Mapping the hyperlinked environment of online news: issues and challenges for the French news sites

Juliette De Maeyer
Université Libre de Bruxelles (ULB), Belgium
Research fellow of the F.R.S. – FNRS
Juliette.De.Maeyer@ulb.ac.be

ABSTRACT: This paper investigates links in the context of news websites. Although hyperlinking is associated with some of the most salient promises coupled with online news (interactivity, credibility, transparency and diversity), research hardly looks into an elementary yet crucial question addressed by our research: where do links found on news websites lead? Our analysis is based on exploratory maps of hyperlink networks. The very purpose of such maps and a number of specific methodological issues are also examined. On empirical grounds, the links found on 18 French news websites are explored. Three subsets are identified, namely links to other media, links to blogs and links that do not provide additional journalistic content. The function of links is discussed while it is established that links with a journalistic purpose are scarce as compared to links with a commercial or practical aim. The possibility of outsourcing the responsibility of linking on blogs is also discussed.

What makes the difference between online news and traditional media? When we look into the specificity of online journalism, hyperlinks occupy a privileged position. This research discusses the issue through the exploration of links around French news sites. The paper is articulated as follows: section (1) delineates what is a stake with links and presents the myths of online journalism embodied in four promises. Section (2) introduces the methods and tools suited to investigate the topic. Section (3) highlights first results of exploratory attempts at mapping French news sites, discussed in section (4), that also focuses on limits and further research.

1) Why do hyperlink matter for online news?

Hyperlinks, those “highlighted words on a webpage that take [readers] to other places on the web” (Turow, 2008, p. 3) are a key issue for journalism as they appear to embody several promises associated with online news. Those promises, when put
together, constitute online journalism myths, which might be understood as a “socially constructed discourse that was shaped under the historical context of the social role of journalism” (Domingo, 2008, p. 683). Four promises related to hyperlinks shape that discourse, namely: interactivity, credibility, transparency and diversity. Following Domingo, they form a “program for creating a more transparent, comprehensive and dialogical reporting that would strengthen democratic participation in plural societies” (Domingo, 2008, p. 683).

Firstly, hyperlinks “create an element of interactivity for the user” (Peng, Tham, & Xiaoming, 1999) by offering them to click and surf to other parts of the web. They are said to “disperse the fundamental linearity” generally associated with traditional media, and especially with print (Dalhgren 1996; quoted in Oblak 2005). Interactivity, going hand in hand with non-linearity, is considered to be “at the core of new media technologies” (Bucy, 2004, p. 373) and therefore to be valuable for news websites (Chung & Nah, 2009) – even though the positive effect of interactive media on readers seems questionable (Tremayne, 2008).

Secondly, hyperlinks are said to improve the credibility of news stories, by providing links to context, facts and sources. They lead to “browsing through reports, archives dating back years and years, official documents and full transcript of interviews and statements” (Deuze, 1999, p. 383) and can thus help the reader to understand an issue in depth (Dimitrova, Connolly-Ahern, Williams, Kaid, & Reid, 2003). In short, they support facticity, which is desirable if one assumes that the process of sourcing is what defines journalism (Tsui, 2008, p. 71). In so doing, hyperlinking also addresses one of the “problems journalists routinely face” (Tremayne, 2005, p. 31), namely how much it is needed to recap previous events. This issue is settled without making news stories heavy with additional – and perhaps optional – text: “the technology of the web allows news presentation that might satisfy both those wanting short, fact-driven accounts and those wanting context, interpretation and opinion” (Tremayne, 2004, p. 238).

