# Methods for mapping hyperlink networks

## Examining the environment of Belgian news websites

---

Juliette De Maeyer
University of Brussels (ULB), Belgium
Research fellow of the F.R.S. – FNRS
Juliette.De.Maeyer@ulb.ac.be

Abstract : This research is an attempt at mapping data retrieved from news websites in order to find one's bearings in the ever-growing complexity of the informational landscape. It posits that drawing maps of the hyperlinks networks in which news websites are entangled will shed new light on the web-based media outlets, by revealing an otherwise concealed dimension of online news. The paper focuses mostly on conceptual foundations and methodological issues, grounded into exploratory attempts of mapping and describing hyperlinks found within selected webpages. It will emphasize a thorough discussion of key concepts and methods exploring: (1) why map? ; (2) why map hyperlinked environments? ; (3) why map the hyperlinked environment of news websites? Questions are raised regarding the nature of links in the context of journalism, and the conception of hyperlink as adding journalistic value to the news website content.

> "*Cyberspace. A consensual hallucination experienced daily by billions of legitimate operators, in every nation, by children being taught mathematical concepts... A graphic representation of data abstracted from banks of every computer in the human system. Unthinkable complexity. Lines of light ranged in the nonspace of the mind, clusters and constellations of data. Like city lights, receding..."*
>
> William Gibson, Neuromancer, 1984

The web is vast. Even when considering controversial or overtly underestimated figures its size makes it impossible to grasp single-handedly. Estimations vary: for example, it is said to count 19,9 billion pages (Wordlwidewebsize.com 2010), or $10^{12}$ hyperlinks (Heymann 2008). In a few words, "the web has exploded before our eyes" (Tremayne 2004, p.234).

Such statement could easily be transposed to the specific domain of online news, as news websites are plentiful. Traditional media organizations not only have companion websites, but new competitors entirely dedicated to online news appear. Other players such as blogs are also marching into the field of news. The readers keen to find their way in that environment, eager to know where these sites are situated in a more global landscape, are inevitably lost.

In this context, the most common way we use to find information on the web, namely search engines, appears to be useless. Search engines index an important part of the sprawling network, but the format chosen to present results is austere: a flat list of URLs of which we seldom consult more than the 20 or 30 first entries (Halavais 2009, p.42). Most importantly, search engines do not even remotely provide a bird's eye view, a global and coherent panorama of the territory we wish to investigate, and idea of the position occupied by every actor and element. Not only is the web largely uncharted, it also lacks geography.

The present paper suggests that maps of hyperlink networks are valuable instruments compensating for this shortage. It also advocates the application of such tools and methods to news websites. The first section delineates the rationale behind the general need for maps. Section 2 consequently focuses on maps of hyperlink networks: previously discovered properties of networks are presented, as well as two examples of existing hyperlink maps. Section 3 addresses what is specifically at stake for news websites when it comes to hyperlink networks. Finally, sections 4 details the methods and tools one can use when attempting to map news sites, while section 5 outlines first results obtained by applying them to the case of Belgian news sites.

## 1. Why map?

In everyday life, maps help us to find our bearings. Applied to the web, they could allow us to grasp a global insight, and to find our way toward specific goals. The intuitive desire to rely on maps is ingrained in the spatial metaphors that we use to talk about the web: we *surf*, *navigate* and *explore* websites that are linked to form a *network*. Common labels to describe the

phenomenon include information *superhighways*, which are part of our informational *landscape*. The words we use to depict the internet have strong spatial and material meanings. As Anne Cauquelin observes, such words are "constantly sliding from one edge to the other of the semantic constellation", making the difference between virtual and physical spaces quite thin (Cauquelin 2007, p.18).

Among that spatial glossary, the *network* might be the most prevailing figure. Webpages are interconnected through hyperlinks, and therefore correspond to the most straightforward definition of a network: "a set of links and relations between points" (Poidevin 1999, p.146). However, it is never obvious or unambiguous to spatially represent the world wide web, whether as a network or in another form.

First and foremost, we have to remember that web pages may have a geographical referent (embodied in their domain names) but that at no point they possess inherent spatial attributes (Dodge & Kitchin 2000, p.72). From then on, if mapping is to be regarded as "a flat, geometrical, simplified and conventional representation of parts or the totality of the surface of the earth, in a ratio of acceptable similarity called the scale" (Joly 1994, p.3) it can only be applied to the most down-to-earth components of the Internet (Dodge & Kitchin 2000, p.72) i.e. its physical infrastructure – the machines, computers and servers that materially compose the network.

Being confronted with data with no spatial properties is an issue map-makers and information visualization specialists are familiar with: "many interesting classes of information have no natural and obvious physical representation. A key research problem is to discover new visual metaphors for representing information and to understand what analytical tasks they support." (Gershon et al. 1998, p.10). Hence, mapping webpages necessarily involves an operation of *spatialization*, i.e. the use of "a number of graphical techniques and visual metaphors" in order to "map data with no inherent spatial properties onto a defined spatial framework so that it might be better understood" (Dodge & Kitchin 2000, p.107).

Consequently, when it comes to map the web, everything is to be invented. There is no single *true* representation of a "reality", but a variety of possible representations depending on what the map-maker wants to show. In other words, we can argue that the quality of a map relies in its adequacy toward needs, and that it is of paramount importance to understand which decisions were made about "what to include and what to exclude, how the map will look and what the map wants to communicate" (Dodge & Kitchin 2000, p.75). Moreover, another crucial question is how to interpret visualizations, or rather how to avoid misinterpreting it. Typically, the way we instinctively read a map – by applying our deeply interiorized habits of 2D or 3D representations of our topographic environment – is not relevant. Namely, some types of visualization imply that the distances or the north/south orientation are meaningless. To put it shortly, maps can be misleading. For instance, some maps choose to show the semantic distance

between the textual content of websites[1]. Others focus on the importance of keywords in the flow of news[2], or the relationships weaved by blogs through reciprocal linking[3]. Therefore, the possibilities of interpretation strongly vary: for some, it is the size of the elements that matters; for others, their relative location, their color, the density of groupings, or a combination of these criteria. In any case, there is no intuitive reading, and every map must come with an explanation about what we might (or might not) read on it, and how to do so.

## 2. Why map hyperlinked networks?

Within the wide range of maps delineated above, my research focuses on those relying on hyperlink networks. Hyperlinks are fascinating objects *per se*, and are regularly mentioned as one of the key features distinguishing online from offline environments. They are said to affect "the overall size and shape of the public sphere" (Turow 2008, p.4). Together, documents and hyperlinks form networks, a concept that has become a leading paradigm for understanding the internet, new media (Gane & Beer 2008), the economy (Kirman 1997) or the society as a whole (Castells 2000). On the web, however, networks are more than a useful metaphor helping us to imagine its informational structure (Ghitalla 2009): recent experimental and empirical works have shown their powerful materiality. Therefore, the prevailing notion of network is questioned as a dominant and fashionable conceptual tool – sometimes marred with ideology (Mattelart 1999; Rumpala 2007; Boltanski & Chiapello 1999) – and its substantial nature is investigated.

### Networks possess remarkable properties

*Network sciences* investigate the properties of networks. After having emerged three centuries ago with the breakthrough in graph theory made by Leonard Euler (Barabasi 2003), the field was expanded by social scientists in the 1960's (Tremayne et al. 2006, p.292). In the 1990's, a set of discoveries that have been labeled "the new science of network" (Watts 2004b; Watts 2004a) revitalized the discipline. Rooted in physics, mathematics, computer science, biology, economics and sociology (Watts 2004b, p.243), this fruitful ground of investigation has given birth to research into patterns of the web, which has "uncovered principles that help to understand networks of all type" (Tremayne 2004, p.234).

