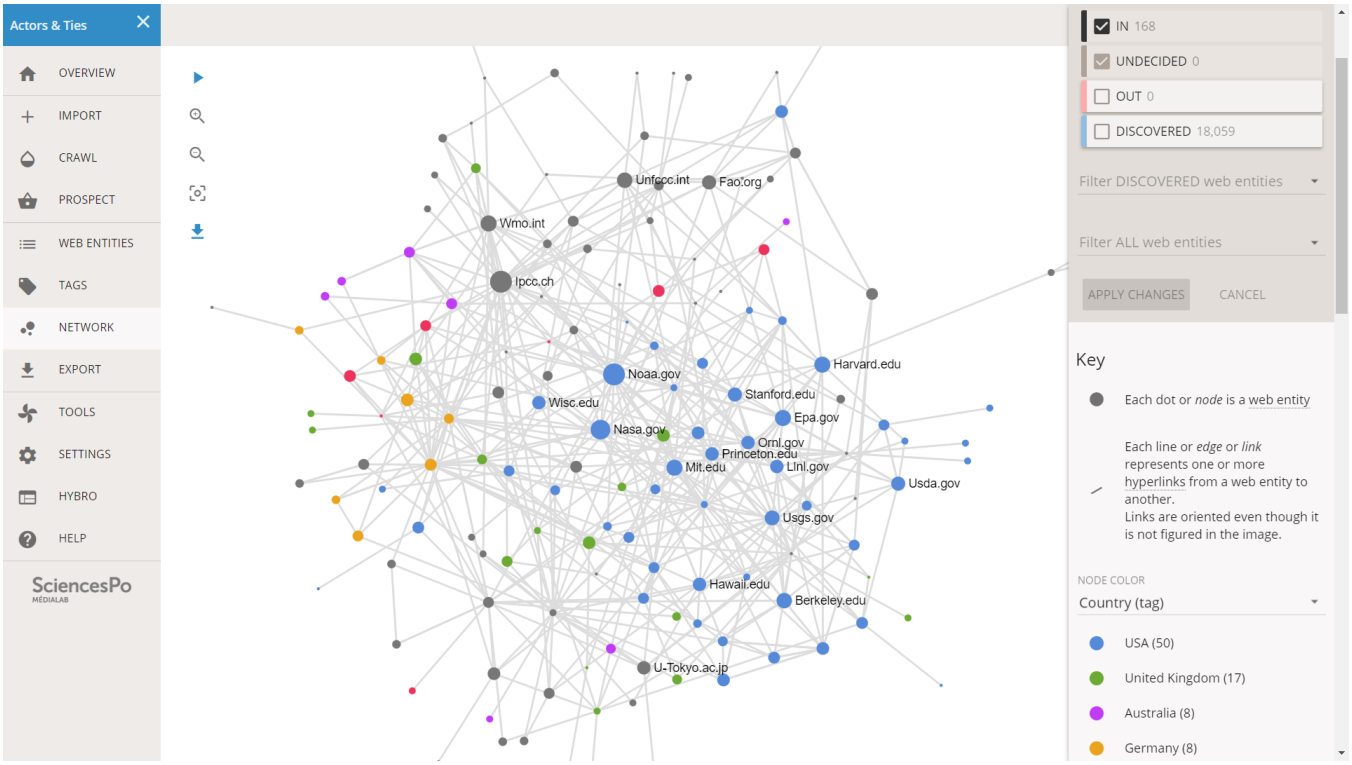

Figure 1: Hyphe's web user interface features interactive WebGL networks using the Sigma.js library

pages as "*iframes*". As a consequence, and contrary to older tools like the Navicrawler [17], Hyphe's main client does not feature browsing. However the ability to browse while curating a corpus is so useful to the researcher's empirical engagement that we developed a second user interface as an augmented web browser relying on Google Chrome's open sourced technology [1]. Plugged to the same back-end as the web interface, Hyphe Browser helps users engage with the web as a digital field: it allows to define, select, crawl and tag web entities while browsing the corpus *in-situ*. It is well attuned to a pedagogical setting and is deployed in multiple college degree courses and experimented within a few french high schools through the FORCCAST programme (Formation par la Cartographie des Controverses à l'Analyse des Sciences et des Techniques) as a way to develop critical analysis of the web and to teach the principles of scientific inquiry through web corpus creation [33].