By providing direct access to sources, hyperlinks increase transparency. It is the third way in which they add value to journalism, by allowing “the reader to trace back the reporting and news gathering process” (Deuze, 1999, p. 383). Newsgathering practices such as finding and selecting valuable sources are no longer invisible. On the contrary, hyperlinks are likely to make them fully perceptible by directly pointing to
them. Links thus constitute “the most significant mechanism of online gatekeeping” (Napoli, 2008, p. 63). Matheson argues that when journalists pick noteworthy links and present them to readers, there is a shift in the way they act as gatekeepers: “the communication interaction invoked here involves the offering of material to the user by the embodied voice of the news producer, rather that the existence of unembodied, self-evident information” (2004, p. 455). Journalistic processes such as sourcing or gatekeeping are made explicit.

Finally, the last myth embodied in hyperlinks is that of an enhanced democracy thanks to diversity. Here, links address another problem journalists routinely face: do they have to report every alternative point of view (Tremayne, 2005, p. 31)? By linking to commentary or opinion that would not have found a place in traditional news stories, this “multiperspectival journalism” (Tremayne, 2005) increases diversity and plurality. Those values are considered as central to our democratic societies.

Beside those four aspects in which links are said to add value to journalism, another pragmatic advantage of hyperlinking is to be remarked: placing links might result in reciprocity. Hence, it raises “important commercial concerns” (Tsui, 2008, p. 75) as the economy of the web is increasingly shaped by the way links are structured (Halavais, 2009; Tsui, 2008, p. 74). For instance, search engines such as Google – that generate traffic and profit – are built on “the assumption that hyperlinks somehow transmit power or credibility” (Halavais, 2008, p. 43). Gathering and trading links is an essential part of the web economy.

In spite of those appealing prophecies, news websites seem, so far, to have failed to embrace hyperlinking. They appear to be especially cautious when it comes to providing external links (i.e. links leading to other websites, as opposed to internal links for navigational purposes within a same site). For example, Dimitrova et al. (2003) examined links stemming from 473 news stories gathered on 15 U.S. websites: only 4,1% of them were external links. Likewise, Tremayne (2005) studied 1147 news stories originating from 10 U.S. sites and concluded that only 17% of the links are external, with web editors favoring “local server content to that which appears elsewhere” (Tremayne, 2005, p. 38). Moreover, he observed that the proportion of external links tended to decrease over time, a trend he proposes to explain by the proportional growth of each site’s archives – providing more internal material to link to. More recently, Tsui (2008)
compared the content of four leading newspaper sites with five leading political blogs. In this case, the percentage of external links goes as high as 35.2%, but it is mainly thanks to blogs. When isolated, three out of the four newspaper sites provide less than 3% of external links. The author reaffirms the lack of hyperlinks to track down original material, and chiefly explains the situation by fear of losing control over audiences, by technical and organizational inertia, and by fear of losing ad revenue.

Even though empirical research indicates that news websites do not live up to the expectations, pundits are still advocating hyperlinks. Linking has not only remained fashionable, it also seems to have been reinvigorated by the advent of the so-called web 2.0. Commentators are pleading in favor of a “link journalism” (Karp, 2008), the progress of which would be smoothed by micro-blogging tools, social bookmarking or collaborative filtering platforms. Similarly, some are supporting “networked journalism” as an engaging future for the profession. For example, Beckett (2010) argues that networked journalism relying on hyperlinks would add value to news in three ways: editorial diversity, relevance, as well as connectivity and interactivity. In this respect, a recent BBC policy shift epitomizes the tendency for news organization to firmly believe in links. Entitled Putting quality first, the 2010 report is aimed at defining BBC's policy for the next years, and present "clear priorities", among which: “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 sites” (BBC, 2010, p. 4).