"Peculiar and fascinating properties" (Ghitalla 2009) were discovered. For example, researchers highlighted the fact that the web is a "scale-free network, dominated by hubs and nodes with a very large number of links" (Barabasi 2003, p.165), and revealed "strong regularities, among which the existence of a 'universal power law'" (Adamic & Huberman 2001, p.131). This topology goes around with Pareto's law, or the "the so-called 80/20 rule"

[1] e.g. http://professeurs.esiea.fr/wassner/?2008/03/25/126-cartographie-smantique-de-blogs, accessed03/10/2010
[2] e.g. http://newsmap.jp/, accessed 03/10/2010
[3] e.g. http://labs.wikio.net/wikiopole/, accessed 03/10/2010

(Tremayne 2004, p.238): 20% of the nodes possess 80% of the links – a small part of websites collects the majority of links while most receive few or no links. For their part, Albert et al. (1999) established the diameter of the web and its small-world nature: despite the impressive total amount of existing websites, there are, on average, nineteen degrees of separation between two pages (Barabasi 2003, p.165), i.e. one webpage is only nineteen clicks away from another. Next to these studies emphasizing low distance as a characteristic of the web, researchers have shown that finding a path between two pages is not that obvious: Broder et al. (2000) proposed a fragmented model of the web corresponding to a bow-tie structure where the direction of the links strongly matters (Barabasi 2003, p.166).

From those large-scale observations, research then goes on investigating smaller subsets: Sunstein (2007) discerned fragmentation within the political domain – as only 15% of the webpages he scrutinized linked to the opposite viewpoints – and others consequently warned against polarization (Webster 2008, p.32), a trend dubbed *cyberbalkanisation*. Finally, more and more researchers focus on web communities, scrutinizing social ties as well as technical characteristics (Adamic et al. 2003; Adamic 1999; Gibson et al. 1998; Ghitalla et al. 2006; Highfield 2009; Bharat et al. 2001). This paper embraces such approaches: focusing on small-scale networks while keeping in mind the general laws discovered by the network sciences. Although applying all their mathematical and statistical tools to modest corpora does not make sense, their lessons nevertheless constitute an important background – similar to the physical laws of which we must be aware when exploring any *offline* phenomenon.

Handling smaller data sets allows a change in methods, a transition from "industrial automation" to "handiwork" (Ghitalla 2009). Our ambition here is to overcome the frailty of previous link studies, i.e. the gap between quantitative and interpretative perspectives (Fragoso 2009, p.4). Semi-automated methods for mapping hyperlinked environments allow constant iterations between structure (the network of hyperlink and its particular topology) and the content (the webpages themselves). Such an approach is consistent with the assumption that "understanding links requires knowing the context and conditions in which they occur, what implies identifying sites and pages where they are located" (Fragoso 2009, p.8). Following Halavais "the universal nature of hyperlinking makes it a very difficult sort of artifact to understand. The question of what someone means when they create a hyperlink or when they activate one is entirely determined by the context" (2008, p.43). In concrete terms, this translates into producing maps on which every node has been identified and meticulously classified by the researcher, and not a network that was solely crawled in an automated way by a machine.

### Maps of hyperlink networks allow to perceive otherwise unnoticed features

Why is this of any interest for research? Maps are first and foremost tools for analytical reading (Ghitalla 2008). The two examples analyzed here show the fruitful intertwinement of the maps, the contents of the websites, and their broader context.

Wikiopole[4] is a monthly updated map showing the current state of the French blogosphere. It proposes a visualization of the Wikio blog ranking which positions blogs according to the number of links pointing to them (*backlinks*). The 1500 first blogs of the ranking are depicted, as well as their mutual links. Each node is a blog, identified by an URL. The color of the node stands for its manually assigned category (e.g. *politics*, *literature*, *high-tech*, etc.). Their position depends on the hyperlinks only: two blogs will be nearby each other if they are linked. "The links spatially distribute URLs and give birth to the general patterns of the representation (…) The graphical territory is neither governed by an outside orientation system (North/South) nor cross-ruled by fixed lines and geodesic marks" (Ghitalla 2009).

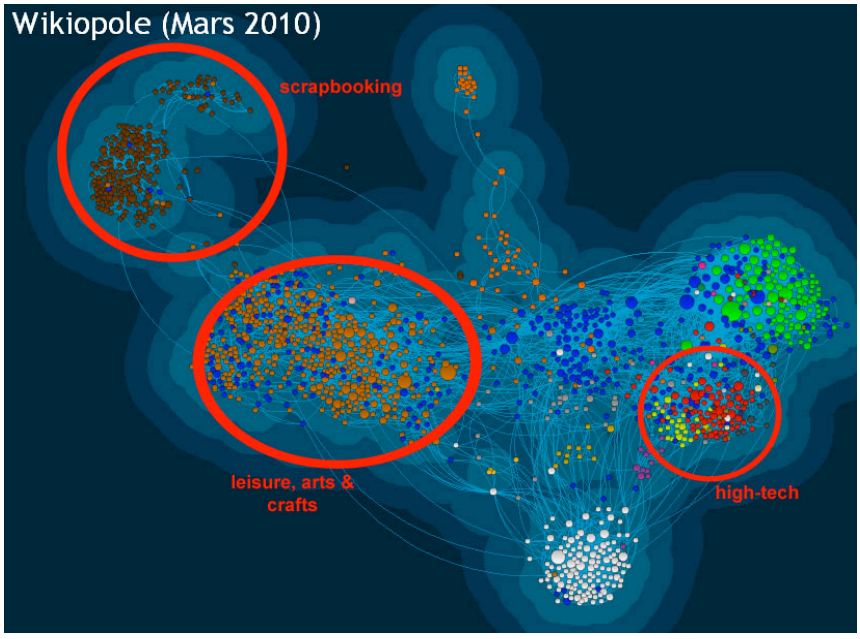


**Figure 1 - Wikiopole**

Figure 1 shows the Wikiopole in March 2010: we can see strong "topical localities" (Davison 2000), i.e. hypertextual proximity corresponding to similarity in content (Ghitalla et al. 2005). Blogs on the same topic (belonging to the same category) are represented next to each other on the map, even though thematic categorization does not determine the position of the nodes. Neat, well-defined islands appear: clusters of blogs, which are heavily interlinked, but loosely connected to other groups. This means, for example, that blogs on literature intensely link to other literary blogs, but sparsely refer to blogs on politics.

Since April 2009, an interesting trend surfaced in Wikiopole, and has been dubbed "the knitters' revolution" (Véronis 2009b). A new island had taken shape and had grown so much that it can be called a continent. It was exclusively formed of blogs about leisure, arts and crafts, and knitting – an entire archipelago was even devoted to scrapbooking. The phenomenon was highly visible on the map and was the sign of a radical change in a blogosphere traditionally dominated by high-tech blogs (Véronis 2009a). Even if other clues could lead to similar

[4] http://labs.wikio.net/wikiopole/, accessed 03/15/2010

observations (such as the top position held by a handful of knitting blogs in the Wikio ranking), the map made it fully visible. When it was published, it made people aware of the unprecedented movement, and triggered noteworthy reactions and analyses: some underlined the gender polarization of the blogosphere (with male bloggers trusting the high-tech island and female bloggers occupying the arts and crafts territory), others tried to explain the phenomenon by the foreseeable seizure of power by a mass of latecomers in a field long subjugated by a minority of early adopters (who might have departed to other territories, such as Twitter) (Véronis 2009a). Some, finally, suggested that the event could be included in the context of media use: the success of arts and crafts blogs could perhaps be explained by the disappearance of periodical press on that topic (Silber 2009).

Next to observations about the evolution of a system at large, maps also allow the emergence of remarkable individual features. For instance, still inside the French blogosphere, it is possible to focus on an individual: *Les coulisses de Bruxelles* (Brussels backstage) is a blog about European politics, held by Jean Quatremer, a journalist for the French daily Libération. This blog appears on the Wikiopole map where it shows all sign of a respectable amount of authority – if we accept the "assumption that hyperlinks somehow transmit power or credibility" (Halavais 2008, p.49) and that the structure of links can be used to impose a structure of relevancy (Finkelstein 2008, p.104).