3 THE MANY ANGLES OF HYPERTEXT STUDIES

Hyphe enables ambitious research protocols but also small-scale projects, such as getting a website's network of internal pages. It also has limitations forcing researchers to involve with other devices, for instance for text mining and language processing. More importantly, digital field work often takes place as a complement to traditional field work.

3.1 From within: web pages network and content

Unlike most other crawlers, Hyphe does not allow users to download the web pages' data, but it can still help them get engaged with their content. The distinction between information *on* the web (stored inside web pages) and information *in* the web (stored as hypertext connections) is fruitful when it comes to interpret web data. Seen as actors, websites can control published contents and who they cite, but not who cites them. The most typical use of a web corpus might be to study how their connections contradict the contents they publish. Whoever an actor pretends to be, who cites them may reveal more about who they really are. Hyphe Browser is in this matter ideal to get engaged with the contents and browse backwards to confront them to the links and their sources.

Other crawlers are more suited than Hyphe for building and storing corpora of web contents to be quantitatively analysed. Archiving contents from a crawl raises legal and technical issues that we chose to avoid (temporal dimension and page reconstruction). We suggest to use Hyphe for sourcing a corpus and then passing it to institutions specialised in archival of the web. On the other hand, full-text search and text analysis features are within reach. First experiments with natural language processing on contents crawled with Hyphe are promising [11] and certainly relevant when it comes to quali-quantitative analysis. Even within the limits of the actual function of Hyphe, it is however possible to engage with substantial social phenomena and carry out in-depth research on them, as illustrated by the two following examples.