As the myths encapsulated in hyperlinks remain lively, it seems more necessary than ever to investigate them on empirical ground. Besides, a fresh scientific field has shed a new light on what is at stake with hyperlinks: the “new science of networks”. This interdisciplinary mix of physics, mathematics, computer science, biology, economics and sociology (Watts, 2004) has witnessed important breakthroughs since the mid-nineties (e.g. Albert et al. 1999; Adamic & Huberman 2001; Broder et al. 2000; Barabasi 2003). By studying all kinds of networks – among which the web holds a central position – they have highlighted “peculiar and fascinating properties” (Ghitalla, 2009) that instilled a new research dynamic. For instance, typical distribution and topology of web-based networks are now known. They constitute the physical laws of cyberspace, of which we must be aware when we observe online phenomena – just as we must be aware of gravity as a basic principle when studying offline phenomena.
On those bases, this research examines the hyperlinks proposed by French news websites, and aims at mapping the hyperlink networks in which they are enmeshed. In so doing, the functions and roles of links for current online journalism are questioned, echoing the issues delineated above. The following question is addressed: what do the French news sites link to?

2) Methods and Tools

Mapping the hyperlink networks surrounding news sites implies two operations, namely: collecting the data (sites and links) and visualizing them. Those two steps come with their own methodological issues, and require specific tools.

Data collection: building a corpus

The tool used to harvest the data is Navicrawler\(^1\), an open-source project presented by its developers as “a tool for exploring the web, that analyses content and structure of pages and hyperlinks (...) it mixes a browser and a crawler” (WebAtlas, 2009). When it comes to methods, building a corpus of sites and links implies to address two questions: (1) where do we start and where do we stop?; (2) how to organize the corpus?

The first question concerns the limits of the dataset. Simply following the cascading links found on websites leads to unmanageable amounts of data and it is therefore required to set a limit to the span of linking to be investigated. First and foremost, we have to determine which news site to study. In this case, no predetermined, exhaustive list of French news sites was established. The research used a snowball sampling scheme (Pickard, 2007, p. 65), and started by exploring the links stemming from the leading French news site, lemonde.fr (OJD, 2010). During the exploration, links to other news sites were discovered, which were consequently explored until no additional news sites appeared in the corpus.

A second border to be settled has to do with the extent to which the news sites are mined. News sites are gigantic and sprawling, with hundred of new stories every day

leading to the creation of new pages. Exhaustively delving into every page is a time and resource-consuming operation, which overwhelms the scope of this research, not to mention the technical features of Navicrawler (which only offers limited functions of large-scale crawling). It was thus decided to solely explore a selected part of the news sites: the most visible top of the iceberg, i.e. the homepage and all the articles that are one click away from it. In so doing, a reasonable puncture of content was achieved.

Finally, the last frontier to determine concerns the linked sites. Which should be included in the corpus and which should be excluded from it? As the research question leading this investigation is open (where do the links found on news sites lead?) such inquiry might seem odd – or at least, the answer obvious: every linked site must be examined and included in the corpus. In this respect, this research differs from other attempts at mapping hyperlink networks (Ghitalla, Jacomy, & Pfaender, 2006; Highfield, 2009; Etling, Kelly, Faris, & Palfrey, 2009; Kelly & Etling, 2008), which are maps of thematically coherent domains. When handling a thematically coherent domain, one must decide if the site she encounters belongs to the field explored or not – with the difficulty of neighbor domains being close to, or sometimes intertwined with, the domain that is explored. Conversely, news websites are in essence not thematically coherent (a characteristic which questions the relevance of analyzing them as networks, an issue discussed in section 4). Every link was therefore included in the corpus, except for those leading to broken or unattainable pages (e.g. those requiring a login and a password) and those clearly originating from ad banners.

The second main operation related to data collection consists of categorizing and tagging sites in order to organize the dataset according to the research question. 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 been refined along the exploration. The organizing process is actually threefold, as it implies constant iterations between the sites, first visualizations and a view of the dataset as statistical tables. Temporary visualizations of the graph open onto the detection of phenomena needing more sophisticated tagging.

The absence of thematic coherence underlined above also implies impediments: it is not relevant to classify sites according to the topic they cover, as this would lead to gigantic categories (as news sites are deemed to cover virtually anything). Moreover, the
topics covered strongly depend on what happens at the time of the data collection, and a thematic categorization would be biased by the hot topics occupying the central stage of news. The chosen categories do therefore rely on more general features likely to cross the whole corpus, such as genre, journalistic function or institutional belonging.