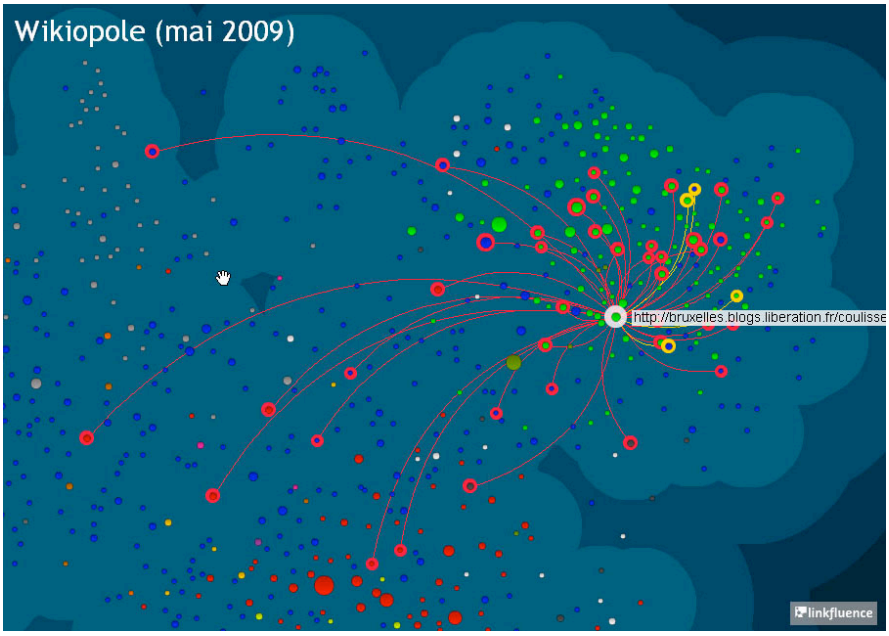


**Figure 2 - Les coulisses de Bruxelles in Wikiopole**

Figure 2 shows *Les coulisses de Bruxelles* and connected blogs. If a decent amount of bloggers link to Quatremer's blog (incoming link in red), it would seem that the journalist-blogger rarely returns the favor: there are only four outgoing (yellow) links from *Les coulisses de Bruxelles* to other blogs. The situation is paradoxical: Jean Quatremer is undeniably *in* the blogosphere, but he does not really *take part* in it.

A glimpse to another map proves useful to further investigate this seeming paradox. The map of the European web[5] is built on the same principle as Wikiopole, but instead of blogs, it depicts 2046 websites devoted to the European Union. Quatremer's blog is also represented on the map, but in this case (Figure 3), the linking pattern differs. Here, the blog is linked to (red links) but there are also some outgoing links (yellow), as well as reciprocal links (green).

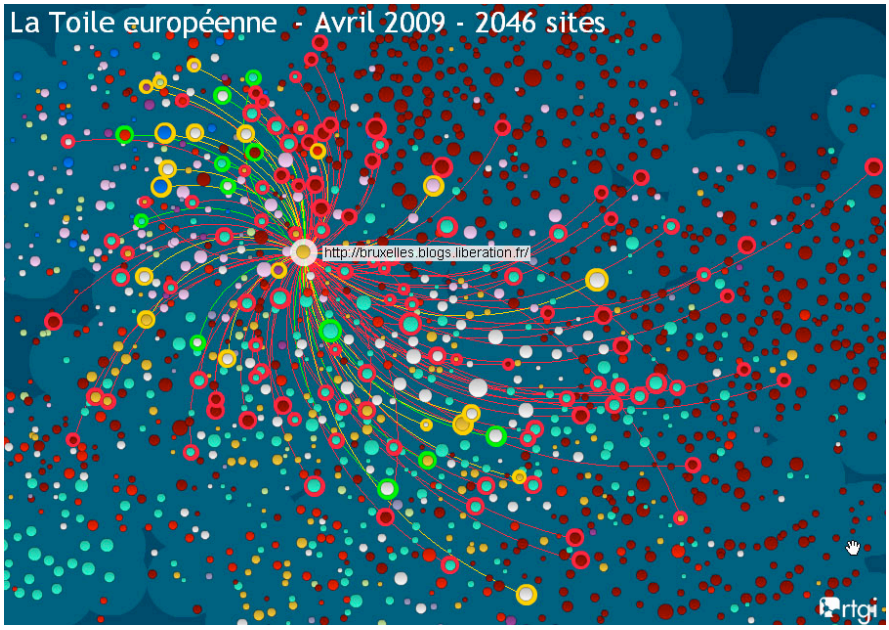


**Figure 3 - Les coulisses de Bruxelles in the European web**

The two maps put side-by-side show that Jean Quatremer does not take actively part in the conversation (as materialized by links) of the top-ranked French blogosphere. However, he seems to belong more dynamically to another sphere, that of the European web. This might question his identity as a journalist-blogger. On the one hand, his linking policy does not match the expected behavior of a blogger, often defined as fostering a conversational system with other bloggers (Domingo & Heinonen 2008, p.6) and relying heavily on hyperlinks, "an important and even essential characteristic" (Tsui 2008, p.72). On the other hand, when linking, he links to other media sources, institutions, or political websites, a practice that resembles the linking behavior of the "traditional" news websites (Tremayne 2005) and the journalistic routines.

From the knitters' revolution to the journalist-blogger duality, hyperlink maps are thus helpful in visualizing phenomena that would have been otherwise hard to detect: "the reason for performing such an analysis is to reveal latent structures that are not already obvious to an observer" (Halavais 2008, p.45). Maps are powerful tools to initiate questionings on identities, at an individual or a more global level: they reflect "deep social and cultural structures" (Halavais 2008, p.39), so that ultimately, "you are what you link" (Adamic & Adar 2001).

[5] http://www.touteleurope.fr/fileadmin/CIEV2/03_actions/construction_euro/toile/cartoavr09/cartographie.html, accessed 03/15/2010

## 3. Why map the hyperlinked environment of online news?

The previous section showed how hyperlink maps contribute to understanding complex phenomena by detecting structural significant trends. The following section will zero in on the particular domain of online news, and argue that hyperlinks are a central issue deserving our full interest.

Why do hyperlinks matter for online news? They are not only able to "create an element of interactivity for the user" (Peng et al. 1999), but also seem to embody some of the promises of online journalism. For instance, two of the problems journalists routinely face are addressed by hyperlinking: "the first is the question of how much of the previous day's events need to be recapped in today's story. On the web, the journalist could link to yesterday's news and dispense with a background paragraph. The second benefit concerns alternative points of view, those that are often excluded in favour of more mainstream views" (Tremayne 2005, p.31).

The added value of links resides thus firstly in the ability to refer to a greater context and to improve or restore transparency as well as credibility (Tremayne 2005, p.32; Tsui 2008, p.70). By stating this, one agrees with the assumption that the process of sourcing is what defines journalism (Tsui 2008, p.71), and that it must be supported by facts in order to be credible. Hyperlinks might "direct readers to additional sources of information" (Dimitrova et al. 2003, p.402) that provide facts and context to back a news story without weighing it down. Concretely, linked documents could be other related articles, or background about the major figures or events mentioned (Franklin 2005, p.105). "The technology of the web allows news presentations that might satisfy both those wanting shorter fact-driven accounts and those wanting context, interpretation and opinion" (Tremayne 2004, p.238). Added value would also lie in the prospect for journalists to echo multiple points of view, embodying the "myth of online journalism", characterized by Domingo as a "program for creating a more transparent, comprehensive and dialogical reporting that would strengthen democratic participation in plural societies" (2008, p.683). An additional advantage of hyperlinking takes the shape of reciprocity, as "important commercial concerns regarding reciprocal linking may guide news web sites" (Tsui 2008, p.75). Some authors also underline the importance of hyperlinks in the process of gatekeeping. A gatekeeper is "an individual who filters out and disregards unwanted/uninteresting and/or unimportant information or stories and attends to information of more import" (Franklin 2005, p.92). By allowing news providers to suggest which voices are worthy of attention (Tsui 2008, p.71), hyperlinking is "perhaps the most significant mechanism of online gatekeeping" (Napoli 2008, p.63). "The decision about which hyperlinks to include in web news stories and which not to include constitutes an additional gatekeeping decision made by web news editors" argue Dimitrova et al. (2003, p.402).

In sum, hyperlinks improve online news by providing context and credibility; by allowing multiperspectival journalism; by initiating links likely to be reciprocated; and by strengthening the process of gatekeeping.

However, despite those appealing promises, news websites seem to have failed, so far, to fully embrace hyperlinks. Empirical research has revealed an overall lack of hypertextuality in online media (Oblak 2005), and when linking was observed, it sparsely led to external sources (Dimitrova et al. 2003). Far from the idea of openness included in the "emergence of a new form of news perhaps best described as contextualized journalism" (Pavlik 2001, p.217), news websites mainly point to internal resources. In that sense, they have been characterized as "gated cybercommunities" (Tremayne 2005) or "walled gardens" (Napoli 2008).