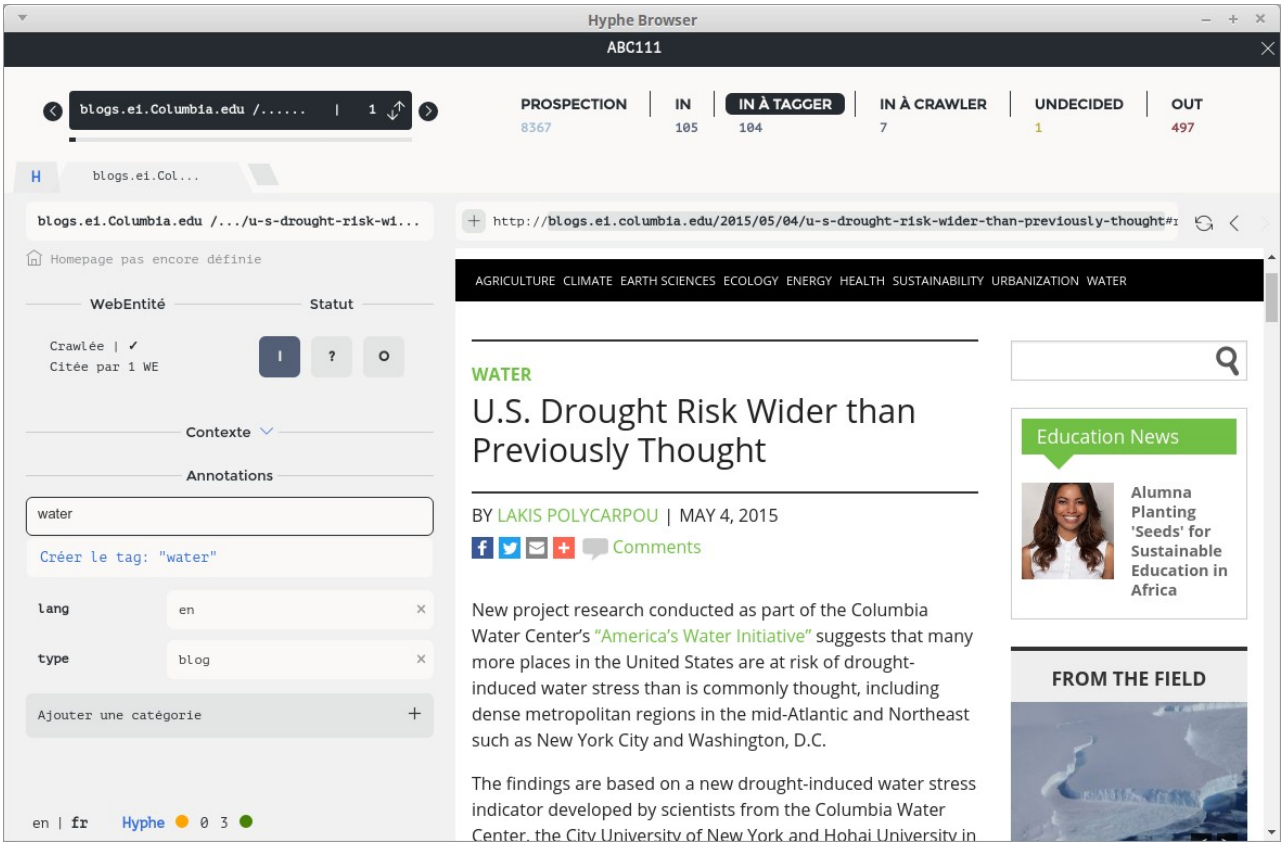

Figure 2: Hyphe's rich client is an augmented web browser

3.2 From above: clusters, opposition and affinity

Hyphe makes it easy to observe clusters of actors within the curated networks of web entities. These networks rarely exceeding a thousand nodes, web technologies can process and visualise them fluidly. Rooted in the notion of "*aggregates*" [9] where similar web pages tend to be more connected, the use of node-link diagrams for hyperlinks networks has proved beneficial when it comes to identify communities and oppositions [8, 12, 26]. Hyphe uses Gephi's layout algorithm Force Atlas 2 [19] to position the web entities so that clusters are visible. They might reveal communities of practice or interest, platforms (eg. Tumblrs tend to cite each others…), deliberate SEO strategies…

3.3 From the side: hierarchy of the digital public space

On the web, anyone can have a voice, but it does not mean that everyone will have the same impact. The web is an extremely skewed public space where a few influencers can reach a public as large as a mass media's, while most voices remain confined to narrow spheres. The direction of hyperlinks is a robust indicator of these hierarchy effects when measured with adapted metrics and visualisations. Unfortunately this directionality falls in a huge blind spot of node-link diagrams. Like paper maps flatten mountains, node-link diagrams flatten the verticality of hierarchical relations which result from the strong asymmetry of hyperlink citations. That is why, Hyphe allows also to combine links analysis with categorisations to explore the asymmetries in hyperlink directions. In an analysis with Hyphe of the Decodex (a set of media curated by journalists from Le Monde and categorised in levels of reliability), Venturini et al. [34] revealed that hyperlinks massively follow the hierarchy from less to most reliable media. This methodology can be easily replicated on other corpora, the data visualisations and statistics being directly embedded in Hyphe. This complements the classic node-link diagrams with visualisations of the hierarchical implications of hyperlinks directionality.