**Data visualization**

Gephi\(^2\), an open-source visualization software, is the tool used to graphically represent the dataset. Representing sites and links in a 2D space is an operation that involves *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, there is no single *true* representation of a “reality”, but a variety of potential representations that depends on what the map-makers want to show. 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). Typically, the maps shown here involve the data being computed through a force-based algorithm, which are conceived as follows: “linked nodes attract each other and non-linked nodes are pushed apart” (Gephi, 2010). Links are the only organizing principle of the map: they “spatially distribute URLs and give birth to 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).

Such particular topology goes hand in hand with reading conventions that might appear as counter-intuitive (RTGI, 2007). For instance, we cannot interpret distances. If a node A is twice as far from node B than from node C, it does not mean that it is twice as distant. Likewise, orientation is meaningless, as there is no significant North-South axis – the map could as well be displayed upside down without losing any of its significance. What strongly matters is density. The maps exhibit zones comparable to wool balls, which denote subsystems of heavily interconnected sites. In the same way, the center-periphery dichotomy is significant: central nodes share links with different areas of the map, whereas peripheral nodes are only connected to a single zone, or even a single site.

3) Results

Eighteen French news sites were explored according to the methods described above. They were not picked in advance as representative of the French mediascape, but discovered along the exploration.

The maps presented in this section display the external links stemming from the news sites explored, as well as any other link found between two nodes of the network (the extent to which each site was mined for links is addressed in the discussion section). 1309 links originate from the 18 news sites. Figure 1 shows the number of links for each of them.

![Figure 1 - Number of external links for each news site](image)

Figure 2 displays the whole network resulting from the exploration, composed of 1076 nodes (i.e. sites) and 7336 edges (links). Its general distribution complies with the physical laws of web-based networks (Barabasi, 2003). For example, the number of links – ingoing links as well as outgoing – follows a power law: a handful of sites link heavily or are heavily linked, while the majority, the long tail, only produces or receives a few links. As the global view is intricate and difficult to read, zooms on subgraphs are needed to better understand the essence of the network. Three specific subgraphs are described here; they are the result of different classifications relying on the function of the links and on the nature or the genre of the site linked.
The media subgraph

Figure 3 shows all the links to media sites. It includes other news sites, sites that constitute a showcase for a media outlet (e.g. the website of a television channel that does not offer news) or media-related sites such as journalists’ union or media regulation organisms. The graph counts 206 nodes and 813 links. Most nodes on this subgraph are French media sites, which form about two-third of the graph, while the last third is composed of international media sites.
This subgraph supports the following observations: Firstly, the important amount of links to other media sites shows that news sites are media-centric. This is consistent with the idea of news workers as a professional community (Deuze, 2005) that fosters links to other members of that community. Secondly, news sites do not seem to be afraid of linking to other media sites, i.e. to potential competitors. Arguing that news sites avoid external linking because it leads to a loss of control over audiences is a line of reasoning that does not apply here: links to other media sites correspond to about one fifth of the whole network. However, a look at roughly the same subgraph with another categorization highlighted must nuance that observation. Figure 4 shows all the French media sites (international and media-related sites have been put aside) and their ownership, represented by the color of the nodes. Media sites belonging to the
same owner tend to be close and to form clusters, which means that they are more connected to each other than they are connected to sites belonging to other owners. A part of the links to other media might thus be explained by commercial concerns and a willingness to reciprocate traffic among media belonging to the same owner: among the total amount of links between French media, 73% are links between sites belonging to the same owner. Commercial closeness therefore explains an important part of the links.

![Figure 4 - French media sites](image)

Similarly, some sites seem to avoid linking to other media sites whereas opportunities exist. For instance, *courrierinternational.com* is mainly constituted of translations of articles from international news media, which is an environment favorable to links as other news media are directly mentioned and used. Even if the source of every article is always clearly mentioned, there is no direct link to the media itself. For example, if the homepage of *courrierinternational.com* features the translation
of an article from the German newspaper *Süddeutsche Zeitung*, it takes three more clicks to find a direct link to the *Süddeutsche Zeitung* website.