The promise of a journalism that would be improved by hyperlinks nonetheless survives, maybe reinforced by the rapid development of the so-called *web 2.0*. Commentators have been calling for the advent of "link journalism", made easier thanks to new tools such as the microblogging site Twitter (Karp 2008), social bookmarking or collaborative filtering platforms. Similarly, some have argued in favor of "networked journalism". For example, Charlie Beckett is claiming that it creates quality by adding value to the news in three ways: editorial diversity, relevance, as well as connectivity and interactivity (Beckett 2010). Further evidence that hyperlinks are still fashionable for news media can be found in the recent BBC policy shift. The March 2010 document entitled "Putting quality first" specifies a very concrete goal: "turning the site into a window on the web by providing at least one external link on every page and doubling monthly 'click-through' to external websites" (BBC 2010, p.4).

Always oscillating between hope and disenchantment in making journalism better, hyperlinks epitomize what is at stake for online news in digital societies. Hyperlink maps of news websites therefore constitute a valuable tool, if only to facilitate the distinction between wishful thinking and real practices, between fantasies of online news prophets and the materiality of the networks we can observe.

## 4. Methods and tools: getting hands-on

If the interest of mapping the hyperlinked environments of news websites has been established above, the researcher nevertheless faces an important question: how to proceed? For this research, two tools were tested, namely Navicrawler[6] and Gephi[7]. The former is a "tool for exploring the web, that analyses content and structure of pages and hyperlinks (…) it mixes a browser and a crawler" (WebAtlas 2009), while the latter is a graph exploration and manipulation software. Both are open-source projects, freely available. The territory to be

[6] http://www.webatlas.fr/index.php?option=com_content&view=article&id=56&Itemid=65, accessed 03/17/2010
[7] http://gephi.org/, accessed 03/17/2010

explored is the Belgian online news landscape; with a simple, yet compelling, research question: what do Belgian online news outlets link to?

This section will not elaborate on Gephi, the visualization tool, which is designed for complex networks and therefore works flawlessly once fed with a well-conceived corpus. The collection of data and the composition of the corpus with Navicrawler, however, are more complex operations. The process of assembling data is nonetheless not completely disconnected from visualization. Frequent iterations are needed, between the websites likely to be added to the corpus and the first graphical overviews of the network. The process could even be considered as threefold, as it is often useful to look at the data in the form of spreadsheets.

The collection process *per se* was elaborated along these lines: (1) opening the homepage of a news website in Navicrawler; (2) collecting all the hyperlinks on that page; (3) browsing all the internal links found on the page (i.e. going one level deeper into the site); (4) browsing all external links found; (5) while doing this, a first categorization emerges, and tags are attributed to each site; (6) in the process, other news websites were identified, which can be used as new starting points for data collection (snowball sampling). So, the data collection is divided into two essential, concomitant operations: deciding what to include in the corpus and what to exclude from it; and classifying the selected websites. This raises a number of methodological concerns.

The first set of issues is connected to the question of where to stop. In this case, the corpus is bound by pre-defined limits: Belgium-based websites, only homepage and first level pages (linked on the homepage). But other criteria are more problematic, namely that of deciding whether a particular links leads to a website likely to be considered as significant in the corpus – as opposed to non significant links such as those stemming from advertisement banners, which are not picked by journalists or even by webmasters and therefore reflect a different linking policy. The case of news websites is more ambiguous than other mapping attempts (e.g. Ghitalla, Jacomy, et Pfaender 2006; Ghitalla, Le Berre, et Renault 2005; Highfield 2009) because the field is not thematically focused. When mapping a thematically coherent domain, such as literature or political blogging, the researcher must decide if the websites he encounters belong to the domain or not. The main difficulty resides in the fact that other nearby domains might be close and sometimes intertwined with the territory being mapped. News websites are thematically open, and the decision to include or exclude a website from the corpus must be based on other criteria. As the research question leading this exploration is open (what do the news websites link to?) and aims at taking stock of the situation, most of the linked websites were included, except the broken or unattainable links (e.g. those leading to pages requiring a login and a password). For instance, even the sites belonging to the "upper layer of the web" were kept, whereas it is commonly recommended to exclude them (Jacomy & Ghitalla 2007, p.6). The upper layer of the web is composed of very generic sites such as search engines or portals. They represent dead-ends for thematic explorations, because they are nonspecific and offer too many links, which lead to dispersion. Yet in this case it seemed relevant to keep

them in the corpus, if only to examine the potential domination of those web giants in the Belgian online news environment.

The second main operation, sorting websites, also requires methodological questioning. The flexible way of tagging and categorizing on Navicrawler encourages emergent classification. Rather than pre-establishing categories, the research started with loose ones, which have become more refined along the exploration. A view of the corpus as a table allows to keep a global overview of the work-in-progress classification, and temporary visualizations of the graph open onto the detection of phenomena needing more sophisticated tagging. The absence of thematic coherence also implies impediments, as news websites can virtually link to anything. Categories do therefore not rely on the content of the linked sites, but rather on more general features likely to cross the whole corpus, such as genres.

The main issue here is the stability of the corpus, with news websites being especially sprawling and constantly evolving. The method described below allows seizing a snapshot of the news sites environment, which could lead to more exhaustive or time-sensitive investigations.

## 5. Observations and preliminary results: the hyperlinked environment of Belgian news websites

Thirteen Belgian news websites were explored following the method delineated in the previous section (they will be referred as "source sites" below). The resulting network is composed of 548 nodes (i.e. sites) and 1848 edges (i.e. hyperlinks), with a diameter of 9 (i.e. the longest distance between any two nodes in the network). The source sites were not picked in advance as representative of the Belgian news landscape, but discovered during exploration, starting from the leading French-speaking news site (lesoir.be). As a result, thirteen source sites were included and further investigated. They offer a nearly exhaustive coverage of the Belgian generalist news websites, with the noticeable exceptions of deredactie.be (news site of the Dutch-speaking public broadcaster) and rtlinfo.be (news site of the private French-speaking broadcaster), which are absent because no other website linked to them at the time of our analysis.