4 CONCLUSION

For social sciences and digital humanities, the hyperlink is alive and well! Despite the growing influence of gated platforms and their strategies of attention economy, the hyperlink is still a cornerstone of the World Wide Web's openness and accessibility. Despite the attempts of taking control over user behaviours, no one can exist on the web without hyperlinks. Even though hypertext methods are blind to some parts of the web and even though hypertext corpora raise a number of technical and methodological issues, hyperlinks remain the main resource to leverage the web as a research field. Platform based studies can be very valuable, but the hyperlink still

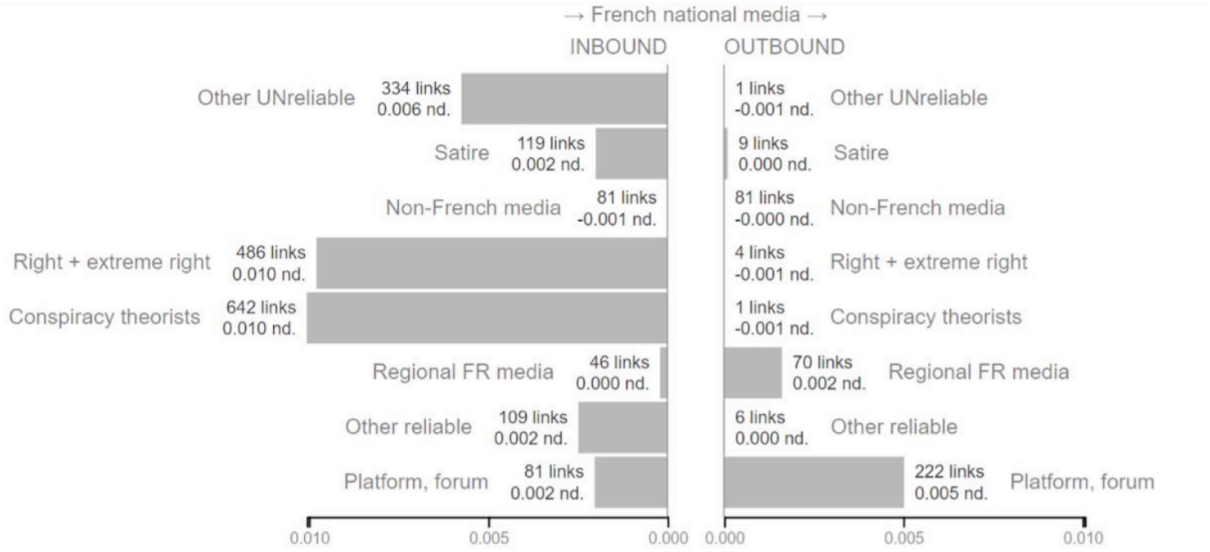

Figure 3: Hyphe's visualisations show hyperlink asymmetries for groups of web entities

has a lot to tell about the local and global structures of our digital public space.

As a scientific instrument, Hyphe suggests (and to some extent imposes) a specific focus on the matters it observes. By choosing to focus on hyperlinks instead of APIs, it encourages to study the web as a hypertext media thanks to a complete methodology including both quantitative and qualitative aspects: data collection, sourcing, iterative corpus building, categorisation through tags or qualitative description, network visualisation and statistical analysis. It still has some limitations (like the technical impossibility yet to mine hyperlinks from JavaScript powered web pages with invisible content to classical crawling technologies) and overcoming them opens new challenges. Building such an ambitious research instrument is indeed a very long journey which we would not have achieved without a long term support from the DIME-SHS EQUIPEX grant.

The World Wide Web was designed as a free and open knowledge space. By providing a free and open source research tool to engage with it as a research field, we aim both at fostering teachers to show students what the WWW is beyond Google or Facebook's interfaces, and at equipping scholars to look beyond the gates of web platforms' APIs, study web actors and analyse their connectivity as well as their hierarchical position. Or, in simpler terms, spreading the word that hyperlink is not dead.

Hyphe has a public demo online: http://hyphe.medialab.sciences-po.fr/demo/

ACKNOWLEDGMENTS

This work was supported by the DIME-SHS research equipment financed by the EQUIPEX program (ANR-10-EQPX-19-01) and the FORCCAST pedagogical equipment financed by the IDEFI program (ANR-11-IDEX-0005-02).