**Links not providing additional journalistic content**

Figure 5 displays all the links that have a function that can only be defined negatively: they do not lead to additional journalistic content. The constitution of this particular subgraph results from the emergence, during the exploration, of links to sites that differed from what was expected to be the function of links for news sites (i.e. providing more context, transparency or diversity to a news story). Among those unanticipated sites we find: services such as weather forecasts, train schedules, phone directories or classified ads; showcases for brands; leisure-related sites (travel agencies, amusement parks, etc.); cultural institutions (solely providing practical information such as opening schedules).

Those links do neither seem to offer context to a news story, nor to lead directly to sources and to diverging opinions. They appear to have another function: providing services to the community of readers, leading them to potentially useful (commercial)
sites. We are far from 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. The idea of news media offering services to the community of users is not new: newspapers feature weather pages, or classified ads. For some media outlets, providing services even seems to be a commercial strategy to attract more subscriptions – or to target a particular audience. For example, the French daily Le Figaro proposes a “business subscription” which also offers a concierge service (Daussy, 2010); “restaurants, travels, daily help... the Figaro’s concierge service offers you its privileged services and a VIP treatment in your private and professional lives” (Figaro, n.d.). The fact that online news sites are not merely about news is therefore not a surprise. The size of the phenomenon, however, might attract attention: such links represent one fifth of the whole network.

When it comes to providing links to various services, another strategic stake must be underlined. Some news sites explored not only propose links to service sites, but also integrate them in some ways. For example, they offer entire subsections devoted to weather forecasts or job ads, the content of which is not produced by the news outlet, but by a partner, third-party site. For example, tf1.fr has a job ads section provided by keljob.com, a weather forecast section in partnership with Météo France, and a real estate ads section furnished by seloger.com. Those subsections have a blurred status: they possess a distinct URL – which is why they have been considered as distinct sites in our network – but they are displayed under the banner of the news site. So, why do the news sites bother to build entire dedicated subsections when a link to the concerned partner would provide exactly the same service to the user? Because such “co-branding” (Scalbert, 2009) boosts their audience measurements (Klein, 2009). Commercial concerns seem to prevail.

The phenomenon, again, is neither new nor revolutionary: partnerships with third-party content providers to produce weather pages or classified ads exist in traditional media. The point is that news sites are not afraid of offering external links and of opening their contents when it comes to leading the users to services.
Links to blogs

A third subset worth of attention is displayed in figure 6, which counts 375 nodes and 2014 edges. Those are links to blogs.

Most of them (372 nodes) are blogs that are hosted by one of the news site explored: that is, the news site offers to its users to create blogs, in a similar fashion to what famous blog platforms (e.g. Wordpress, Blogger) do. The blogs are then hosted on the news site; they possess a URL in which the news site’s name is clearly mentioned, and a banner with the look-and-feel of the news site is generally displayed at the top of the blog. Blogs hosted by news sites can be written by journalists belonging to that news organization, or by anyone subscribing to the blogging service.
The phenomenon of blog hosting is interesting *per se*. Why do news sites open up to such user-generated content on the one hand when, on the other hand, the fear of losing control is one of the main arguments against external linking? Firstly, blogs generate traffic, and hosting them allows news sites to have their audiences swelled. Secondly, the fact that news sites attempt to become actors in the blogging market might be understood as an effort to incorporate what stands at the margins of the professional journalistic sphere. In so doing, they comply with the “professional fuzziness” described by Denis Ruellan (2007): as a profession, journalism annexes new fields of practice and incorporate new tasks that have appeared at it borders.