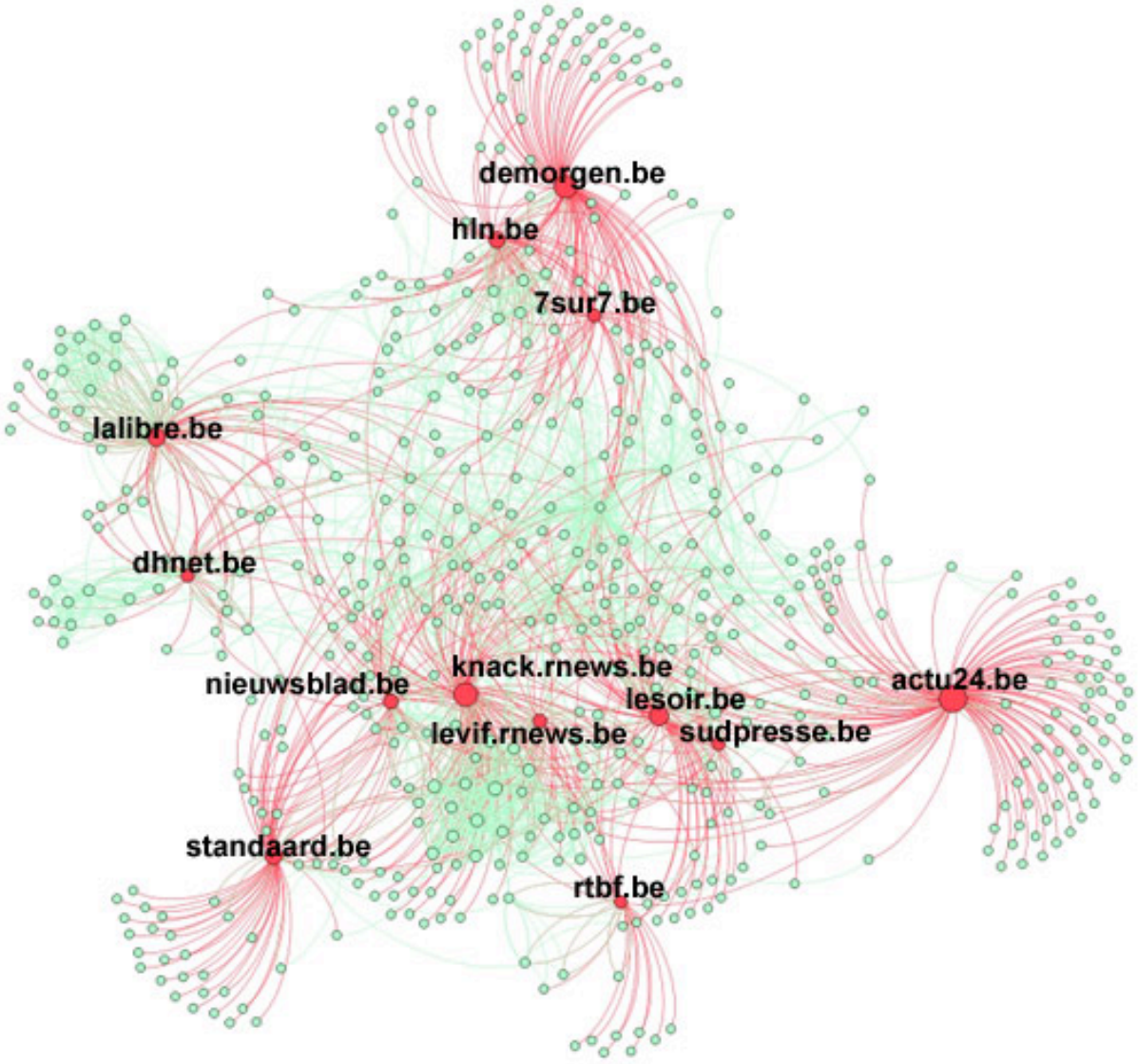


**Figure 4 - the whole network**

Figure 4 shows an overview of the network, with the source sites in red and the other sites in light green. The size of the nodes depends on the number of outgoing links: the more it links to other sites, the bigger a node is. Some links are grouped on the outer circle of the map, whereas others are central. Islands on the edges are sites only linked by one of the source sites, which means that the source sites both share a part of their links and have links specific to them.

When examining the link structure on a global scale we notice that it complies with the general laws of web-based networks. For instance, the number of links (ingoing as well as outgoing) abides by a power law (Figure5): a handful of sites links heavily or are heavily linked, while the majority – the long tail – only features a few links.

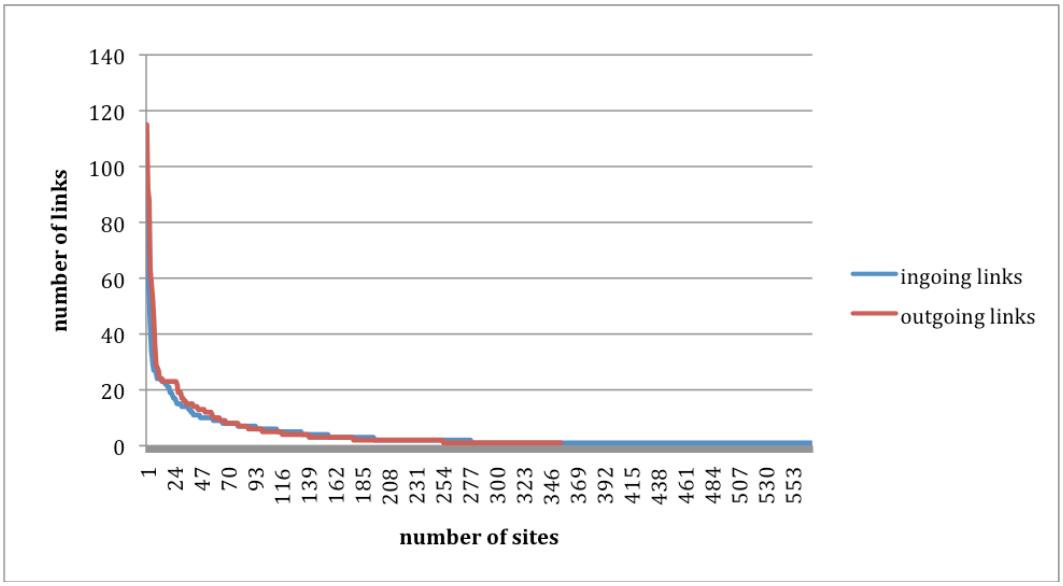


**Figure 5 - total amount of ingoing & outgoing links**

Zooms on subgraphs are needed to better understand the essence of the network, as the global view is intricate and difficult to read. For example, an important portion of links leads to media websites, such as international or local news sites, or the corporate showcase of traditional media companies. The subgraph containing those sites and the source sites is composed of 84 nodes and 234 edges. When adding pure-players, i.e. news sites not affiliated to a traditional media organization, and other sites directly related to professional media – such as journalists union, media regulation organisms or portals promoting the press – the subgraph comprises 117 nodes and 329 edges (Figure 6 where some URLs are displayed as examples). In this case, the size of the nodes depends on the ingoing links (the more a site is linked to, the bigger the node), and can be considered as a rough estimation of the sites' authority in that particular subgraph. The biggest node on this graph is the CIM the Belgian body responsible for measuring media audience and circulation. Its dominant position can easily be explained by the affiliation of all major websites to the CIM for the measurement of the traffic on their pages.

This abundance of links to other media websites go against the assumption that "a standard argument against providing external links is that readers may never come back to your site" and that "the unwillingness to give up control of the visitor's news experience explains the lack of hyperlinks to outside websites" (Dimitrova et al. 2003, p.409), all the more so as linking to friend or partner media, e.g. headed by the same news organization, does only explain a small part of the links.