The Digital Methods Initiative in Amsterdam, Dominique Cardon, Jonathan Gray and Tommaso Venturini inspired part of this paper thanks to fruitful discussions regarding hyperlinks.

REFERENCES

[1] 2018. electron: Build cross-platform desktop apps with JavaScript, HTML, and CSS. https://github.com/electron/electron original-date: 2013-04-12T01:47:36Z.
[2] Félix Arias-Robles and José Alberto García-Avilés. 2017. Many Short Links. *Digital Journalism* 5, 9 (Oct. 2017), 1205–1225. https://doi.org/10.1080/21670811.2016.1240014
[3] Judit Bar-Ilan. 2007. Manipulating search engine algorithms: the case of Google. *Journal of Information, Communication and Ethics in Society* 5, 2/3 (2007), 155–166.
[4] A. L. Barabási and R. Albert. 1999. Emergence of scaling in random networks. *Science* 5439, 286 (1999), 509–512.
[5] Thierry Bardini. 1997. Bridging the Gulfs: From Hypertext to Cyberspace[1]. *Journal of Computer-Mediated Communication* 3, 2 (1997), 0–0. https://doi.org/10.1111/j.1083-6101.1997.tb00069.x
[6] Tim Berners-Lee and Robert Cailliau. 1990. *WorldWideWeb: Proposal for a HyperText Project.* Technical Report. https://www.w3.org/Proposal.html
[7] Tim Berners-Lee, Roy T. Fielding, and Larry M. Masinter. 2005. *Uniform Resource Identifier (URI): Generic Syntax.* Number 3986 in Request for Comments. RFC Editor. https://doi.org/10.17487/RFC3986 Published: RFC 3986.
[8] Marie-Aimée Berthelot, Marta Severo, and Eric Kergosien. 2016. Cartographier les acteurs d'un territoire : une approche appliquée au patrimoine industriel textile du Nord-Pas-de-Calais. In *CIST2016 - En quÃête de territoire(s) ?* CollÃÍge international des sciences du territoire (CIST), Grenoble, France, 66–72. https://hal.archives-ouvertes.fr/hal-01353660
[9] Rodrigo A. Botafogo and Ben Shneiderman. 1991. Identifying Aggregates in Hypertext Structures. In *Proceedings of the Third Annual ACM Conference on Hypertext (HYPERTEXT '91).* ACM, New York, NY, USA, 63–74. https://doi.org/10.1145/122974.122981
[10] Soumen Chakrabarti. 2001. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. ACM Press, 211–220. https://doi.org/10.1145/371920.372054
[11] Maxime Crépel, Dominique Boullier, Mathieu Jacomy, Benjamin Ooghe-Tabanou, Diégo Antolinos-Basso, and Paul Monsallier. 2017. Privacy web corpus. http://tools.medialab.sciences-po.fr/privacy/
[12] Caterina Froio. 2017. Nous et les autres. *Réseaux* 202-203 (June 2017), 39–78. https://doi.org/10.3917/res.202.0039
[13] Carolin Gerlitz and Anne Helmond. 2013. The like economy: Social buttons and the data-intensive web, The like economy: Social buttons and the data-intensive web. *New Media & Society* 15, 8 (Dec. 2013), 1348–1365. https://doi.org/10.1177/1461444812472322
[14] Anne Helmond. 2013. The Algorithmization of the Hyperlink : Computational Culture. *Computational Culture, a journal of software studies* (Nov. 2013). http://computationalculture.net/the-algorithmization-of-the-hyperlink/
[15] Anne Helmond. 2015. The Platformization of the Web: Making Web Data Platform Ready. *Social Media + Society* 1, 2 (July 2015), 2056305115603080. https://doi.org/10.1177/2056305115603080

[16] Bernie Hogan. 2018. Digital Traces in Context| Social Media Giveth, Social Media Taketh Away: Facebook, Friendships, and APIs. *International Journal of Communication* 12, 0 (Jan. 2018), 20. http://ijoc.org/index.php/ijoc/article/view/6724

[17] Mathieu Jacomy, Franck Ghitalla, and Dana Diminescu. 2007. Méthodologies d'analyse de corpus en sciences humaines à l'aide du Navicrawler. In *Rapport final du programme TIC & Migrations*. Fondation Maison des Sciences de l'Homme, Paris, France.