Blogs are also interesting in terms of links, as linking is entirely part of the blog genre. Not only are spaces dedicated to links designed in their structure (with the “blogroll” column for example), but linking is also highlighted as a characteristic of bloggers’ writing style. In this respect, blogs are a window on the web. For instance, if we focus on the 97 blogs hosted by *20minutes.fr* that were discovered in this exploration, and if we dig one level deeper and mine them for links, we find a resulting network of 1477 nodes and 11,690 edges (figure 7). A whole new and complex universe of links appears, only a few clicks away from the news site’s homepage. Sheltering blogs might therefore constitute a way for news sites to integrate into the link ecology while keeping their distance – as the responsibility of linking is somewhat outsourced on blogs.

![Figure 7 - blogs hosted by 20minutes.fr and their links](image-url)
4) Discussion

Put altogether, the three subsets of links examined here (links to other media, links to blogs, and links to services) account for about ¾ of the whole network. They can be examined in terms of promises associated with links for online journalism, namely interactivity, credibility, transparency and diversity.

Links to other media present an apparent diversity. However, a large amount of them might be explained by commercial closeness, which makes the diversity claim fragile. Links to services do not embody functions of credibility, transparency or diversity. They nevertheless suggest some levels of interactivity. Those interactive qualities do not pursue journalistic goals; they rather put the emphasis on the role of news media as caring for their community of users – perhaps with commercial interests in mind. Links to blogs, finally, fulfill the expectation of diversity by leading the users to the bloggers’ various opinions. News sites nonetheless show their appetite for control, as they mainly link to blogs hosted on their own platform.

The reluctance to link externally, highlighted by previous research, must therefore be slightly nuanced. News site have proven their ability to link externally, but mostly on grounds that do not relate to the issues raised by links for journalism. Links with a journalistic purpose are scarce as compared to links with a commercial or practical aim.

The results presented here have to be handled with care in regard to the following limits: first, there is a risk inherent to the visualization method chosen. We should remain wary of the aesthetic appeal of the maps, all the more so that they are not easy to decipher. Our intuitive urge to interpret elements such as distance or orientation must be restricted, because they do not mean anything in this context.

There are also technical limits of which we must be aware. For instance, Navicrawler has a particular way of pondering URLs and their hierarchy. As a consequence, “http://blog.newsmedia.com” is considered as a distinct URL – and therefore as a distinct node in the network – whereas
“http://www.newsmedia.com/blog” is considered as a subdomain of “newsmedia.com”, and thus ignored. Further research must overcome those difficulties.

Thirdly, the following question should be kept in mind: is it really a network? The exploration that has led to the maps presented here was neither flat nor egalitarian. As some nodes (the news sites) were mined deeper than the others, they inevitably have more links. In this respect, the maps presented here are not strictly networks, but rather explorations focused on some determined nodes, puncture in an environment that was only partially probed. This is the reason why the analysis relied solely on basic descriptive parameters such as the number of nodes and links, and not on more complex parameters (e.g. diameter, authorities, hubs) aimed at understanding the dynamics of networks.

As for further research, it converges on the following elements. First, the dataset used here was of exploratory nature. Investigating a larger sample is the next logical step. Besides, only sites related to traditional media were taken into consideration. Pure-players and independent news sites could be scanned, and could offer a fruitful ground for comparison.

In this study, links were considered at large. All the links found within news sites were taken into account (except for those stemming from advertisement). That is, the links structurally present on a page – in the menu or sidebars, for example – as well as the links in the news items themselves. A focus on the latter could constitute a challenging way of investigating issues specifically related to journalism. Finally, exploring the blogosphere and its relation with news sites is a stimulating perspective, as link studies around blogs have proven to be a fertile area of research.

The peek at the hyperlink networks around French news sites proposed by this research is exploratory, partial and perfectible, just as early Renaissance maps were only a first step towards the atlases that we know today. There is an entire universe of links to be discovered and mapped.
References


http://cyber.law.harvard.edu/publications/2009/Mapping_the_Arabic_Blogosphere