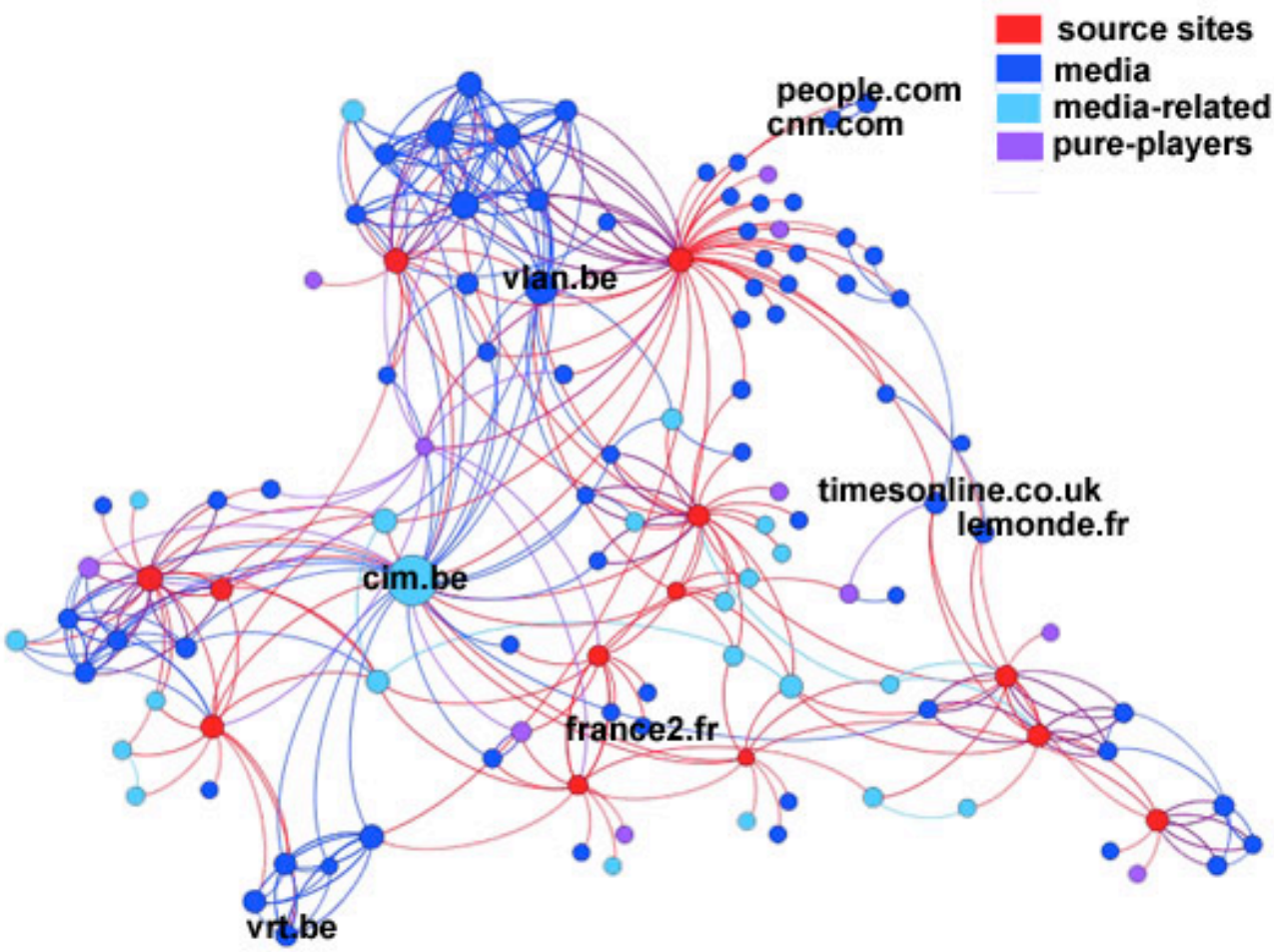


**Figure 6 - media subgraph**

Another noteworthy set of links consists of resources specific to, and centered on, the web. These are search engines, social networking sites, file-sharing platforms, open-source portals, RSS feed readers, aggregators, etc. Some of those links are present to offer useful functions to the reader (search or personalization), while others aim at encouraging visitors to share with their social network and give visibility to the content offered. The subgraph (Figure 7) counts 66 nodes and 187 edges. The dominant sites are giants of the web industry, such as Facebook, Google, Youtube or Twitter – at the expense of competitors such as Yahoo (3 links) or Netlog (1 link).

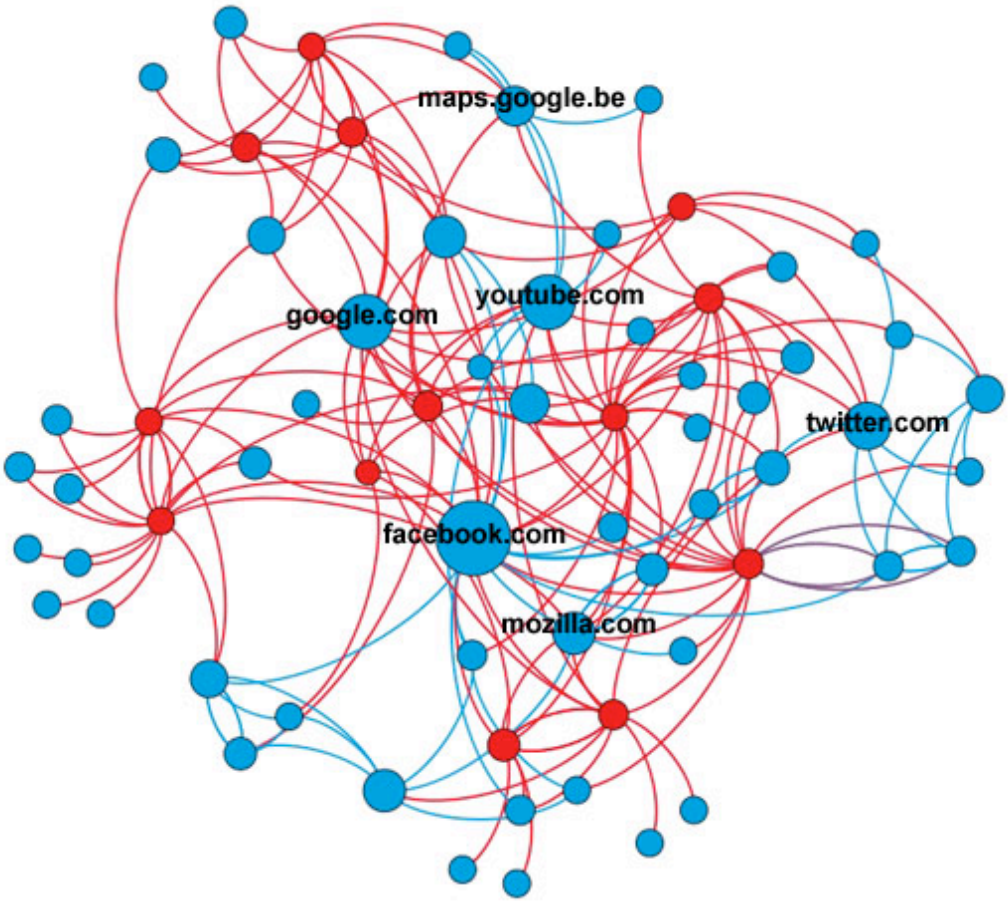


**Figure 7 – web-related resources**

An additional group of sites prompts curiosity: links that do not propose additional informational content, but redirect the user towards sites of practical utility. Concretely, these are links to e-commerce platforms, to general service-providing sites (weather forecasts, job ads, train schedules, e-booking, etc.), to sites related to leisure (travel agencies, amusement parks, tourism-related resources), or to the showcase sites for brands or cultural institutions. These links, however, do not stem from advertisements: they are part of the sites' content. For example, links to the official websites of Xbox and Playstation are found in the column "useful links" of an article on video games. Similarly, links pointing to the sites of Mobistar, Proximus and Base (mobile phone companies) originate from a page explaining how to use the mobile version of the news site. This subgraph (Figure 9) counts 169 nodes and 280 edges.