[18] Mathieu Jacomy, Paul Girard, Benjamin Ooghe-Tabanou, and Tommaso Venturini. 2016. Hyphe, a Curation-Oriented Approach to Web Crawling for the Social Sciences. AAAI, Cologne, Allemagne. https://spire.sciencespo.fr/hdl:/2441/6obemb2hsj9pboj9bbvc7sftne

[19] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE* 9, 6 (June 2014), e98679. https://doi.org/10.1371/journal.pone.0098679

[20] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. 2016. Jupyter Notebooks: a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides and B. Schmidt (Eds.). IOS Press, 87 – 90.

[21] Ruud Koopmans and Ann Zimmermann. 2007. Visibility and Communication Networks on the Internet: The Role of Search Engines and Hyperlinks. 213–264.

[22] Bruno Latour, Pablo Jensen, Tommaso Venturini, Sébastian Grauwin, and Dominique Boullier. 2012. 'The whole is always smaller than its parts'–a digital test of Gabriel Tardes' monads. *The British journal of sociology* 63, 4 (2012), 590–615.

[23] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology* 59, 7 (May 2008), 1041–1052. https://doi.org/10.1002/asi.20794

[24] Ross A. Malaga. 2008. Worst practices in search engine optimization. *Commun. ACM* 51, 12 (Dec. 2008), 147. https://doi.org/10.1145/1409360.1409388

[25] Sciences Po médialab. [n. d.]. Hyphe: web corpus curation tool & links crawler. https://github.com/medialab/hyphe

[26] Anders Kristian Munk. 2014. Mapping Wind Energy Controversies Online: Introduction to Methods and Datasets. *SSRN Electronic Journal* (2014). https://doi.org/10.2139/ssrn.2595287

[27] Theodor Holm Nelson. 1993. *Literary machines: the report on, and of, Project Xanadu concerning word processing, electronic publishing, hypertext, thinkertoys, tomorrow's intellectual revolution, and certain other topics including knowledge, education and freedom.* Mindful Press, Sausalito (Ca.). OCLC: 272513536.

[28] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report. http://ilpubs.stanford.edu:8090/422/

[29] Guillaume Plique, Mathieu Jacomy, Benjamin Ooghe-Tabanou, and Paul Girard. 2018. It's a Trie... it's a Graph... it's a Traph! https://fosdem.org/2018/schedule/event/multi_level_graph_index/

[30] Bernhard Rieder. 2015. the end of Netvizz (?). http://thepoliticsofsystems.net/2015/01/the-end-of-netvizz/

[31] Richard Rogers. 2013. Mapping public Web space with the Issuecrawler. *Digital cognitive technologies: Epistemology and the knowledge economy* (2013), 89–99.

[32] Richard Rogers. 2017. Digital methods for cross-platform analysis. *The SAGE handbook of social media* (2017), 91–110.

[33] Forccast team. [n. d.]. Formation par la cartographie des controverses à l'analyse des sciences et des techniques. http://controverses.org/

[34] Tommaso Venturini, Mathieu Jacomy, Liliana Bounegru, and Jonathan Gray. 2017. Visual Network Exploration for Data Journalists. In *The Routledge Handbook of Developments in Digital Journalism Studies* (abingdon: routledge ed.). Scott Eldridge II and Bob Franklin, Rochester, NY. https://papers.ssrn.com/abstract=3043912

[35] Esther Weltevrede. 2016. Repurposing digital methods: The research affordances of platforms and engines.