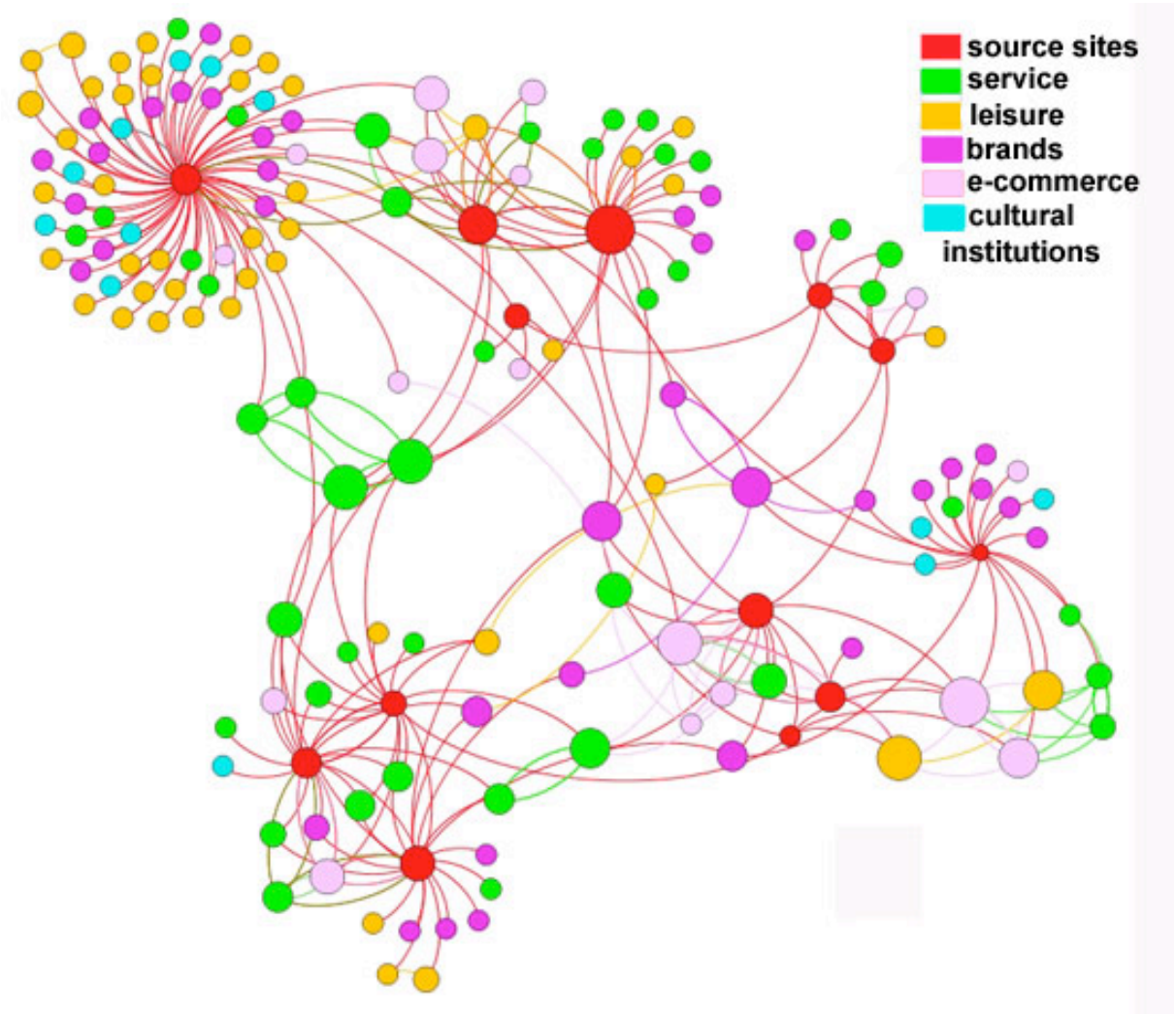


**Figure 8 – links to sites not providing additional journalistic content**

Along with the previously highlighted web-related resources, those sites trigger questions on the function of hyperlinks. For example, questions are raised regarding the use of product placement or the way media discusses brands. But above all, the nature of those links is challenging: they do not provide context or depth to news stories. They do not correspond, for instance, to the guidance on hyperlinks provided by BBC editor Steve Hermann (2010), for whom three sorts of links matter: source material for government reports or science papers, other related news coverage, and related commentary and articles. They do not lead to "browsing through reports, archives dating back years and years, official documents and full transcripts of interviews and statements", nor allow "the reader to trace back the reporting and news gathering process" (Deuze 1999, p.383). The links discussed here are different in nature, they are mainly practical and might come handy for the community of readers – emphasizing the role of media as community managers – but they do not seem to provide strictly *journalistic* content.

Finally, some of the source sites host blogs. On the global map, they have been considered as external sites – they often benefit from a specific URL – but it must be underlined that their status is hybrid. They do not belong to the media organization, yet they are somehow

endorsed by it – with some blogs being written by journalists and others by unknown internet users. For a news website, blogs may represent an opportunity to getting more closely connected to the rest of the web. The blog genre is deeply entrenched in a culture of links, and each blog is generally in liaison with other sites. For instance, Figure 10 shows the 104 blogs hosted by the French-speaking daily's website lalibre.be (in red), and the whole new universe of links they propose (830 nodes and 3551 edges in that graph). Therefore, sheltering blogs might represent a way for news websites to integrate into the link ecology of the web while keeping their distance, as the responsibility of linking is somewhat outsourced on blogs. The whole new possibilities of navigation are at the same time remote and close to the news websites' homepage – they are only one or two click away from it, but do not *exactly* belong to the content of the site.

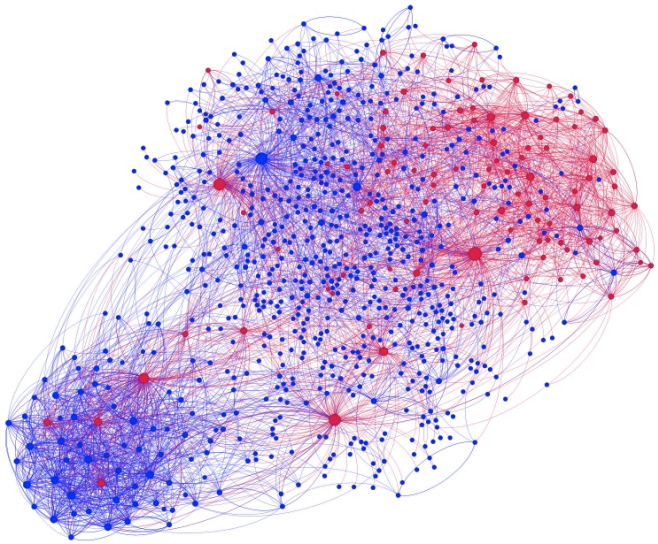

**Figure 9 - network around blogs hosted by lalibre.be**

## 6. Discussion, limitations and concluding remarks

Thanks to hyperlink maps, it is possible to highlight categories of links that challenge common conception about the role of hyperlinks in journalism. First, links to media and media-related sites have shown news websites as centered on their profession, and unexpectedly open to their competitors. Second, links to web-related services and practical information questioned the function of hyperlinks as added journalistic value. Altogether, those links roughly represent half of the global network. The other half still needs to be scrutinized and explained. Further research has to focus on refining the reflection on the nature of links as providing journalistic content or as carrying another function.

On a more general level, the hyperlink map has proven useful to stir questioning and to draw attention to interesting phenomena. However, limitations do exist and network analysts

must be cautious. Firstly, hyperlink maps are not easy to decipher. Our intuitive urge to interpret elements such as orientation or distance must be restricted, because they don't mean anything in this context. When handling complex networks, an important pedagogical effort must be made in order to explain the map. Secondly, we should remain wary of the aesthetic appeal of the maps. They are indeed pleasant to look at, they look rigorous and sophisticated (or "sciencey"), but we must not use them when they do not add value to our analysis or illustrate it clearly. Thirdly, maps will never be self-sufficient: they have to be associated with other ways of presenting data (charts, tables, etc.) as well as with other ways of conducting research (for example, our investigation of Belgian news sites link policy could lead to ethnographic inquiries in the concerned newsrooms in order to confront the results with the journalists' practices). Finally, maps should not be seen as an objective picture of an unambiguously depicted situation. We must not forget that they are largely determined by the arbitrary construction of a research corpus. Similarly, visualizations are not neutral: zooming on a subgraph or highlighting a specific property is a deliberate choice that needs to be clearly stated.

Mapping hyperlink environments therefore constitutes a challenge, drenched in the enthusiasm stirred by new and not well-tried experimental tools. Methods for mapping still need to mature, but they nevertheless represent thought-provoking ways of exploring a set of objects whose complexity, immateriality and pace of change require powerful tools and innovative methods.

---

**References**


Adamic, L.A., 1999. The Small World Web. In *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*. Springer-Verlag, p. 443-452. Available at: http://portal.acm.org/citation.cfm?id=699477 [Accessed March 10, 2010].

Adamic, L.A. & Adar, E., 2001. You are What You Link. Dans 10th international World Wide Web Conference. Hong Kong. Available at: http://www10.org/program/society/yawyl/YouAreWhatYouLink.htm [Accessed October 20, 2009].

Adamic, L.A., Buyukkokten, O. & Adar, E., 2003. A social network caught in the web. *First Monday*, 8(6). Available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1057/977 [Accessed February 23, 2010].

Adamic, L.A. & Huberman, B.A., 2001. The Web's hidden order. *Commun. ACM*, 44(9), 55-60.

Albert, R., Jeong, H. & Barabasi, A., 1999. Internet: Diameter of the World-Wide Web. *Nature*, 401(6749), 130-131.

Barabasi, A., 2003. *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life,* Cambridge MA, Perseus Publishing.

BBC, 2010. Strategy review - Putting quality first. Available at: http://www.bbc.co.uk/bbctrust/our_work/strategy_review/index.shtml [Accessed March 16, 2010].

Beckett, C., 2010. Editorial Diversity: Quality Networked Journalism. Available at: http://www.charliebeckett.org/?p=2575 [Accessed March 15, 2010].

Bharat, K. et al., 2001. Who Links to Whom: Mining Linkage between Web Sites. Dans *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM)*. ICDM. p. 51-58.

Boltanski, L. & Chiapello, E., 1999. *Le Nouvel Esprit Du Capitalisme*, Paris: Gallimard.

Broder, A. et al., 2000. Graph structure in the Web. *Comput. Netw.*, 33(1-6), 309-320.

Castells, M., 2000. *The Rise of the Network Society* 2 éd., Oxford: Blackwell.

Cauquelin, A., 2007. *Le site et le paysage* 2 éd., Paris: Presses universitaires de France.

Davison, B.D., 2000. Topical locality in the Web. Dans *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. Athens, Greece: ACM, p. 272-279.

Deuze, M., 1999. Journalism and the Web: An Analysis of Skills and Standards in an Online Environment. *International Communication Gazette*, 61(5), 373-390.

Dimitrova, D.V. et al., 2003. Hyperlinking as Gatekeeping: online newspaper coverage of the execution of an American terrorist. *Journalism Studies*, 4(3), 401.

Dodge, M. & Kitchin, R., 2000. *Mapping cyberspace*, New York: Routledge.

Domingo, D., 2008. Interactivity in the daily routines of online newsrooms: dealing with an uncomfortable myth. *Journal of Computer-Mediated Communication*, 13(3), 680-704.

Domingo, D. & Heinonen, A., 2008. Weblogs and journalism. A typology to explore the blurring boundaries. *Nordicom Review*, 29(1), 3-15.

Finkelstein, S., 2008. Google, links and popularity versus authority. In J. Turow & L. Tsui (ed). *The hyperlinked society*. Ann Arbor, p. 104-120.

Fragoso, S., 2009. Making Sense of Website Connectivity: a theoretical-methodological proposal for the study of networks of websites and the links that bind them. *Association of Internet Researchers - IR10*. Milwaukee.

Franklin, B., 2005. *Key Concepts in Journalism Studies*, London ; Thousand Oaks: SAGE Publications.

Gane, N. & Beer, D., 2008. *New Media: The Key Concepts* English ed., Oxford ; New York: Berg.

Gershon, N., Eick, S. & Card, S., 1998. Information Visualization. *Interactions*, March-April 1998, 9-15.

Ghitalla, F., 2009. Du nuage aux abymes. Dimensions heuristique et expérimentale des modèles du web. Available at: http://www.webatlas.fr/download/DuNuageAuxAbymes.pdf?aa3b70196c0dc6fb4c4810f9d1c623ff=3e3ddf360a9c31c20dcbe3a3f678d2c6 [Accessed December 22, 2009].

Ghitalla, F., 2008. L'atelier de cartographie. Pratique et enjeux des cartographies thématiques de documents web. Available at: http://www.webatlas.fr/download/docs/ateliercartographie.pdf?aa3b70196c0dc6fb4c4810f9d1c623ff=3e3ddf360a9c31c20dcbe3a3f678d2c6 [Accessed December 22, 2009].

Ghitalla, F., Jacomy, M. & Pfaender, F., 2006. Détection et visualisation d'agrégats de documents web L'exemple du domaine thématique de la Culture Scientifique, Technique et Industrielle. Available at: http://www.webatlas.fr/download/docs/agregatCSTI.pdf?aa3b70196c0dc6fb4c4810f9d1c623ff=3e3ddf360a9c31c20dcbe3a3f678d2c6 [Accessed December 22, 2009].

Ghitalla, F., Le Berre, A. & Renault, M., 2005. Des documents, des liens et des acteurs. Expérimentations autour de radiographies documentaires du web. *Conference H2PTM.*

Gibson, D., Kleinberg, J. & Raghavan, P., 1998. Inferring Web communities from link topology. Dans *Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space---structure in hypermedia systems: links, objects, time and space---structure in hypermedia systems*. Pittsburgh, Pennsylvania, United States: ACM, p. 225-234.

Halavais, A., 2008. The hyperlink as organizing principle. In J. Turow & L. Tsui (ed). *The hyperlinked society*. Ann Arbor, p. 39-55.

Halavais, A., 2009. *Search Engine Society*, Cambridge: Polity.

Hermann, S., 2010. BBC News linking policy. *BBC - The Editors blog*. Available at: http://www.bbc.co.uk/blogs/theeditors/2010/03/bbc_news_linking_policy.html [Accessed March 20, 2010].

Heymann, S., 2008. Du Web à l'idée du Web : conception d'outils pour les sciences humaines. *web-mining.fr*. Available at: http://web-mining.fr/science/du-web-%C3%A0-lid%C3%A9e-du-web-conception-doutils-pour-les-sciences-humaines [Accessed February 7, 2010].

Highfield, T., 2009. Linking to the concerted? Mapping the structure of the French and Australian political blogospheres. *Association of Internet researchers - IR10 Doctoral Colloquium*. Milwaukee.

Jacomy, M. & Ghitalla, F., 2007. Méthodologies d'analyse de corpus en sciences humaines à l'aide du Navicrawler (Rapport final) D. Diminescu, éd. Available at: http://www.webatlas.fr/download/methodo_shs_navicrawler.pdf?aa3b70196c0dc6fb4c4810f9d1c623ff=90d626e304d42bbf1fdba8397c44a569 [Accessed March 7, 2010].

Joly, F., 1994. *La cartographie* 2 éd., Paris: Presses universitaires de France.

Karp, S., 2008. Newsrooms Can Grow Twitter Followers By Using Twitter For Link Journalism. *Publishing 2.0*. Available at: http://publishing2.com/2008/10/29/newsrooms-can-grow-twitter-followers-by-using-twitter-for-link-journalism/ [Accessed March 9, 2010].

Kirman, A., 1997. The economy as an evolving network. *Journal of Evolutionary Economics*, 7(4), 339-353.

Mattelart, A., 1999. Mapping modernity: Utopia and communication networks. in *Mappings*. London: Cosgrove Denis, p. 167-192.

Napoli, P.N., 2008. Hyperlinking and the forces of "massification". In J. Turow & L. Tsui (ed) *The hyperlinked society*. Ann Arbor, p. 56-69.

Oblak, T., 2005. The Lack of Interactivity and Hypertextuality in Online Media. *Gazette*, 67(1), 87-106.

Pavlik, J.V., 2001. *Journalism and New Media*, Columbia, NY: Columbia University Press.

Peng, F.Y., Tham, N.I. & Xiaoming, H., 1999. Trends in Online Newspapers: a look at the U.S. web. *Newspaper Research Journal*, 20(2), 52-64.

Poidevin, D., 1999. *La carte, moyen d'action : conception-réalisation* Ellipses., Paris.

Rumpala, Y., 2007. La connaissance et la praxis des réseaux comme projet politique. *Raison publique*, 7, 199-220.

Silber, M., 2009. Quand les blogs remplacent la presse disparue. *marsupilamima*. Available at: http://marsupilamima.blogspot.com/2009/04/quand-les-blogs-remplacent-la-presse.html [Accessed March 16, 2010].

Sunstein, C.R., 2007. *Republic.com 2.0*, Princeton: Princeton University Press.

Tremayne, M., 2005. News Websites as Gated Cybercommunities. *Convergence*, 11(3), 28-39.

Tremayne, M., 2004. The web of context : applying network theory to the use of hyperlinks in journalism on the web. *Journalism and Mass Communication Quarterly*, 81(2), 237.

Tremayne, M. et al., 2006. Issue Publics on the Web: Applying Network Theory to the War Blogosphere. *Journal of Computer-Mediated Communication*, 12(1), 290-310.

Tsui, L., 2008. The hyperlink in newspapers and blogs. In J. Turow & L. Tsui (ed) *The hyperlinked society*. Ann Arbor, p. 70-83.

Turow, J., 2008. Introduction: On not taking the hyperlink for granted. In J. Turow & L. Tsui (ed) *The hyperlinked society*. Ann Arbor ,p. 1-17.

Véronis, J., 2009a. Blogs: Déclin de la High-Tech ? *Technologies du langage*. Available at: http://blog.veronis.fr/2009/11/blogs-declin-de-la-high-tech.html [Accessed March 16, 2010].

Véronis, J., 2009b. Lexique: La révolution des tricoteuses. *Technologies du langage*. Available at: http://blog.veronis.fr/2009/04/lexique-la-revolution-des-tricoteuses.html [Accessed March 16, 2010].

Watts, D.J., 2004a. *Six Degrees: The New Science of Networks*, Vintage.

Watts, D.J., 2004b. The "New" Science of Networks. *Annual Review of Sociology*, 30(1), 243-270.

WebAtlas, 2009. Navicrawler. Available at: http://www.webatlas.fr/index.php?option=com_content&view=article&id=56&Itemid=65 [Accessed March 31, 2010].

Webster, J.G., 2008. Structuring a marketplace of attention. In J. Turow & L. Tsui (ed) *The hyperlinked society*. Ann Arbor, p. 23-38.

Wordlwidewebsize.com, 2010. The size of the indexed world wide web. Available at: http://www.worldwidewebsize.com/ [Accessed March 26, 2010].